INTERPRETABLE FEATURE DISCOVERY FROM UNSTRUCTURED DATA USING LLMS AND STATISTICAL CHOICE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature extraction from unstructured data typically relies on opaque embeddings unsuited to interpretable decision-making. Interpretable features, on the other hand, typically require costly, and often infeasible, hand-engineering. Large language models (LLMs) offer a natural bridge, as they can read raw data and describe it in interpretable terms. One could simply ask an LLM to extract features from a large sample of the dataset, but context limitations and diluted attention can lead to unsatisfactory results. We introduce a method that automatically discovers interpretable, decision-relevant features from unstructured data. Our key idea is to perform pairwise comparisons between items, asking an LLM to articulate, in interpretable terms, the features distinguishing each pair. We treat these explanations, or *rationales*, as data: we develop a statistical choice model that combines these rationales across the comparisons to identify relevant features and estimate per-item, per-feature scores. We first test our method on a synthetic hiring dataset of cover letters with known ground-truth features and show that it significantly outperforms direct LLM feature extraction in feature recovery. We then apply it to the ChangeMyView (Reddit) debate dataset, where it discovers several interpretable features associated with persuasion while reaching predictive accuracy comparable to established baselines. We also show how the same rationales can be used to audit LLM decisions, yielding a holistic score whose decomposition over interpretable features reveals what the LLM rewards.

1 INTRODUCTION

Consider an instructor tasked with grading essays. She typically first inspects a sample of the essays to inform her grading policy: some may show excellent writing while others have an impressive vocabulary, etc. These comparable attributes between essays are called *features*. From this small sample, she can draw up a rubric, namely the detailed features she will use to grade with *limited bias*. As she grades, she is likely to realize that the original sample that first calibrated her grading policy was too small to contain every important dimension: as the grading goes on, she realizes that grammar is also distinguishing and important, so she needs to go back to the first essays to revise their grading.

This is clearly inefficient, and the instructor would hope to use AI tools to assist in this process of discriminative feature extraction, namely the extraction of the relevant features of the data on which she can base her decisions. If she had essays from previous years together with the grade they obtained, she could train a machine learning model to predict the essays' grades from their *embeddings*, which would serve as features for grading. This would automate the grading, but the features used would be opaque and uninterpretable, which is incompatible with transparent decision-making: a crucial aspect in such applications. Current techniques cannot automate the *interpretable rubric extraction* step. A crucial challenge for decision-makers is therefore: given an unstructured set of items, discover interpretable and discriminative features *for this dataset*, along with per-item per-feature scores, that they can reliably base their decisions on. Additionally, when decision-making is delegated to an LLM, the same features should make the LLM's decisions auditable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A natural approach is to give an LLM the whole set of items and ask it directly for the features that distinguish them (a grading rubric, in the case of essays). However, processing an entire dataset requires an extremely long context, and LLM performance typically degrades as input length increases (Liu et al., 2024; Levy et al., 2024). Consistent with this, we find that direct LLM extraction misses features that are present in the data and conflates others (Section 4). The difficulty is therefore one of *context management*: how to use LLMs on the set of items so that they can reliably identify the features that matter. We introduce an end-to-end approach that, given an unstructured dataset, discovers human-interpretable, dataset-relevant features using LLMs.

Contributions. Our idea is that the relevance of a feature can be revealed by a set of pairwise comparisons of items, and that LLM judges can articulate these comparisons in human language. We call such an articulated comparison a *rationale*: which features the judge cites, and which item it favors on each. The key is that, with enough comparisons, a statistical model can reverse-engineer the judge's rationales and recover which features distinguish the items and how each item scores along each feature.

We treat the rationales as data. A statistical choice model (Section 3) describes each rationale as a two-stage outcome: whether the LLM cites a feature, and, when it does, which item wins on it. The model absorbs the LLM's potential biases, such as a preference for the first or second position in the prompt. Fitting it yields per-item, per-feature scores with confidence intervals, together with diagnostics that filter out features with unreliably estimated scores. Estimation is non-trivial: the likelihood is non-convex and the number of parameters grows with the number of items. We nonetheless prove, under mild assumptions, that the likelihood admits a local maximizer that is consistent for every item score at once in the regime where the number of items grows (Theorems 1 and 2).

We validate our approach extensively in three settings. First, we benchmark it on a synthetic hiring dataset of cover letters with known ground-truth features, where our method correctly recovers all planted features, and most discovered features track a single planted one, whereas methods that prompt an LLM with the whole dataset/a sample of it miss some features and blur several together (Section 4). We then apply it to a real dataset of online debates (the ChangeMyView subreddit) to discover and quantify which argumentation techniques are the most effective. We find that our method reaches an accuracy comparable to the state of the art while providing interpretable features (Section 5) along with insights on argumentation techniques. Finally, we show in Appendix A how an LLM's rationales can also be used to audit its decisions and represent, in interpretable terms, what actually mattered in its decision when comparing two items, yielding interpretability insights for LLM-based scoring.

Our approach provides a concrete procedure that can be supplied as instructions to an LLM when tasked with feature discovery. It offers a principled way to organize multiple LLM agents for this purpose: a lead agent delegates pairwise comparisons along a graph to parallel sub-agents and aggregates their rationales through a statistical model. This aligns with recent efforts to scale AI performance through the orchestration of multiple agents (Li et al., 2024a), where how the work is split and recombined is critical to performance (Kim et al., 2026).

2 Related work

The LLM-as-a-Judge paradigm The *LLM-as-a-Judge* paradigm is an approach to several problems that uses LLMs to evaluate or rank items where one could have solicited a human judge. Examples include model benchmarking: in their foundational work, Zheng et al. (2023) asked both LLMs and humans which of two AI-generated answers to a prompt they preferred, and found substantial agreement between the two, suggesting that LLMs could replace humans for comparing items. Nowadays, LLMs are used as judges in a variety of domains, spanning information retrieval (Sun et al., 2024), recommendation systems (Hou et al., 2024), alignment training (Bai et al., 2022), automated essay scoring (Hou et al., 2025), etc.

LLM biases. Nonetheless, several works (Zheng et al., 2023; Ye et al., 2024a; Gu et al., 2025) document biases in LLMs, such as a *positional bias* (when asked to compare two items, the LLM prefers the item that appears first/second in the prompt for no particular reason), a *verbosity bias* (a

tendency to prefer more verbose items, such as longer essays if the items being assessed are essays), etc. These phenomena can severely undermine LLM-based decision-making if not accounted for.

Feature evaluation with LLMs. Closer to our work, general LLM-based approaches typically work from a predetermined pool of features: the features are scored, but never *discovered from the data*. Evaluation features are either fixed in the prompt (Zheng et al., 2023; Liu et al., 2023), drawn from a predefined skill taxonomy (Ye et al., 2024b), or supplied by the user as a scoring rubric (Kim et al., 2024). These approaches work by asking the judge to output a few tokens per feature: either a pointwise score (Liu et al., 2023; Kim et al., 2024) or the verdict of a pairwise comparison (Zheng et al., 2023; Liusie et al., 2024) ("A" if item A is better, "B" otherwise). The judge's *rationales* are often discarded or used as a qualitative confirmation, while we argue that they could be treated as data useful for feature discovery. Some approaches (Xie et al., 2026) leverage them to create scoring guidelines to explain existing scores and preference labels, but to our knowledge no approach uses them for interpretable feature extraction.

3 METHOD

Our method proceeds in four steps, illustrated in Figure 1. First, LLMs compare pairs of items along a comparison graph, describing the features on which they differ and indicating which item shows more of each feature. These local comparisons require only a small context but are myopic, missing dataset-level information. Second, we group these descriptions into a shared set of features. Third, we fit a statistical choice model that aggregates these local observations into dataset-level, per-item feature scores, using both the cited differences and the features left unmentioned. Finally, we derive per-feature identifiability and reliability diagnostics that flag features whose scores are not pinned down by the observations.

Rationale collection. Let $1, \dots, n$ denote the items. We compare them along a *comparison graph*, a set of N pairs $p = (i, j)$ where item i is shown first. For each pair, an LLM judge is asked how the two items differ: it answers in free text, naming the features on which they differ and which item is stronger on each. These raw features, phrased differently across pairs, are then passed to a second LLM call that groups them into K shared features. The outcome is a sparse *verdict tensor* $\boldsymbol{T} \in \{-1, 0, +1\}^{N \times K}$: $T_{p,k} = 0$ if feature k was not invoked for pair p, and $T_{p,k} = +1$ (resp. -1) if it was invoked and the first (resp. second) item of the pair was cited as stronger.

Statistical choice model. Given the verdicts $\boldsymbol{T}$, our objective is to reverse-engineer the items' scores along each feature and the judge's biases. Our statistical model takes inspiration from Bradley–Terry models (Bradley and Terry, 1952) to estimate latent scores from pairwise comparisons (see Appendix B.1 for related work).

We assume that each item has a *latent score* along each feature, collected in a *score matrix* $\boldsymbol{S} \in \mathbb{R}^{n \times K}$. If pair p compares items i (shown first) and j, we write $\Delta_{p,k} = S_{i,k} - S_{j,k}$ for the true difference of their scores along feature k. Our model posits that these score differences drive the LLM's behavior in two ways: a) the more the items' scores differ on k, the more likely the LLM is to cite feature k; b) when feature k is cited, the item having the larger score along it is more likely to be chosen.

Importantly, we do not assume the LLM does any of this reliably: its citations and choices are noisy and may also carry the judge's biases. We hence model the citation of feature k on pair p as a coin flip of probability $\sigma(f(\Delta_{p,k}) + \gamma_k)$, where σ is the sigmoid function and f is an even function increasing in $|\Delta|$ (we use $f = \log(2\cosh)$, a smooth surrogate of the absolute value), so that the *non-citation* probability is

$$\mathbb{P}(T_{p,k} = 0) = 1 - \sigma(f(\Delta_{p,k}) + \gamma_k). \tag{1}$$

Here γ_k is the baseline citation propensity of feature k, which accounts for a feature-specific eagerness or reluctance of the LLM to cite it. Under this model, silence is also informative: a non-citation suggests that the judge did not find the pair's difference on that feature salient enough to cite it.

Our model also assumes that, conditioned on citing feature k, the LLM declares the first item the winner with probability $\sigma(\Delta_{p,k} + \beta)$, where β captures the positional bias widely reported in the

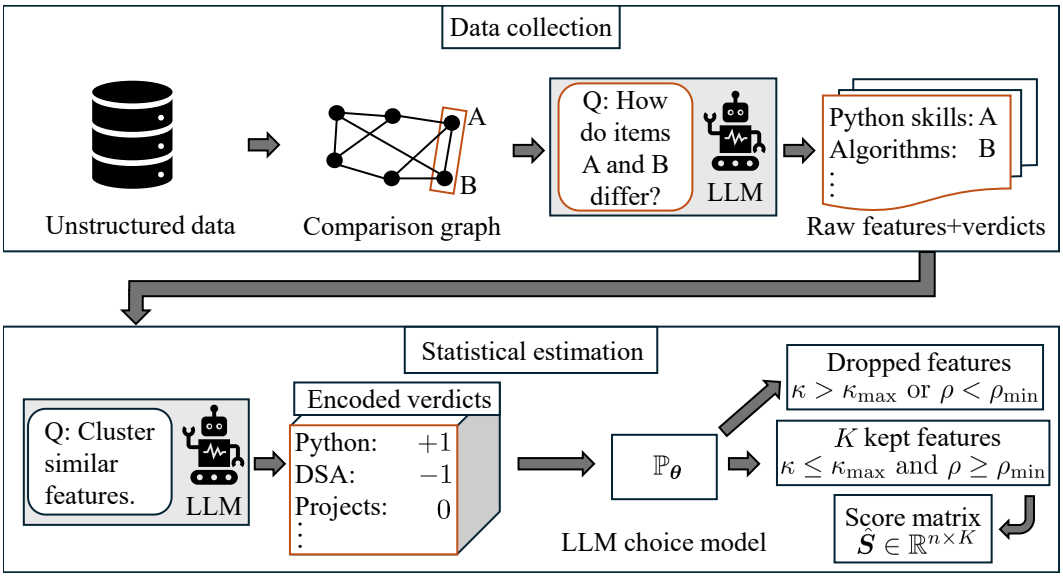

Figure 1: **Overview of the pipeline described in Section 3.** Items are first compared in pairs by the LLM, cited features are then clustered, and a statistical model learns the relevant features from the LLM's choices. Finally, statistical diagnostics identify the reliable scores. The orange running example follows one pair of cover letters (the setup of Section 4) through the pipeline.

literature (Zheng et al., 2023; Ye et al., 2024a; Gu et al., 2025), which we also observe in our experiments. Multiplying the two gives:

$$\begin{aligned} \mathbb{P}(T_{p,k} = 1) &= \sigma(\Delta_{p,k} + \beta)\, \sigma(f(\Delta_{p,k}) + \gamma_k) \\ \mathbb{P}(T_{p,k} = -1) &= \big(1 - \sigma(\Delta_{p,k} + \beta)\big)\, \sigma(f(\Delta_{p,k}) + \gamma_k). \end{aligned} \tag{2}$$

Hence, β and γ_k capture the LLM's biases (a preference for a position in the prompt, and feature-specific citation habits) so that they do not contaminate the estimated scores. The position coefficient β is shared across features; as we show in Appendix B.2, performing comparisons with presentation order randomized makes the scores essentially insensitive to this choice, and a per-feature refit confirms it empirically.

Parameter estimation. We estimate $\boldsymbol{\theta} := (\mathrm{vec}(\boldsymbol{S}), (\gamma_k)_{k=1}^K, \beta) \in \mathbb{R}^{nK+K+1}$ from the verdict tensor $\boldsymbol{T}$ by penalized maximum likelihood,

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}} \Big\{ \mathcal{L}(\boldsymbol{\theta}) + \frac{\lambda}{2} \|\boldsymbol{S}\|_F^2 \Big\}, \tag{3}$$

where $\mathcal{L}$ is the negative log-likelihood of $\boldsymbol{T}$ and $\lambda \ge 0$ is a ridge parameter. Under conditional independence of the verdicts across pairs and features given $\boldsymbol{\theta}$, $\mathcal{L}$ factorizes as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) = -\sum_{\text{pairs } p} \sum_{\text{features } k} \Big[& \mathbf{1}_{\{T_{p,k} \neq 0\}} \log \sigma\big(f(\Delta_{p,k}) + \gamma_k\big) + \mathbf{1}_{\{T_{p,k}=0\}} \log\big(1 - \sigma(f(\Delta_{p,k}) + \gamma_k)\big) \\ & + \mathbf{1}_{\{T_{p,k}=1\}} \log \sigma\big(\Delta_{p,k} + \beta\big) + \mathbf{1}_{\{T_{p,k}=-1\}} \log\big(1 - \sigma(\Delta_{p,k} + \beta)\big) \Big]. \end{aligned} \tag{4}$$

The likelihood is invariant to shifting all scores of a feature by a common scalar; the ridge penalty resolves this by selecting, within each invariance class, the centered representative $\mathbf{1}^\top \boldsymbol{s}_k = 0$, where $\boldsymbol{s}_k$ is the k-th column of $\boldsymbol{S}$.

Theoretical guarantees. Our estimator raises a nonstandard consistency question: the number of parameters grows with the number of items n, and the negative log-likelihood (4) is mildly non-convex. We prove, under reasonable assumptions, that the likelihood admits a local maximizer that is consistent for every item score at once as n grows with a bounded number of comparisons per pair, in the case $K = 1$ (single feature).

Theorem 1 (Local MLE and uniform consistency, dense regime)**.** *Assume A1–A3 (every pair is compared at least once and at most a fixed number of times, independent of n; the true scores are bounded; f is even and C^3) and A6 (the population information for (γ, β), after profiling out $\boldsymbol{s}$, has smallest eigenvalue of order n^2). For every $c > 0$ there exist constants C, n_0 depending only on c and the constants in the assumptions such that for all $n \geq n_0$, with probability at least $1 - O(n^{-c})$, the negative log-likelihood* (4) *admits a stationary point $(\hat{\boldsymbol{s}}, \hat{\gamma}, \hat{\beta})$ on the hyperplane $\{\mathbf{1}^\top \boldsymbol{s} = 0\}$ satisfying*

$$\|\hat{\boldsymbol{s}} - \boldsymbol{s}^*\|_\infty \leq C\sqrt{\frac{\log n}{n}}, \qquad \max\left(|\hat{\gamma} - \gamma^*|, |\hat{\beta} - \beta^*|\right) \leq C\sqrt{\frac{\log n}{n}}. \tag{5}$$

Moreover, this point is the unique stationary point in a neighborhood $\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n) \times \mathcal{B}(\gamma^*, r_n) \times \mathcal{B}(\beta^*, r_n)$ with $\varepsilon_n, r_n = \Theta\left(\sqrt{\frac{\log n}{n}}\right)$, and it maximizes the likelihood there: it is the local MLE.

In Appendix B, we prove a more general version of Theorem 1 (Theorem 2) in which both channels carry arbitrary comparison-level covariates and γ and β are coefficient vectors.

In practice, we optimize the regularized likelihood (3) numerically on sparse comparison graphs, and the fitted model is well calibrated out of fold (Figures 13 and 14). Theorem 1 assumes a dense design and does not cover this regime; what it guarantees is that, near the truth, the non-convex likelihood has a unique stationary point, which is a consistent local maximizer, so a local optimizer targets a well-defined estimator rather than an arbitrary critical point.

Extracting score reliability. Because $\hat{\boldsymbol{\theta}}$ is an M-estimator, it satisfies a central limit theorem as the number of comparisons grows, which yields confidence intervals on every parameter (see Appendix B.3 for more details). The same machinery gives two per-feature diagnostics: a condition number κ_k of the score block of the Fisher information, which flags features whose scores are poorly identified by the data (Rothenberg, 1971), and a reliability coefficient ρ_k, the fraction of the observed score spread attributable to true between-item differences rather than estimation noise. Features with $\kappa_k > \kappa_{\max}$ or $\rho_k < \rho_{\min}$, for thresholds fixed a priori, are filtered out before a final refit.

Scoring new items. Running the full pipeline on the entire dataset allows the statistical model to aggregate comparisons globally, correct the judge's biases, and provide reliability diagnostics. With d comparisons per item, this requires $dn/2$ pairwise LLM calls. A cheaper alternative is to discover features on a representative subset and then ask an LLM to score additional items independently along those features, requiring one call per item. This pointwise scoring requires specifying precisely how each feature should be scored, which may require human contextual knowledge of the dataset. It is also subject to well-documented biases of pointwise LLM scoring (Liusie et al., 2024). Alternatively, one can compare each new item pairwise against a fixed number of already-scored items along the fixed features and fit a Bradley–Terry model for each feature. This retains dataset-level aggregation and position-bias correction without repeating feature discovery. The choice trades cost against robustness to pointwise biases; both are evaluated in Section 5.

4 Empirical Results on Interpretable Feature Extraction

We first evaluate our approach in a synthetic hiring application where we control the ground-truth features, allowing us to test which features each method recovers and how accurately it recovers them. We construct a dataset of $n = 200$ cover letters responding to a real Google security software engineering job advertisement, with known features (which we call *dials*) embedded in each letter. See Appendix C for details about the experiment.

Dataset. Each letter realizes 18 engineered ground-truth *dials*: 8 primary ones (*years of experience*, *education level*, *school prestige*, *number of languages*, *production launches*, *skill coverage*, *employer prestige* and *prose fluency*), 6 binary role-skill presences (e.g. Python, *security assessments*), and 4 three-level niche dials (e.g. *open-source involvement*, *security certifications*) exactly two of which are present in each letter.

Dials are drawn near-orthogonally and with capped correlation to letter length, so that each dial is a separately recoverable signal and length cannot proxy for quality. Of the 18 generated features,

$D = 16$ are relevant recovery targets, while two serve as controls: *skill coverage* is a redundant composite of the individual skill features, and *prose fluency* is irrelevant to the job.

Metrics. A discovered feature is said to *track* a dial if the cover letters' scores along the feature correlate sufficiently with the dial. A discovered feature is said to be *about* the dial if its name and the dial's name are sufficiently aligned in TEXT-EMBEDDING-3-LARGE embedding space. We will compare methods along the following metrics (defined mathematically in Appendix C.2.1).

- **Cost.** Total method cost in USD.
- **Statistical recovery (Tracked & Stat.R.).** The fraction of tracked dials (Tracked), and how strongly on average each dial is tracked (Stat.R.).
- **Semantic recovery (Sem.R.).** The fraction of dials that have a discovered feature that is both about it and tracks it. By definition, it is at most Tracked: the gap between the two counts dials tracked under a misaligned name.
- **Attributable features (Attrib.).** The fraction of the discovered features that track a single dial.
- **Leakage (LK).** The average correlation between a dial and the attributable features other than the one that best tracks it. Since dials are near-orthogonal by construction, a feature that measures one dial should be uncorrelated with the others; leakage quantifies how much of the other dials nonetheless bleeds into its scores (lower is better).

Benchmarks. We compare our approach to direct LLM feature extraction and its variants. Specifically, we provide a powerful LLM (GPT-5.4 (HIGH REASONING)) with the whole dataset or a subset and ask it to identify features and formulate dataset-specific guidelines for scoring them. We then use a cheaper model (GPT-5.4 MINI) to score each cover letter on each feature. More precisely, we consider the following variants:

- **Direct LLM:** An LLM is queried on the full dataset to discover features, after which a cheaper LLM scores each item along each feature. We also consider **Direct LLM–Multi-score**, where multiple agents score each item with slightly different prompts, and their scores are averaged. This variant approximately matches our approach's cost, allowing us to evaluate whether the benefits arise from using multiple agents or from our specific way of orchestrating them.
- **Local LLMs:** Multiple agents each examine a randomly selected subset of the dataset and discover features. We pool their features and use an LLM to merge equivalent descriptions into a shared set (see Remark 7 in Appendix C for the choice of merging procedure). A single agent then scores each item along these features. Analogously, we also consider **Local LLMs–Multi-score**. These variants help assess whether the performance benefits arise from distributing feature discovery across multiple agents or from our pairwise comparisons and statistical aggregation.

Ablations. To examine the contributions of statistical scoring and reliability filtering, we retain the features discovered through our sub-agents' pairwise comparisons and consider two ablations, each with single-agent and multi-agent scoring.

- **Pairwise comparisons + LLM scoring:** LLMs directly score all our discovered features, replacing statistical scoring and its reliability filtering.
- **Reliable features + LLM scoring:** An LLM scores only the features retained by our statistical model's reliability filter. Comparing these two ablations isolates the contribution of filtering under LLM scoring.

Results. Table 1 reports the metrics of each benchmark and ablation. Our approach outperforms direct LLM feature extraction, even though that benchmark uses one of OPENAI's most capable models available at the time of the experiments. It tracks all 16 ground-truth features, compared with 13 for Direct LLM, and achieves complete semantic recovery: the discovered features both track the underlying signals and describe them meaningfully. The three missing dials are a discovery failure rather than a scoring failure: at matched cost, Direct LLM–Multi-score still tracks the same 13. More of our features are attributable to a single ground-truth feature (16 of 21 versus 13 of 21), and leakage is lower (0.090 versus 0.109). Direct LLM is however cheaper ($0.26 versus $3.45).

	Method	Cost	K	Tracked	Stat.R.	Sem.R.	Attrib.	LK↓
	Ours	3.45	21	**100.0%**	0.850	**100.0%**	**76.2%**	**0.090**
Benchmarks	Direct LLM	**0.26**	21	81.2%	0.732	81.2%	61.9%	0.109
	Direct LLM–Multi-score	3.42	21	81.2%	0.787	81.2%	66.7%	0.118
	Local LLMs	2.35	23	75.0%	0.702	68.8%	47.8%	0.110
	Local LLMs–Multi-score	3.27	23	75.0%	0.737	68.8%	47.8%	0.113
Ablations	Pairwise comparisons + LLM scoring	3.67	32	93.8%	0.763	87.5%	40.6%	0.093
	Pairwise comparisons + LLM scoring–Multi-score	6.55	32	**100.0%**	0.860	**100.0%**	43.8%	0.104
	Reliable features + LLM scoring	3.60	21	93.8%	0.792	93.8%	66.7%	0.098
	Reliable features + LLM scoring–Multi-score	6.54	21	**100.0%**	**0.868**	**100.0%**	66.7%	0.103

Table 1: Per-method estimates on the single full-data run. Costs are Batch-API prices in USD; K is the number of features the method reports. Tracked, Sem.R., Attrib. in %; Stat.R. and LK are correlations.

Our approach also outperforms the multi-agent alternatives, whether multiple agents are used for feature discovery (Local LLMs), scoring (Multi-score), or both, with statistically significant gains in tracking, attribution, and leakage over Local LLMs (Figure 15). Multi-agent scoring improves statistical recovery but increases leakage, while multi-agent feature discovery reduces recovery compared with a single agent with the full dataset. The benefits therefore do not follow simply from using multiple agents: how we divide their work into pairwise comparisons and aggregate their outputs through a statistical model matters.

The ablations also achieve high tracking and semantic recovery, consistent with their use of the same features as the full pipeline, but without statistical scoring and its reliability filter, attribution is substantially lower (13–14 attributable features versus 16) and leakage higher (0.103–0.104 for Multi-score methods versus 0.090), despite a slightly higher Stat.R. (0.860–0.868 versus 0.850) and twice the cost. The resulting scores therefore separate the ground-truth signals less clearly. Filtering alone raises the attributable fraction under multi-agent scoring from 14 of 32 to 14 of 21 while preserving complete tracking and semantic recovery, showing that it removes features that track no dial. Statistical scoring on the same 21 features then raises it to 16.

5 Application to a Real Dataset

We next evaluate our approach on a real-life dataset whose underlying features are unknown: online debates from Reddit's ChangeMyView forum. The central question is whether automatically discovered features provide interpretable insights into persuasion while retaining sufficient signal to predict which arguments change a participant's view.

The ChangeMyView dataset. In the /r/ChangeMyView (CMV) subreddit, a user (the original poster, OP) posts an opinion with their reasoning and invites counterarguments. The OP then awards an explicit delta (Δ) to any reply that changed their view, providing labels of persuasion success. Tan et al. (2016) turned this signal into a controlled paired prediction task: predicting which of two responses to a post actually convinced the OP. Hence, the items being assessed by our feature discovery pipeline will be *exchanges*, namely pairs of the form (OP post, reply). See Appendix D for details about the experiment. Prior work on this task either hand-engineers features (word overlap, style and bag-of-words statistics in Tan et al. (2016); argumentation-theoretic features in Wachsmuth et al. (2017)) or learns non-interpretable embeddings (Durmus and Cardie, 2018).

Extracting interpretable features. We apply our pairwise feature-discovery procedure to a subset of 300 training exchanges, grouping the resulting descriptions into 29 reply-side features, of which our reliability diagnostics retain 16. These features capture human-recognizable aspects of argumentation, including *concrete illustrations*, *structured rebuttals*, *conceptual clarification*, and *confrontational tone*, without requiring hand-engineering. Table 10 in Appendix D.3.2 lists all 29 features and their descriptions. Figure 3 illustrates their associations with persuasion: it reports how often a successful reply scores higher on each feature than its matched unsuccessful reply. For instance, features related to *evidence and cases* show stronger positive associations with persuasion

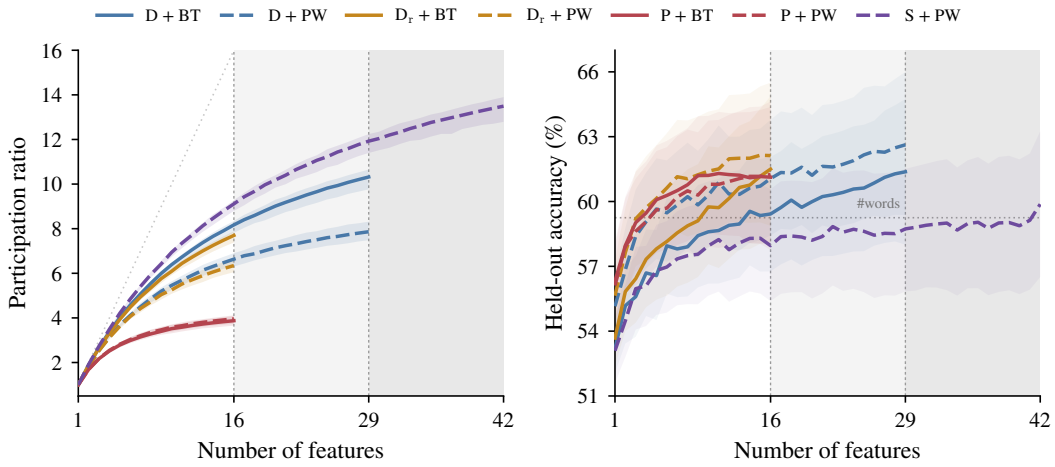

Figure 2: **Discovered features span more effective dimensions than prior ones while retaining predictive accuracy.** In both panels, each point is an average over random k-subsets of the feature set, so that sets of different sizes are compared at matched size. *Left:* noise-corrected participation ratio of the feature-score correlation matrix (Spearman) over the 1,600 held-out exchanges; dotted: $y = k$. *Right:* held-out accuracy of the ℓ_1 logistic regression on the 800 held-out pairs; chance $= 50\%$. D: discovered (29 features); D_r: reliable subset of D (16); P: prior (16); S: sample-based (42); BT/PW: Bradley–Terry/pointwise scoring; #words: reply length. Bands: 95% bootstrap percentile intervals over held-out exchanges (left) and exchanges + pairs (right). Shaded: $k > 16$, where P and D_r are not defined, and $k > 29$, where only S is defined. Details in Appendix D.3.1.

than those related to *tone and appeals*. *Confrontational tone* shows a negative association with persuasion, although this association is not statistically significant.

Benchmarks and ablations. We write D for the 29 discovered features and D_r for the 16 that pass the reliability filter, and compare both to isolate the contribution of the statistical model. In addition, we compare our approach to features obtained from two types of direct LLM call. In the first, similarly to the benchmarks of Section 4, GPT-5.4 (HIGH REASONING) is given a sample of exchanges and asked to propose distinct features, each with a name and a short definition (S, for samples; 42 features). In the second, the model is given no data, only a brief description of the persuasion task (P, for prior; 16 features).

For each feature set, the features are held fixed and the training and held-out exchanges are scored along them with GPT-5.4 MINI, following the two scoring protocols of Section 3 (*Scoring new items*). In pointwise scoring (**PW**), one call per exchange assigns it a 0–10 score on every feature (Appendix D.2). In pairwise scoring (**BT**), the LLM compares two exchanges on every feature at once and a Bradley–Terry model with a positional-bias term is fitted per feature: training exchanges are compared among themselves, and each held-out exchange is then compared against 16 training exchanges only, whose scores are kept frozen, so that all scores lie on a common scale and new items are scored at fixed cost. Since S has 42 features, its pairwise calls would be prohibitively costly, so we report S under PW only.

Expressiveness of the discovered features. The left panel of Figure 2 reports the noise-corrected participation ratio (Gao et al., 2017) of random subsets of each feature set, which summarizes the effective number of uncorrelated dimensions in the feature scores. Features derived from the data, whether from a sample of exchanges (S) or from our pairwise comparisons (D, D_r), are far less redundant than features elicited from the model's prior (P): at 16 features, D and D_r span about eight effective dimensions under BT scoring and six to seven under pointwise scoring, against about four for P, whose curve has flattened before 16. The higher dimensionality under BT scoring is consistent with our intuition that aggregating comparisons at the dataset level preserves distinctions that myopic pointwise LLM scoring blurs.

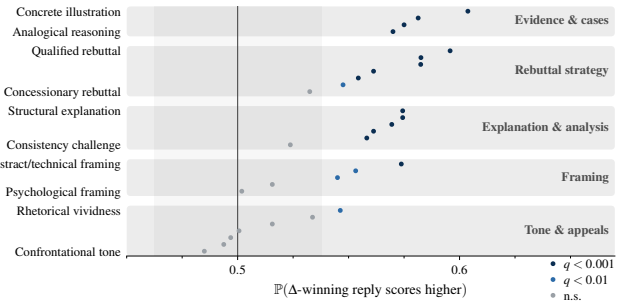

Figure 3: **Effectiveness of the discovered argumentation techniques.** Discovered features are gathered in five categories. For each discovered feature (D + PW), we report the *effectiveness*, i.e. the fraction of the test pairs in which the reply that earned the Δ scores higher on that feature; 0.5 is chance. BH q is the Benjamini–Hochberg false-discovery rate across the 29 tests. See Figure 21 for the figure showcasing every feature's effectiveness with confidence intervals.

Arm	Dims	Accuracy
Tan	3,205	63.75% [60.38, 67.00]
D + PW	29	62.63% [59.38, 66.00]
D_r + PW	16	62.13% [58.88, 65.50]
Emb.	3,072	61.63% [58.25, 65.00]
D_r + BT	16	61.50% [58.13, 64.88]
D + BT	29	61.38% [58.00, 64.75]
P + PW	16	61.25% [58.00, 64.63]
P + BT	16	61.13% [57.75, 64.38]
S + PW	42	59.88% [56.50, 63.25]
#words	1	59.25% [55.88, 62.75]
Zero-shot	–	56.94% [54.19, 59.88]

Table 2: **Held-out accuracy of each method**, with 95% bootstrap percentile intervals over the 800 held-out pairs; chance = 50%. Emb.: TEXT-EMBEDDING-3-LARGE embeddings. Tan: hand-engineered features of Tan et al. (2016). Zero-shot: the judge predicts the Δ-winning reply. #words: reply length.

Predictive signal. We fit an ℓ_1-regularized logistic regression on the scores of each feature set to predict which of two replies to a post persuaded the OP. Table 2 reports the held-out accuracy of each approach, alongside hand-engineered features, embeddings, and simpler baselines. Our approach (D and D_r) achieves accuracy comparable to the best methods (see Table 9 for full data estimates) while providing interpretable features.

Beyond raw accuracy, how much predictive signal a feature set carries relative to its size matters for interpretability. The right panel of Figure 2 therefore reports held-out accuracy against the number k of features, averaged over random k-subsets of each set. Among the expressive, non-redundant feature sets identified in Figure 2 (S, D, D_r), ours are the most accurate: the direct-LLM sample-elicited set S has more features and a high effective dimensionality, yet it is the least accurate at almost every k, 4.2 points below D_r + PW at $k = 16$ and 3.6 points below D_r + PW on average over k (paired 95% CI $[1.8, 5.4]$; see Figure 19 in Appendix D.3.1). This indicates that part of the variation S captures is irrelevant to persuasion. Conversely, the features elicited from the LLM without data access (P) are among the most predictive at low k but plateau quickly, illustrating the limits of data-blind feature elicitation.

The results also show that the statistical model plays an important role. The filtered set D_r has about half as many features as D, yet loses at most half a point of accuracy. Hence, the statistical model retains the informative features, yielding a compact, highly interpretable set with a strong predictive signal. The right panel of Figure 2 confirms this: at every k, random subsets of D_r are more accurate than equally sized subsets of D, under both scorings, by one point on average over $k \leq 16$ (paired 95% CIs $[-0.1, 2.2]$ for PW and $[0.0, 2.1]$ for BT), showing that features retained by the filter carry more signal than the unfiltered ones.

6 Conclusion

We introduced an approach that discovers human-interpretable features from unstructured data using LLMs: judges compare items pairwise, and a statistical choice model aggregates their rationales. On both a controlled experiment and real-life data, it recovers features significantly better than a direct LLM call, even with GPT-5.4 (HIGH REASONING). The recovered features capture most of the predictive signal with fewer, less redundant features, while matching the accuracy of established baselines. Appendix A further extends the framework to auditing an LLM's own decisions.

Reproducibility statement

Code to reproduce every experiment in the paper is available at the anonymized repository `https://anonymous.4open.science/r/Interpretable-Feature-Discovery-LLM-2D23`, and will be released publicly upon publication. The mathematical guarantees of the introduced statistical choice model are proved in Appendix B, and full details of every experiment are given in Appendices C and D.

Statement on LLM Usage

Because LLMs are the object of study, the judge outputs, pairwise comparisons and pointwise scores analyzed throughout, as well as the synthetic cover-letter corpus of Section 4, were produced by LLMs through API calls.

Beyond this experimental role, our use of generative AI was as follows. The text of the paper was written by the authors, with LLMs used for occasional rewording. Portions of the experimental code were written with LLM assistance; all AI-assisted code was reviewed and tested by the authors, who take full responsibility for its correctness. The mathematical claims and proofs were developed by the authors without AI assistance; as an additional verification step, we used Claude Fable 5 (Anthropic) to audit the proofs, which surfaced an error; the authors diagnosed and corrected it by hand.

REFERENCES

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, 2024.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need, 2024a. URL https://arxiv.org/abs/2402.05120.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2026. URL https://arxiv.org/abs/2512.08296.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agents, 2024. URL https://arxiv.org/abs/2304.09542.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems, 2024. URL https://arxiv.org/abs/2305.08845.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Zhaoyi Joey Hou, Alejandro Ciuba, and Xiang Lorraine Li. Improve llm-based automatic essay scoring with linguistic features, 2025. URL https://arxiv.org/abs/2502.09497.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024a. URL https://arxiv.org/abs/2410.02736.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets, 2024b. URL https://arxiv.org/abs/2307.10928.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL https://arxiv.org/abs/2310.08491.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models, 2024. URL https://arxiv.org/abs/2307.07889.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling, 2026. URL https://arxiv.org/abs/2510.17314.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624. International World Wide Web Conferences Steering Committee, April 2016. doi: 10.1145/2872427.2883081. URL http://dx.doi.org/10.1145/2872427.2883081.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1017/.

Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1094. URL https://aclanthology.org/N18-1094/.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262. URL https://www.biorxiv.org/content/early/2017/11/12/214262.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024b. URL https://arxiv.org/abs/2406.11939.

Scott A. Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269, 2023. doi: 10.1075/ijlcr.22026.cro.

Scott Crossley. A large-scale corpus for assessing source-based writing quality: Asap 2.0, January 2025. URL https://doi.org/10.5281/zenodo.14781349.

W. A. Glenn and H. A. David. Ties in paired-comparison experiments using a modified thurstone-mosteller model. *Biometrics*, 16(1):86–109, 1960. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2527957.

P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967. doi: 10.1080/01621459.1967.10482901. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1967.10482901.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2283595.

Roger R. Davidson and Robert J. Beaver. On extending the bradley-terry model to incorporate within-pair order effects. *Biometrics*, 33(4):693–702, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529467.

Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3), August 2012. ISSN 0883-4237. doi: 10.1214/12-sts396. URL http://dx.doi.org/10.1214/12-STS396.

Daniel McFadden. Modelling the choice of residential location. In Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen W. Weibull, editors, *Spatial Interaction Theory and Planning Models*, pages 75–96. North-Holland, Amsterdam, 1978.

Axel Börsch-Supan. On the compatibility of nested logit models with utility maximization. *Journal of Econometrics*, 43(3):373–388, 1990. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(90)90126-E. URL https://www.sciencedirect.com/science/article/pii/030440769090126E.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/120149.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017. doi: 10.1287/opre.2016.1534. URL https://doi.org/10.1287/opre.2016.1534.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence, 2016. URL https://arxiv.org/abs/1505.01462.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *The Annals of Statistics*, 47(4), August 2019. ISSN 0090-5364. doi: 10.1214/18-aos1745. URL http://dx.doi.org/10.1214/18-AOS1745.

Ting Yan. Inference in a generalized Bradley-Terry model for paired comparisons with covariates and a growing number of subjects. arXiv preprint arXiv:2507.22472, 2025.

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233, 1967.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610, 2007.

CONTENTS OF THE APPENDIX

A Extension: Interpretable LLM-based scoring **15**

A.1 Details of the extension 17

B Statistical Model Theory **28**

B.1 The Bradley–Terry with ties literature 28

B.2 Why a shared position bias coefficient? 29

B.3 Uncertainty quantification 30

B.4 Reliability diagnostics 30

B.5 Model, likelihood, score, and Hessian 31

B.6 Notation and external tools 33

B.7 Assumptions and the population curvature floor 34

B.8 The leave-one-out construction 35

B.9 Concentration 36

B.10 Hessian estimates 39

B.11 Stage 1: existence of a local consistent estimator and ℓ_∞ rates 43

B.12 Stage 2: the profiled likelihood and joint consistency 51

C Experiment Details on Interpretable Feature Recovery **57**

C.1 Dataset 57

C.2 Methods Details 60

C.3 Additional Results 67

C.4 Run-variance analysis 71

D Experiment Details on the ChangeMyView Dataset **76**

D.1 Dataset 76

D.2 Methods Details 77

D.3 Additional results 78

A EXTENSION: INTERPRETABLE LLM-BASED SCORING

The goal of this small extension is to provide a method inspired by the feature discovery pipeline but well-suited to the goal of interpretable scoring. LLMs are increasingly used to score items (essays, résumés, or other LLMs) through pairwise comparisons, and the resulting verdicts are typically aggregated by fitting a Bradley–Terry model. This yields a single scalar score per item, whereas a decision-maker would want an overall score that decomposes into per-feature scores along interpretable dimensions. For example, for an LLM grading scientific research papers, have an overall score per-paper that aggregates scores on atomic features *discovered from the data*: scientific rigor, novelty of the results, etc.

Method. Established methods (e.g. Li et al. (2024b)) have the LLM make pairwise comparisons and fit a Bradley–Terry model on their outputs to yield scores. Performing this on many features gives several scores that can be aggregated to one overall score that decomposes cleanly on atomic features, but it requires knowing the assessable features.

Our method works by asking the LLM judge for a holistic verdict, and asking it to justify it by specifying *along which feature* the winning item is better than the losing one. The canonicalization of cited qualities into K shared features is identical to Section 3. The data for a compared pair p are therefore the overall verdict $w_p = \mathbf{1}_{\text{first-presented item wins}}$ and the citation indicators $\boldsymbol{c}_p \in \{0,1\}^K$, where $c_{p,k} = 1$ if feature k was cited as a quality the winner exhibits more; the direction of each citation is thus implied by w_p.

Statistical model. To infer per-item per-feature scores from the judge's justifications, our model links the holistic verdict to the per-feature scores by setting:

$$\mathbb{P}(w_p = 1) = \sigma\left(\sum_{k=1}^{K} \Delta_{p,k} + \beta\right), \tag{A.1}$$

where $\Delta_{p,k}$ is the score gap on feature k (first minus second) and β is the position coefficient. Hence, the model simply posits that the larger the holistic score, the likelier an item is to win the comparisons it is involved in.

To model citations, the model posits that (conditionally on the overall verdict) each feature is cited independently, with a probability that increases in the margin by which that feature favors the *winner* (because we ask the LLM to justify its winning-item choice):

$$\mathbb{P}(c_{p,k} = 1 \mid w_p) = \sigma\big((2w_p - 1)\Delta_{p,k} + \gamma_k\big). \tag{A.2}$$

Here $2w_p - 1$ is 1 when the first item won and -1 otherwise.

The likelihood of an observed pair is therefore:

$$\mathbb{P}(w_p, \boldsymbol{c}_p) = \sigma\Big(\sum_{k=1}^{K} \Delta_{p,k} + \beta\Big)^{w_p} \Big(1 - \sigma\Big(\sum_{k=1}^{K} \Delta_{p,k} + \beta\Big)\Big)^{1-w_p} \prod_{k=1}^{K} \sigma(\eta_{p,k})^{c_{p,k}} \big(1 - \sigma(\eta_{p,k})\big)^{1-c_{p,k}}, \tag{A.3}$$

with $\eta_{p,k} = (2w_p - 1)\Delta_{p,k} + \gamma_k$, and the NLL sums $-\log$ of (A.3) over pairs, ridge-penalized as in Eq. (3).

The method therefore not only discovers the features and estimates each item's score on them, but also links these per-feature scores to a holistic verdict, so that the overall judgment decomposes additively into named, interpretable components.

Interpretability. To ensure the model's identifiability, the fitted scores are mean-zero. If one sets $\sigma_k := \mathrm{std}(\boldsymbol{S}_{\cdot,k})$, $\tilde{S}_{i,k} := \frac{S_{i,k}}{\sigma_k}$ and $\tilde{\Delta}_{p,k} = \frac{\Delta_{p,k}}{\sigma_k}$, Equation (A.1) becomes:

$$\mathbb{P}(w_p = 1) = \sigma\left(\sum_{k=1}^{K} \sigma_k \tilde{\Delta}_{p,k} + \beta\right), \tag{A.4}$$

with each $\tilde{S}_{i,k}$ having mean zero and variance 1. One can hence interpret σ_k as the judge's perceived importance of feature k in the overall verdict, yielding interesting interpretability insights about what each judge rewards.

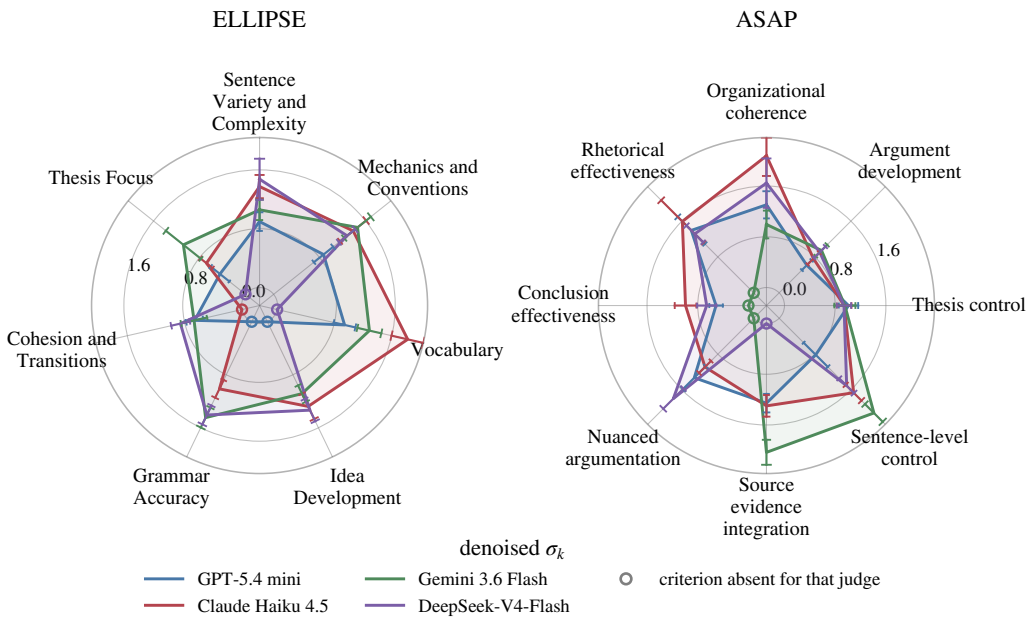

Figure 4: **Which features matter for each judge on each dataset.** Radial axis is σ_k (with estimation noise removed) on each judge's reliable features ($\rho > 0.5$, $\kappa < 100$). Bars are 95% delta-method intervals from the sandwich covariance. Axes are cross-judge clusters formed by meaning (GPT-5.4 (HIGH REASONING)), restricted to constructs at least three of the four judges discovered (see Figures 6 and 7 for the full plots). An open marker denotes a construct absent from that judge's reliable basis (never surfaced, or removed by the identifiability screen), not a zero estimate.

Application. We demonstrate the method on two essay corpora with human holistic scores: (i) $n = 100$ essays from ELLIPSE (Crossley et al., 2023), a corpus of English-language-learner writing from grades 8–12 scored for overall language proficiency (as well as six analytic dimensions), and (ii) $n = 100$ essays from ASAP 2.0 (Crossley, 2025), a corpus of persuasive essays from grades 6–10 scored for overall writing quality. Each corpus is compared along a $d = 40$-regular graph. To assess whether the discovered constructs and their importance depend on the judge, we run the pipeline with four judges from four model families: GPT-5.4 MINI, CLAUDE HAIKU 4.5, GEMINI 3.6 FLASH, and DEEPSEEK-V4-FLASH. Canonicalization is performed by GPT-5.4 (HIGH REASONING) in every case, and every feature passes through reliability gates analogous to those of Section 3. To compare the judges' feature sets, we match features across judges semantically with GPT-5.4 (HIGH REASONING) using Prompt 3; Figure 4 shows the resulting shared constructs.

Results. Figure 4 shows that the judges agree on *which* constructs structure their verdicts far more than on *how much* each one separates essays. On ELLIPSE, every construct shared by three judges or more is lexical, mechanical, or organizational, consistent with what varies in L2 learner writing; on ASAP 2.0, the four-judge constructs are discourse-level (thesis control, organizational coherence, argument development, sentence-level control). Since the same pipeline, pairs, and prompts (up to corpus framing) produce both bases, this reorganization is attributable to the corpora rather than to any judge's habits: the method recovers what the task actually varies on. Within the shared constructs, magnitudes disagree: GPT-5.4 MINI is the most compressed judge throughout (smallest σ_k on 5 of 7 ELLIPSE and 4 of 8 ASAP 2.0 consensus axes), thesis control on ASAP 2.0 is the one construct on which all four judges essentially coincide ($\sigma_k \in [0.94, 1.01]$), and sentence-level control is the most contested (0.82 to 2.11). Interestingly, the largest spreads on both corpora sit on features surfaced by a *single judge*: the sign of an idiosyncratic emphasis. Figure 5 reports each judge's fitted total $\sum_k S_{i,k}$ against the human holistic score: rank agreement is high on both ELLIPSE (Spearman rank correlation $r_s = 0.89$–0.92) and ASAP 2.0 ($r_s = 0.77$–0.86). Each total decomposes exactly into named per-feature scores, so unlike a scalar Bradley–Terry fit, every verdict can be audited feature by feature: the property that motivated the extension.

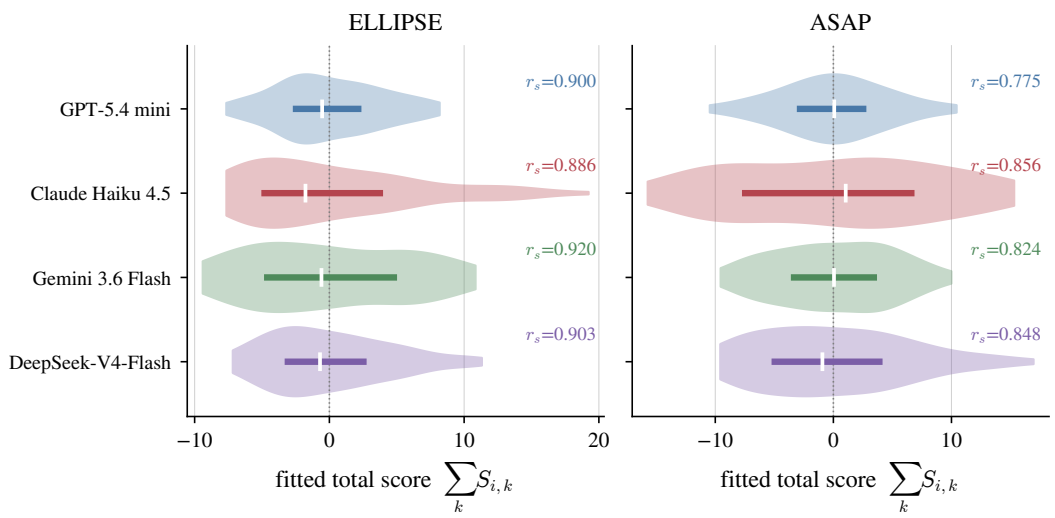

Figure 5: **Per-judge overall score distribution and its agreement (Spearman rank correlation) with human scores.**

A.1 Details of the extension

Data and comparison design. For each corpus we sample $n = 100$ essays, band-balanced on the human holistic score, and build 2,000 pairs as the union of 40 random perfect matchings, so every essay appears in exactly 40 comparisons. The essay sample, the pair list, and the presentation order are generated once and are byte-identical across the four judges of a corpus: within a corpus, only the judge varies. Between corpora, only the corpus framing of the elicitation prompt changes (ELLIPSE frames English-language proficiency of grade 8–12 learners; ASAP 2.0 frames source-based argumentative quality and inlines the source article and assignment verbatim); schema, field order, parser, and all downstream prompts are unchanged, so the ELLIPSE–ASAP 2.0 contrast is a corpus contrast, not a prompt contrast.

Elicitation. One call per pair elicits a forced binary verdict (Prompts 1 and 2) followed by the qualities the winner exhibits more than the loser, each with evidence, a short label, and a one-sentence description. The prompt names no features and contains no worked examples (to avoid priming the constructs under study), forbids bundled qualities ("and", "/", "&"), and rules out of scope the topic, the stance, length in itself, and the corpora's anonymization placeholders. The output schema emits the verdict *before* the qualities, matching the factorization of the likelihood (A.3). Reasoning modes are disabled on all four judges.

Canonicalization and reliability gates. Per judge, cited phrases with at least 3 citations are grouped by meaning into features by GPT-5.4 (HIGH REASONING), with a catch-all "Other" bucket dropped downstream; group membership is validated programmatically. The model of Equation (A.3) is fitted by joint MLE (L-BFGS, analytic gradients, ridge 10^{-3}), and each feature passes a reliability screen built on the same principle as Section 3: a condition number κ_k of the score block of the Fisher information of model (A.3), and a signal-to-noise ratio ρ_k obtained by subtracting the sandwich estimation variance from the observed score spread. The statistics differ from those of the main pipeline because the likelihood does; we retain features with $\rho > 0.50$ and $\kappa < 100$, then refit the model on the surviving features. Across the eight judge–corpus cells, the screen retains between 6 and 16 features.

Cross-judge consolidation. The four judges' reliable features are grouped into shared constructs by a single GPT-5.4 (HIGH REASONING) call per corpus, with at most one feature per judge and per construct. The prompt requires singletons to remain visible (a construct found by one judge is a finding, not noise) and receives no statistics so semantic grouping cannot be contaminated by how essays scored. Figure 4 displays constructs recovered by at least three judges.

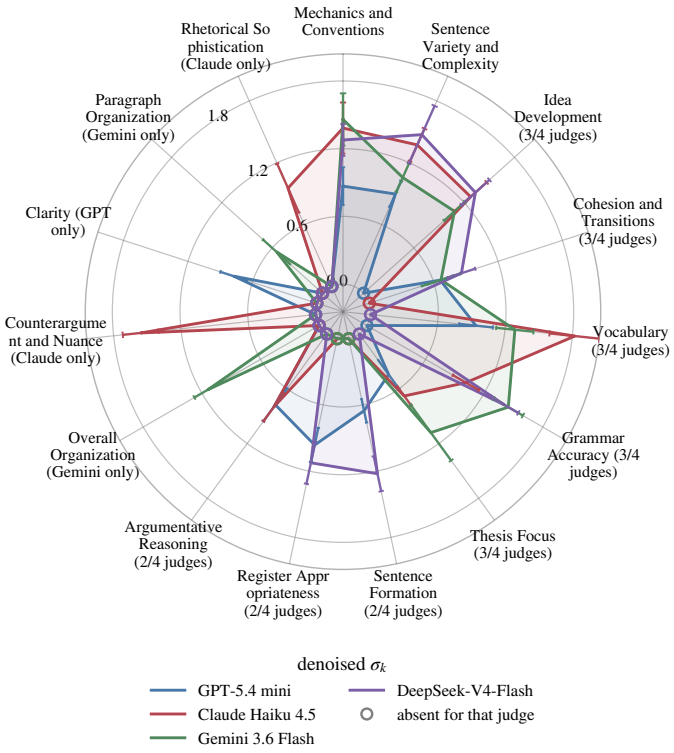

Figure 6: **Denoised per-feature spread by judge on ELLIPSE.** Axes are cross-judge constructs formed by meaning (GPT-5.4 (HIGH REASONING)); annotations give how many of the four judges discovered each construct. The radial coordinate is the denoised spread σ_k of each judge's feature mapped to that construct, on per-judge standardized reliable feature sets ($\rho > 0.5$, $\kappa < 100$); bars are 95% delta-method intervals from the sandwich covariance. An open marker denotes a construct absent from that judge's reliable basis (never surfaced, or removed by the identifiability screen), not a zero estimate.

Reported quantity and denoising. The radial axis reports a noise-corrected spread: the observed σ_k mixes essay-to-essay variation with estimation error, so we subtract the mean sandwich variance (Appendix B.3) of the fitted scores, with delta-method standard errors from the sandwich covariance. Denoising shrinks ELLIPSE spreads by 10.5% on average and reorders 10 of 34 features within their own judge.

Fit quality and position bias. Spearman correlation between the fitted total $\sum_k S_{i,k}$ and the human holistic score is 0.89–0.92 on ELLIPSE and 0.77–0.86 on ASAP 2.0. The fitted position coefficient β varies in sign and magnitude across judges ($|\hat{\beta}|$ up to 2.6) and tracks each judge's raw first-position win rate on both corpora: position bias is a genuine property that depends on the judge and on the dataset.

Sensitivity and design caveats. We verified the sensitivity of the reliability gates: at $\rho > 0.70, \kappa < 50$ the ELLIPSE reliable feature sets shrink from $9/9/9/7$ to $6/6/6/7$ features (GPT-5.4 MINI, CLAUDE HAIKU 4.5, GEMINI 3.6 FLASH, DEEPSEEK-V4-FLASH) and the constructs shared by three judges or more fall from 7 to 5, so consensus counts depend on the gates and we report the setting used. Because different judges lose different features to κ, the consensus counts *understate* agreement (e.g., Vocabulary is a $3/4$ construct on ELLIPSE only because the screen removed DeepSeek's). The citation floor interacts similarly: in a preliminary three-judge run, a floor of 5 let thinly spread phrasing make constructs appear less shared than they were, which motivated the minimum-citation threshold of 3.

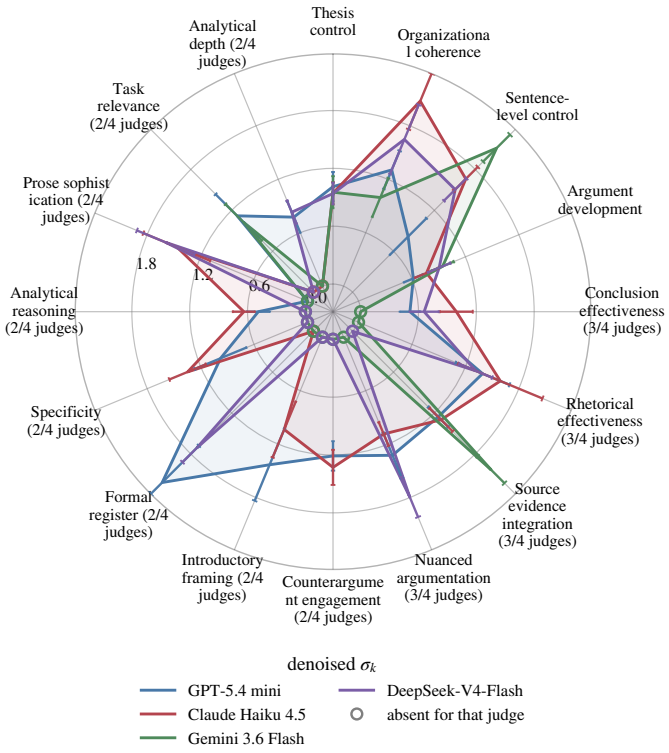

Figure 7: **Denoised per-feature spread by judge on ASAP 2.0.** Same construction, quantities, and conventions as Figure 6, on the ASAP 2.0 corpus.

Model calibration. The extension model assigns two probability forecasts of observable binary events, one per channel. The winner gate, $\mathbb{P}(w_p = 1) = \sigma\big(\sum_k \hat{\Delta}_{p,k} + \hat{\beta}\big)$, is the probability that the first-shown essay wins the overall verdict, scored once per comparison; the citation gate, $\mathbb{P}(c_{p,k} = 1 \mid w_p) = \sigma\big((2w_p - 1)\hat{\Delta}_{p,k} + \hat{\gamma}_k\big)$, is the probability that feature k is cited given the verdict, increasing in the margin by which k favors the declared winner, scored on every (comparison, feature) cell of the reliable feature set. We score both as forecasts, out of fold (5-fold cross-validation over the 2,000 comparisons, so that every prediction comes from parameters estimated without it, with 95% bootstrap intervals over comparisons) independently for each of the four judges and on both corpora, so calibration is also a robustness check across judges.

Figures 8 and 10 show reliability diagrams for ELLIPSE and ASAP 2.0: within ten equal-width bins of predicted probability, the observed event rate is plotted against the predicted one, so a perfectly calibrated gate lies on the diagonal. The expected calibration error (ECE) summarizes each panel as the mean absolute gap between the two, weighted by bin occupancy. The winner gate is calibrated for every judge: out-of-fold ECE ranges over 0.012–0.028 on ELLIPSE and 0.010–0.019 on ASAP 2.0, in no case worse than in sample. Its predictions concentrate near the endpoints of the unit interval, so the diagnostic is exercised where it is most demanding: the model is confident on most comparisons and right about how confident it is. The citation gate is accurate in sample (ECE 0.003–0.011) but slightly degrades out of fold (ECE 0.027–0.041), with the gap concentrated in the upper bins, where observed citation rates fall short of the predictions for all four judges and on both corpora.

Figure 9 reports discrimination through the Brier skill score against chance ($p = 1/2$, Brier 0.25) and against the empirical rate estimated on the training folds (the rate at which the first-shown essay wins for the winner gate, the per-feature citation rate for the citation gate). The second reference is the demanding one for the citation channel: it is the maximum-likelihood fit of the gap-independent submodel in which $\mathbb{P}(c_{p,k} = 1 \mid w_p) = \sigma(\gamma_k)$, citation as pure per-feature habit, so skill over it is a direct out-of-fold test of whether citations track the winner-oriented gaps.

The winner gate's skill scores against the two references coincide (slot balance keeps the first-shown rate near 1/2) and are large for every judge (0.80–0.84 on ELLIPSE, 0.72–0.85 on ASAP 2.0). The citation gate scores 0.03–0.23 against the habit-only submodel across judges and corpora, rejecting it throughout: citations do track the gaps, though the contribution is second-order next to the verdict channel and shrinks from its in-sample values in line with the reliability curves.

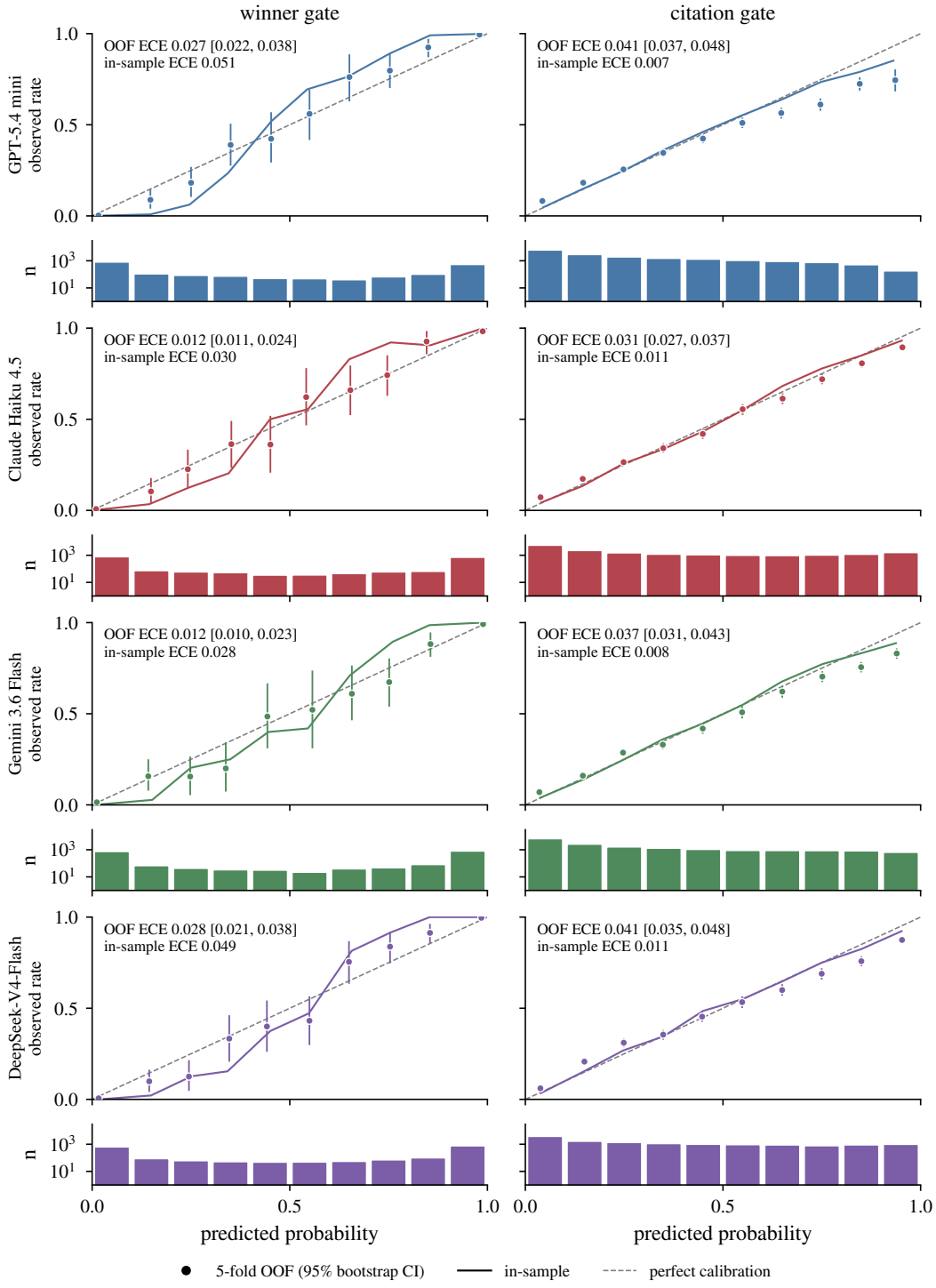

Figure 8: Out-of-fold reliability diagrams for the winner gate (left; unit: comparison) and the citation gate (right; unit: cell) of the model on ELLIPSE, one row per judge. Units are grouped into ten equal-width bins $B_1, \ldots, B_{10}$ of predicted probability $\hat{p}_i$; points show the observed event rate $\bar{y}_b = \frac{1}{|B_b|} \sum_{i \in B_b} y_i$ against the mean prediction $\bar{p}_b = \frac{1}{|B_b|} \sum_{i \in B_b} \hat{p}_i$, with 95% bootstrap intervals (resampling comparisons, resp. cells). Faint line: in-sample fit; dashed: perfect calibration ($\bar{y}_b = \bar{p}_b$). The expected calibration error is $\mathrm{ECE} = \sum_b \frac{|B_b|}{N} |\bar{y}_b - \bar{p}_b|$, with $N = \sum_b |B_b|$. Bottom strips: $|B_b|$ per bin, log scale.

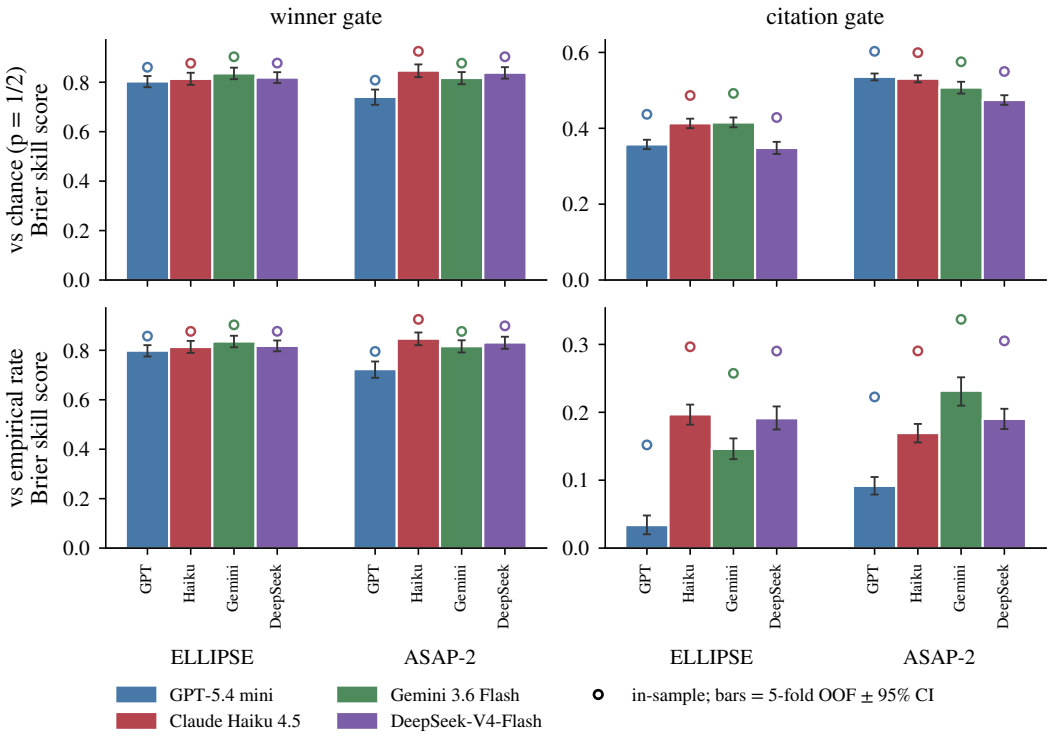

Figure 9: Out-of-fold Brier skill of the two gates, per judge and corpus, with in-sample values overlaid. With $\hat{p}_i$ the predicted probability and $y_i \in \{0, 1\}$ the gated event, the Brier score is $\mathrm{BS} = N^{-1} \sum_i (\hat{p}_i - y_i)^2$ and the skill score against a reference forecast q_i is $\mathrm{BSS} = 1 - \mathrm{BS}/\mathrm{BS}_{\mathrm{ref}}$, where $\mathrm{BS}_{\mathrm{ref}} = N^{-1} \sum_i (q_i - y_i)^2$; $\mathrm{BSS} = 1$ is perfect, 0 matches the reference, and negative values are worse than it. Two references: chance ($q_i = \frac{1}{2}$) and the empirical event rate of the training fold ($q_i = \bar{y}_{\mathrm{train}}$). Intervals as in Figure 8.

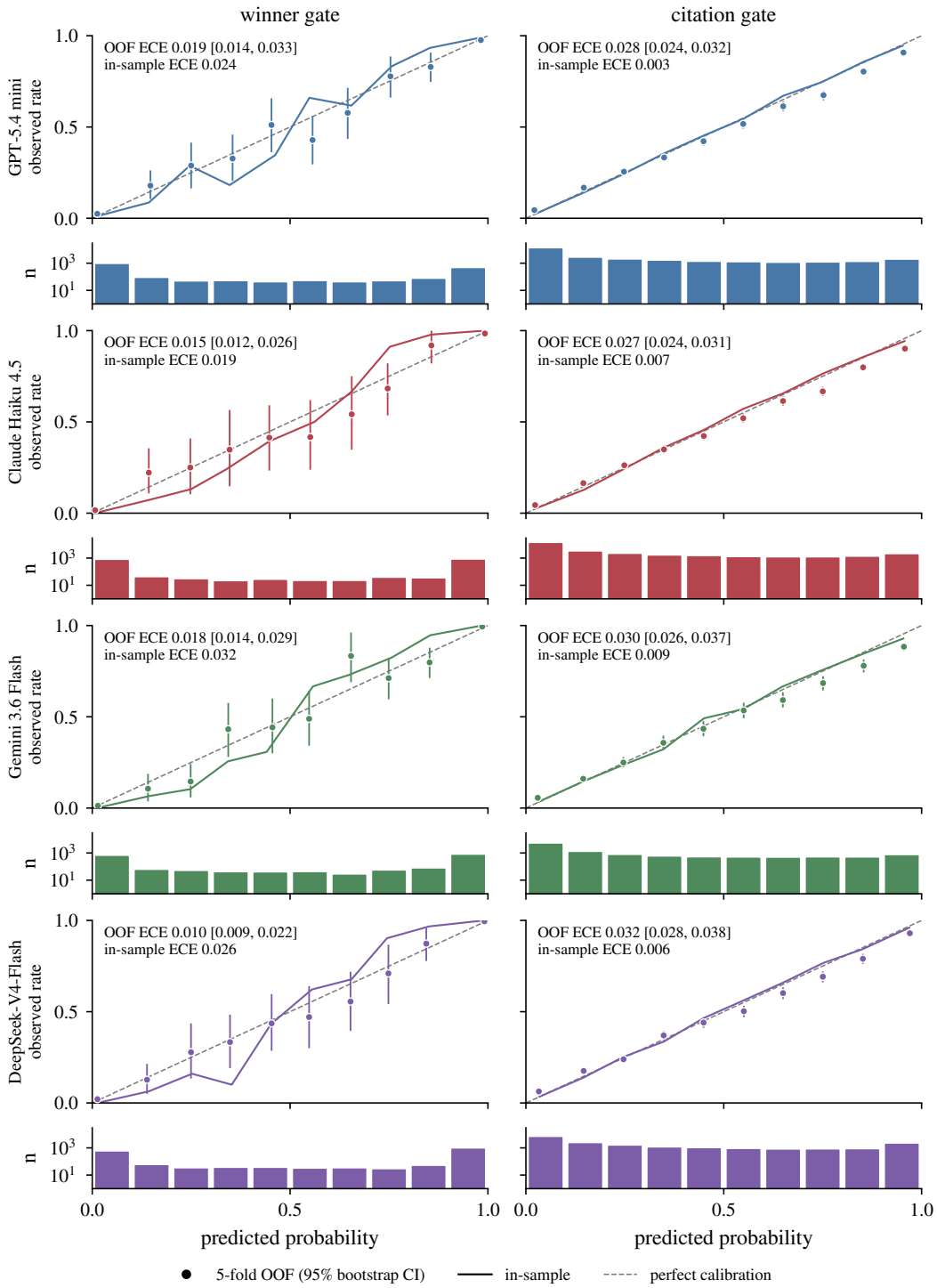

Figure 10: As Figure 8, on ASAP 2.0.

Prompts. The prompts used are reported below.

Prompt 1: Comparative elicitation – ELLIPSE.

```
=== SYSTEM ===
You are helping build an interpretable scoring framework by examining how student essays
differ in English language proficiency. The framework works by collecting many pairwise
comparisons: for each pair, a judgment of which essay is stronger overall is recorded
together with the qualities that essay has more of. Aggregated across many pairs, these
reveal the distinct dimensions along which student writing varies. There is no fixed rubric
--- the qualities are recovered from comparisons like this one, not specified in advance.

These are essays written by English language learners in grades 8 to 12, during
standardized writing assessments. They span a wide range of proficiency, from beginning to
quite advanced English writers. Each essay responds to an independent writing prompt that
requires no specialized background knowledge; different essays may respond to different
prompts. Read and weigh the essays as you would when assessing the English writing
proficiency of language learners at this level.

Your task. You will be shown two student essays, Essay A and Essay B. They may respond to
different writing prompts. This pair is one comparison in a larger collection.

First, decide which essay demonstrates stronger English language proficiency overall. This
is a forced choice: pick A or B even when the two are close.

Then identify the qualities that the essay you chose has more of than the other essay.
Report the ones that genuinely stand out to you on reading this pair --- bring up whatever
you find salient, and do not work from a predetermined list. Favor real, noticeable
differences over minor ones, but do not force the list to be short: report as many as
actually distinguish these two essays.

For each quality, first name the concrete, specific feature you actually saw in the essays
--- a short phrase pointing to the specific evidence. Then, grounded in that evidence, name
the quality and describe in one sentence what it captures.

Name only qualities on which the essay you chose is genuinely ahead. If the two essays are
comparable on some aspect of proficiency, leave it out rather than listing it. Do not pad
the list to justify your choice, and do not list a quality on which the other essay is in
fact stronger.

Out of scope --- do not treat these as differences in proficiency:
- the writing prompt an essay was responding to, or the topic or difficulty of that prompt
--- judge proficiency, not the task an essay happened to be assigned;
- the position or opinion an essay takes --- any stance can be expressed proficiently;
- length in itself;
- anonymization placeholders inserted by the corpus builders to remove identifying
information, such as Generic_Name, Generic_City, Generic_School, PROPER_NAME,
LOCATION_NAME, STUDENT_NAME, TEACHER_NAME or SCHOOL_NAME. Ignore them entirely: they are
not the student's words. Never count them as spelling, capitalization, spacing or
word-choice errors --- including when a placeholder is misspelled or run together with an
adjacent word (for example "Generic_Namehad" or "wasGeneric_Name").

How to report. List each quality as exactly one item --- do not combine qualities with
"and", "/", "&", or a comma; split related qualities into separate items. For the evidence
field, give a short phrase naming the concrete feature you saw. For the quality field, give
a concise general noun phrase naming the quality (the kind of short label you might use as
a rubric column). For the description field, give one sentence saying what that quality
captures.

Output only JSON of the form:
{"overall_winner": "A" | "B", "qualities": [{"evidence": "<short phrase>", "quality":
"<concise noun phrase>", "description": "<one sentence>"}, ...]}
=== USER ===
Essay A:
```

⟨*first essay*⟩

Essay B:
⟨*second essay*⟩

Following the standard above, first give the essay that demonstrates stronger English language proficiency overall, then list the qualities that essay has more of than the other --- for each, first the concrete feature you saw, then the quality and a one-sentence description. Respond with the JSON object only.

Prompt 2: Comparative elicitation – ASAP 2.0.

=== SYSTEM ===
You are helping build an interpretable scoring framework by examining how student argumentative essays differ in quality. The framework works by collecting many pairwise comparisons: for each pair, a judgment of which essay is stronger overall is recorded together with the qualities that essay has more of. Aggregated across many pairs, these reveal the distinct dimensions along which student writing varies. There is no fixed rubric --- the qualities are recovered from comparisons like this one, not specified in advance.

These are argumentative essays written by students in United States schools, primarily in grades 6, 8, 9 and 10. They span a wide range of writing ability and maturity, from beginning writers to quite proficient ones. Every essay responds to the same source-based argumentation task: the students read the article below and were asked to take and defend a position on it, using details from that article as evidence. Read and weigh the essays as you would when assessing argumentative writing at this level.

--- SOURCE ARTICLE ---
⟨*source article, inlined verbatim (1,120 words, "Driverless Cars Are Coming")*⟩

--- ASSIGNMENT GIVEN TO STUDENTS ---
In the article "Driverless Cars are Coming," the author presents both positive and negative aspects of driverless cars. Using details from the article, create an argument for or against the development of these cars. Be sure to include: your position on driverless cars; appropriate details from the article that support your position; an introduction, a body, and a conclusion to your argumentative essay.

Your task. You will be shown two student essays, Essay A and Essay B, both responding to this assignment. This pair is one comparison in a larger collection.

First, decide which essay is the stronger argumentative essay overall. This is a forced choice: pick A or B even when the two are close.

Then identify the qualities that the essay you chose has more of than the other essay. Report the ones that genuinely stand out to you on reading this pair --- bring up whatever you find salient, and do not work from a predetermined list. Favor real, noticeable differences over minor ones, but do not force the list to be short: report as many as actually distinguish these two essays.

For each quality, first name the concrete, specific feature you actually saw in the essays --- a short phrase pointing to the specific evidence. Then, grounded in that evidence, name the quality and describe in one sentence what it captures.

Name only qualities on which the essay you chose is genuinely ahead. If the two essays are comparable on some aspect of quality, leave it out rather than listing it. Do not pad the list to justify your choice, and do not list a quality on which the other essay is in fact stronger.

Out of scope --- do not treat these as differences in quality:
- the side an essay takes, for or against the development of driverless cars --- either position can be argued well;

- the topic itself, or which particular details from the article an essay chose to use, except insofar as those choices affect the quality of the argument;
- length in itself;
- anonymization placeholders inserted by the corpus builders to remove identifying information, such as PROPER_NAME, OTHER_PII, STUDENT_NAME, LOCATION_NAME or SCHOOL_NAME. Ignore them entirely: they are not the student's words. Never count them as spelling, capitalization, spacing or word-choice errors --- including when a placeholder is misspelled or run together with an adjacent word.

How to report. List each quality as exactly one item --- do not combine qualities with "and", "/", "&", or a comma; split related qualities into separate items. For the evidence field, give a short phrase naming the concrete feature you saw. For the quality field, give a concise general noun phrase naming the quality (the kind of short label you might use as a rubric column). For the description field, give one sentence saying what that quality captures.

Output only JSON of the form:
{"overall_winner": "A" | "B", "qualities": [{"evidence": "<short phrase>", "quality": "<concise noun phrase>", "description": "<one sentence>"}, ...]}
=== USER ===
Essay A:
⟨*first essay*⟩

Essay B:
⟨*second essay*⟩

Following the standard above, first give the essay that is the stronger argumentative essay overall, then list the qualities that essay has more of than the other --- for each, first the concrete feature you saw, then the quality and a one-sentence description. Respond with the JSON object only.

Prompt 3: Cross-judge consolidation prompt.

=== SYSTEM ===
You are a careful taxonomist of writing-assessment constructs. Output only valid JSON matching the requested schema.
=== USER ===
You are consolidating evaluation features discovered independently by several LLM judges into a set of shared constructs.

Each judge compared ⟨*corpus framing*⟩ and, with no fixed rubric, named the qualities that distinguished the stronger essay. Each judge's raw phrasings have already been canonicalized separately, so within a single judge the features below are already distinct from one another. Your job is to identify, ACROSS judges, which features denote the same underlying construct.

Below is the full list. Each entry has an INDEX in square brackets, the JUDGE that produced it, its NAME, and a one-sentence DEFINITION.

⟨*indexed feature list — one entry per line: [i] (judge) name: definition*⟩

Instructions:
- Group entries that denote the SAME underlying evaluative construct, judging by MEANING, not by surface wording. Different words for the same idea belong in one group; the same word used for different ideas belongs in different groups.
- Keep genuinely DISTINCT constructs separate, even when related. Two constructs a writing expert would score on separate rubric lines should be separate groups.
- A group may contain AT MOST ONE feature per judge. The features within a judge were already canonicalized, so two of them never denote the same construct; if you find yourself wanting to place two features from one judge together, they belong in different groups.

```
- A feature with no counterpart in any other judge forms a group of ONE. Do not absorb it
into a loosely related group to tidy up, and do not create a catch-all: a construct only
one judge discovered is a real finding and must stand alone.
- Do NOT try to balance the groups. There is no target number of groups and no expectation
that every group contains every judge. Groups of one, two, three, and four are all equally
valid outcomes; forcing coverage would fabricate agreement that is not there.
- Do NOT merge two constructs just because they often co-occur or because strong essays
tend to score well on both. Group only by shared MEANING.
- The judge labels are present only so you can respect the one-per-judge rule. They carry
no information about meaning; never group or split on the basis of which judge produced an
entry.

For each group, give:
- a canonical NAME: a concise noun phrase naming the single construct the group captures.
Name the construct itself, not the union of its members' wordings; avoid conjunctions
unless the construct genuinely has no single name.
- a one-sentence DEFINITION of that construct.
- the list of members, each as its INDEX, its JUDGE, and its NAME copied verbatim.

Rules on membership (these will be checked programmatically):
- Every input index must appear in exactly one group, exactly once. Place all of them; drop
none; duplicate none.
- Do not invent indices that are not in the input.
- Each member's "judge" and "name" must be the exact input values at that index --- this is
a cross-check, so copy them from the bracketed entry, do not paraphrase.
- No group may contain two members with the same judge.

Output only JSON:
{"clusters": [{"name": "<canonical noun phrase>", "definition": "<one sentence>",
"members": [{"index": <int>, "judge": "<verbatim>", "name": "<verbatim>"}, ...]}, ...]}
```

B Statistical Model Theory

B.1 The Bradley–Terry with ties literature

Classical tie models. The Bradley–Terry model (Bradley and Terry, 1952) assigns each item a latent merit and models a binary win/loss outcome through the logistic link. Classical extensions accommodate a third outcome per comparison: Glenn and David (1960) add a discrimination threshold to the Thurstone–Mosteller model, Rao and Kupper (1967) do the same on the logistic scale, and Davidson (1970) models the tie probability as proportional to the geometric mean of the two merits, governed by a tie intensity ν. Davidson and Beaver (1977) further incorporate multiplicative within-pair order effects, the classical ancestor of our position coefficient β; see Cattelan (2012) for a survey. Since each of our per-feature observations is trinomial ("i wins", "j wins", "not cited"), our model is naturally read against this literature, and this subsection records the relationship.

Davidson's model is a gated model. Write $\pi_i = e^{s_i}$. Davidson's model puts $\mathbb{P}(i \text{ wins}) \propto e^{s_i}$, $\mathbb{P}(j \text{ wins}) \propto e^{s_j}$, and $\mathbb{P}(\text{tie}) \propto \nu\, e^{(s_i+s_j)/2}$. Dividing the three weights by $e^{(s_i+s_j)/2}$ and writing $\Delta = s_i - s_j$,

$$\mathbb{P}(\text{decisive}) = \frac{2\cosh(\Delta/2)}{2\cosh(\Delta/2)+\nu} = \sigma\Big(\log\big(2\cosh(\Delta/2)\big) - \log\nu\Big), \qquad \mathbb{P}(i \mid \text{decisive}) = \sigma(\Delta)$$

Davidson's model is exactly the single-feature instance of (B.8) with gate link $f(u) = \log(2\cosh(u/2))$, gate intercept $\gamma = -\log\nu$, and no position effect. Our link $f = \log(2\cosh)$ is a variant of the same choice, and the general model of Appendix B.5.1 is a strict generalization of Davidson's. The Rao–Kupper model admits no such factorization: it is the ordered logit with symmetric cutpoints, $\mathbb{P}(i \succ j) = \sigma(\Delta - \log\theta)$ and $\mathbb{P}(i \succ j \text{ or tie}) = \sigma(\Delta + \log\theta)$, so the threshold couples the two channels and the direction given a decisive outcome is not of logistic form in Δ.

Why tie models do not transfer off the shelf. Since Davidson's model is a gated model, a battery of K independent Davidson fits, one per feature, lies inside our model class ($\gamma_k = -\log\nu_k$) and is the natural classical baseline. Three things separate our model from it.

First, the regime and the semantics of the null outcome: the tie literature was developed where ties are a minority outcome among near-equal players and assert near-equality, whereas non-citation is our dominant outcome and signals lack of salience as much as similarity. The fitted intercepts of Table 4 correspond to tie intensities $\nu_k = e^{-\gamma_k}$ ranging over two orders of magnitude, which is to our knowledge far outside anything that literature contemplates.

Second, the position effect. The classical route (Davidson and Beaver, 1977) multiplies the first-presented merit, $\pi_i \mapsto e^{\beta}\pi_i$, leaving the tie weight $\nu\sqrt{\pi_i\pi_j}$ unchanged; the model stays log-linear, with

$$\mathbb{P}(\text{decisive}) = \sigma\Big(\log 2\cosh\big(\tfrac{\Delta+\beta}{2}\big) + \tfrac{\beta}{2} - \log\nu\Big), \qquad \mathbb{P}(i \mid \text{decisive}) = \sigma(\Delta+\beta).$$

The conditional direction is exactly our direction channel, but the gate argument is no longer even in Δ: presentation order alters the citation probability, and the null outcome is most likely at $\Delta = -\beta$ rather than at a genuine tie. Attaching the order effect symmetrically to both merits changes only the constant $\beta/2$, absorbed into ν; the non-evenness remains. Our parity structure (even f and $\boldsymbol{Y}_{ijk}$, odd $\boldsymbol{Z}_{ijk}$ defined in Appendix B.5) instead keeps the direction channel $\sigma(\Delta+\beta)$ while confining the position effect to it by construction.

The trade-off is exact: an order effect on the merits is what keeps the trinomial model log-linear, so an order-invariant gate combined with a position-biased direction channel forces the likelihood *outside* the multinomial-logit family – which is precisely where concavity is lost (the signed curvature term of (B.15)) and why the estimation theory of Appendices B.9–B.12 cannot be imported from that literature.

Third, the K features are estimated jointly with a shared position coefficient, so the joint likelihood is not a product of per-feature tie models, both channels carry general covariates, and the inferential targets of Section 3 – per-feature Laplacian condition numbers, reliability coefficients, and misspecification-robust intervals – have no counterpart in the paired-comparison ties literature.

A discrete-choice reading, and scope beyond LLM judges. The gate/direction factorization can also be read as a two-level discrete choice with an outside option: the judge first "chooses" whether the feature is worth mentioning, then chooses a winner, analogous to a nested logit in which non-citation is the outside nest (McFadden, 1978; Börsch-Supan, 1990). It is by this analogy that we elaborated the model in the first place. We do not claim a formal random-utility foundation – our gate replaces the inclusive value by the even link f – but the analogy delimits the scope of the framework: it applies to any rater who, per comparison, reports verdicts only on the attributes salient enough to mention.

Estimation theory. Modern non-asymptotic theory for the binary Bradley–Terry model studies the growing-dimension regime: Simons and Yao (1999) establish consistency and asymptotic normality as $n \to \infty$ in the dense case, Negahban et al. (2017) and Shah et al. (2016) analyze spectral and minimax aspects, Chen et al. (2019) obtain ℓ_∞ guarantees through the leave-one-out technique we adapt, and Yan (2025) provides the uniform-consistency template our Stage-1 analysis builds on. For tie models, we are not aware of comparable growing-dimension ℓ_∞ guarantees. The technical landscape also differs: Davidson's model is a log-linear multinomial logit and Rao–Kupper an ordered logit, so both have concave log-likelihoods in the natural parametrization, whereas the free even link and the gate covariates make our likelihood non-concave – the signed curvature term of (B.15) – which is the difficulty the machinery of Appendices B.9–B.12 is built to handle.

B.2 Why a shared position bias coefficient?

The direction channel of the model includes a position coefficient β: for a pair p on feature k, given citation, the probability that the first-presented item wins is $d_{p,k} = \sigma(\Delta_{p,k} + \beta)$, where pairs are written in presentation order, so that $\Delta_{p,k} = s_{i,k} - s_{j,k}$ with i the item shown first. The coefficient β is shared across features. A natural alternative is a per-feature coefficient β_k; we first explain why this choice has almost no effect on the estimated scores, then verify it empirically.

Setup: how β enters the score equations. Fix a feature k (we drop it from the notation). The direction log-likelihood is

$$\ell(\boldsymbol{s}, \beta) \;=\; \sum_{p\,:\,c_p=1} \Big[\, w_p \log \sigma(\Delta_p + \beta) + (1 - w_p) \log\big(1 - \sigma(\Delta_p + \beta)\big)\Big],$$

where $w_p = 1$ if the first-presented item of pair p won. Consider an item i, and let P_i be the pairs containing i. In each $p \in P_i$, the gap Δ_p contains s_i with sign

$$\varepsilon_p(i) \;=\; \begin{cases} +1 & \text{if } i \text{ was presented first in } p, \\ -1 & \text{if } i \text{ was presented second,} \end{cases}$$

The score equation for s_i is therefore

$$\frac{\partial \ell}{\partial s_i} = \sum_{p \in P_i} \varepsilon_p(i)\big(w_p - d_p\big) \;=\; 0, \qquad d_p = \sigma(\Delta_p + \beta). \tag{B.1}$$

This is the standard Bradley–Terry balance condition (observed wins minus expected wins, counted from i's side of each pair) except that every expected win d_p is tilted by the common offset β.

First-order effect of changing β. Suppose β moves by an amount δ (for instance, from the shared estimate $\hat\beta$ to a per-feature estimate $\hat\beta_k$, so $\delta = \hat\beta_k - \hat\beta$). Each expected win moves by $\mathrm{d}d_p = w_{\theta,p}\delta$ with $w_{\theta,p} = d_p(1 - d_p)$, so the score equation (B.1) is thrown off balance by

$$-\,\delta \sum_{p \in P_i} \varepsilon_p(i)\, w_{\theta,p}.$$

To restore balance, s_i must adjust. Differentiating (B.1) in s_i (and neglecting the adjustment of the other items' scores, which enters only through the off-diagonal Laplacian terms), the required adjustment is, to first order,

$$\Delta \hat s_i \approx -\delta \frac{\displaystyle\sum_{p \in P_i} \varepsilon_p(i)\, w_{\theta,p}}{\displaystyle\sum_{p \in P_i} w_{\theta,p}}. \tag{B.2}$$

We write $\approx$ because modifying β also modifies the other s_j, which in turn modify s_i: we neglect these second-order perturbations. The denominator is a sum of positive weights; the numerator is the *same* weights, but signed by presentation order. The score shift is thus δ times a weighted average of ± 1's: $+1$ for each pair where i was shown first, -1 where it was shown second.

Cancellation under the balanced design. Equation (B.2) exposes exactly what the design controls. If the weights were equal, the numerator would be $n_{\text{first}}(i) - n_{\text{second}}(i)$: the imbalance in how often item i appeared in each presentation slot. Our comparison graph assigns presentation order uniformly at random, independently of the items; with m comparisons per item, each item is first $\text{Binomial}(m, \frac{1}{2})$ times, so the slot fraction deviates from $\frac{1}{2}$ by $O_p(m^{-1/2})$ (for $m = 40$: standard deviation 0.079, matching the realized imbalance). Hence $\Delta\hat{s}_i = O_p(\delta/\sqrt{m})$: the effect of any change in β on any score is reduced by the square root of the number of comparisons, *because presentation order is randomized*.

Two remarks. First, the cancellation is not exact: under randomization the two presentation orders of a pair carry weights $\sigma'(\Delta+\beta)$ and $\sigma'(\Delta-\beta)$, whose signed average is $\frac{1}{2}(\sigma'(\Delta+\beta)-\sigma'(\Delta-\beta)) = \beta\,\sigma''(\Delta) + O(\beta^3)$; the numerator of (B.2) thus retains a deterministic residual of order $O(\beta)$ per pair, on top of the $O_p(m^{-1/2})$ stochastic imbalance. Both are negligible here ($\hat{\beta}$ and $|\sigma''|$ are small), but the orthogonality is first-order, not exact. Second, the citation channel contains no position term at all (the gate is $\sigma(f(\Delta) + \gamma_k)$), so the information that citations carry about score magnitudes is untouched by β under either specification.

Empirical check. When we refit the data of Section 4 with per-feature β_k, the positional bias estimates differ significantly across features, yet the scores are *essentially unchanged*: the median absolute shift is 1% of the per-feature score spread (90th percentile: 5%). The residual movement is attributable to the imperfect slot balance: across features, the score shift correlates with $|\hat{\beta}_k|$ at $r = 0.87$, as the cancellation argument predicts.

We therefore retain the shared β in the main model; the per-feature $\hat{\beta}_k$ remain interpretable as an audit of the judge (features differ in their susceptibility to position bias) but the design prevents this heterogeneity from reaching the scores.

B.3 Uncertainty quantification

B.3.1 Confidence intervals

The estimator $\hat{\boldsymbol{\theta}}$ is an M-estimator, which lets us adopt the weaker, standard view from robust statistics (Huber, 1967; White, 1982) to account for potential model misspecifications in our estimates. $\hat{\boldsymbol{\theta}}$ estimates the *pseudo-true* parameter $\boldsymbol{\theta}^\star$ (which is the Kullback–Leibler projection of the judge's actual comparison distribution onto our model class). In the regime where the item set is held fixed and the number N of edges of the comparison graph grows, the usual M-estimation argument yields

$$\sqrt{N}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\big) \xrightarrow[N\to\infty]{d} \mathcal{N}\big(0, \boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}\big), \qquad \begin{aligned} \boldsymbol{H} &= \mathbb{E}\big[\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^\star)\big], \\ \boldsymbol{G} &= \mathbb{E}\big[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^\star)\,\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^\star)^\top\big], \end{aligned} \tag{B.3}$$

estimated by their empirical counterparts $\widehat{\boldsymbol{H}}$ (average Hessian) and $\widehat{\boldsymbol{G}}$ (average outer product of per-comparison gradients) at $\hat{\boldsymbol{\theta}}$. The resulting *sandwich* covariance $\widehat{\boldsymbol{\Sigma}}^{\text{sand}} = \frac{1}{N}\widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{G}}\widehat{\boldsymbol{H}}^{-1}$ coincides with the inverse Fisher information when the model is correct (by the information-matrix equality $\boldsymbol{G} = \boldsymbol{H}$), and remains an asymptotically valid covariance for the projection $\boldsymbol{\theta}^\star$ when it is not. Hence, the confidence intervals on scores are statements about this projection, namely the score configuration that best summarizes the judge's comparisons within our model class.

B.4 Reliability diagnostics

Since citation is sporadic, there may be features that are not cited enough or whose winners are inconsistent so that the scores are not precisely pinned down by the data. One advantage of our method is that it rests on a statistical model: we can hence inspect the *condition numbers* of blocks of the Hessian to diagnose poorly identified features (Rothenberg, 1971). Indeed, for each feature k,

the unpenalized score block of the Fisher information is a weighted Laplacian $\boldsymbol{L}_k$. By computing:

$$\kappa_k = \frac{\lambda_{\max}(\boldsymbol{L}_k)}{\lambda_2(\boldsymbol{L}_k)} \tag{B.4}$$

and keeping features that have $\kappa \leq \kappa_{\max}$ (where $\kappa_{\max}$ is specified a priori) we can filter out poorly identified features. Features that have $\kappa > \kappa_{\max}$ are called *unidentified*.

Moreover, an interesting quantity to compute is the spread of the scores, as it tells us how markedly items differ along each feature in the dataset. However, a high spread could also be due to noisy estimates. We can hence denoise the estimated spreads of the scores by subtracting the analytic estimation variance:

$$\nu_k := \frac{1}{n} \operatorname{Tr}\left(\widehat{\boldsymbol{\Sigma}}_k^{\text{sand}}\right) \tag{B.5}$$

where $\widehat{\boldsymbol{\Sigma}}_k^{\text{sand}}$ is the k-th score block of the sandwich covariance. Thus $\rho_k := \frac{\operatorname{Var}(\hat{\boldsymbol{s}}_k) - \nu_k}{\operatorname{Var}(\hat{\boldsymbol{s}}_k)}$ is the fraction of observed score spread attributable to genuine between-item signal. Keeping features with $\rho_k > \rho_{\min}$ (where $\rho_{\min}$ is specified a priori) filters out *unreliable* features.

Having fitted the model a first time, spotted unidentified and unreliable features, and possibly chosen features that we want to drop or fuse, we can then refit the model and obtain scores on reliable (and chosen) features.

B.5 Model, likelihood, score, and Hessian

B.5.1 Model

We study the model where there is only one feature (case $K = 1$). There are $n + 1$ items indexed by $\{0, 1, \ldots, n\}$. For each unordered pair $\{i, j\}$ with $0 \leq i < j \leq n$ we observe m_{ij} comparisons, indexed $k = 1, \ldots, m_{ij}$. Each comparison yields a trinomial outcome

$$r_{ijk} \in \{0, +1, -1\} : \qquad 0 = \text{“pair not cited”}, \quad +1 = \text{“}i\text{ wins”}, \quad -1 = \text{“}j\text{ wins”}.$$

We encode it by a *citation* indicator and a *direction* indicator,

$$c_{ijk} := \mathbf{1}_{\{r_{ijk} \neq 0\}} \in \{0, 1\}, \qquad a^1_{ijk} := \mathbf{1}_{\{r_{ijk} = +1\}}, \quad a^{-1}_{ijk} := \mathbf{1}_{\{r_{ijk} = -1\}},$$

so that $a^1_{ijk} + a^{-1}_{ijk} = c_{ijk}$.

Each item i carries a latent merit $s_i \in \mathbb{R}$; we write $\boldsymbol{s} = (s_0, \ldots, s_n)^\top \in \mathbb{R}^{n+1}$ for the merit vector (suppressing the dependence on n when clear) and $\Delta_{ij} := s_i - s_j$ for the gap. There are two fixed-dimension structural parameters: the gate coefficient $\boldsymbol{\gamma} \in \mathbb{R}^{p_\gamma}$ and the direction (position-bias) coefficient $\boldsymbol{\beta} \in \mathbb{R}^{p_\beta}$. Each comparison carries covariates $\boldsymbol{Y}_{ijk} \in \mathbb{R}^{p_\gamma}$ (gate) and $\boldsymbol{Z}_{ijk} \in \mathbb{R}^{p_\beta}$ (direction), with parities

$$\boldsymbol{Y}_{jik} = \boldsymbol{Y}_{ijk} \text{ (even)}, \qquad \boldsymbol{Z}_{jik} = -\boldsymbol{Z}_{ijk} \text{ (odd)}, \qquad \Delta_{ji} = -\Delta_{ij}. \tag{B.6}$$

Let $\sigma(x) = e^x/(1 + e^x)$ and let $f : \mathbb{R} \to \mathbb{R}$ be a fixed *even* link. Define the two linear predictors and their sigmoids,

$$\eta_{ijk} := f(\Delta_{ij}) + \boldsymbol{Y}_{ijk}^\top \boldsymbol{\gamma}, \quad g_{ijk} := \sigma(\eta_{ijk}); \qquad \theta_{ijk} := \Delta_{ij} + \boldsymbol{Z}_{ijk}^\top \boldsymbol{\beta}, \quad d_{ijk} := \sigma(\theta_{ijk}). \tag{B.7}$$

The model is, conditionally independently across all (i, j, k),

$$\mathbb{P}(r_{ijk} = 0) = 1 - g_{ijk}, \quad \mathbb{P}(r_{ijk} = +1) = g_{ijk}\, d_{ijk}, \quad \mathbb{P}(r_{ijk} = -1) = g_{ijk}\,(1 - d_{ijk}). \tag{B.8}$$

Equivalently $c_{ijk} \sim \text{Bernoulli}(g_{ijk})$, and conditionally on $c_{ijk} = 1$, $a^1_{ijk} \sim \text{Bernoulli}(d_{ijk})$. The parities (B.6) make (B.8) invariant under $i \leftrightarrow j$ (the gate is invariant; the direction flips), since f is even while Δ and $\boldsymbol{Z}$ are odd.

Identification (the $\mathbf{1}^\perp$ gauge). Adding a constant to every s_i leaves (B.8) unchanged (only gaps enter), so we impose the gauge $\boldsymbol{s} \in \mathbf{1}^\perp := \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \mathbf{1}^\top \boldsymbol{x} = 0\}$ throughout. This choice is natural here: every merit-direction score and Hessian block below is a linear combination of the centered vectors $\boldsymbol{\omega}_{ij} := \boldsymbol{e}_i - \boldsymbol{e}_j \in \mathbf{1}^\perp$, so the gauge is preserved automatically by the estimating equations and by the gradient iterations of Appendix B.11, with no explicit projection.

B.5.2 LOG-LIKELIHOOD

From (B.8), the per-comparison log-likelihood is

$$\ell_{ijk} = c_{ijk} \log g_{ijk} + (1 - c_{ijk}) \log(1 - g_{ijk}) + a^1_{ijk} \log d_{ijk} + a^{-1}_{ijk} \log(1 - d_{ijk}),$$

and the total is

$$\ell(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{0 \le i < j \le n} \sum_{k=1}^{m_{ij}} \ell_{ijk}. \tag{B.9}$$

B.5.3 SCORES

Using $\sigma' = \sigma(1 - \sigma)$ and the chain rule through (B.7) ($\frac{\partial \eta_{ijk}}{\partial \Delta_{ij}} = f'(\Delta_{ij})$, $\frac{\partial \theta_{ijk}}{\partial \Delta_{ij}} = 1$), the per-comparison scores are

$$\frac{\partial \ell_{ijk}}{\partial \eta_{ijk}} = c_{ijk} - g_{ijk}, \qquad \frac{\partial \ell_{ijk}}{\partial \theta_{ijk}} = a^1_{ijk} - c_{ijk} d_{ijk}.$$

The three score blocks are

$$\nabla_{\boldsymbol{s}} \ell = \sum_{i<j} \sum_{k} \Big[(c_{ijk} - g_{ijk})\, f'(\Delta_{ij}) + \big(a^1_{ijk} - c_{ijk} d_{ijk}\big) \Big]\, \boldsymbol{\omega}_{ij} \ \in \mathbb{R}^{n+1}, \tag{B.10}$$

$$\nabla_{\boldsymbol{\gamma}} \ell = \sum_{i<j} \sum_{k} (c_{ijk} - g_{ijk})\, \boldsymbol{Y}_{ijk} \ \in \mathbb{R}^{p_\gamma}, \tag{B.11}$$

$$\nabla_{\boldsymbol{\beta}} \ell = \sum_{i<j} \sum_{k} \big(a^1_{ijk} - c_{ijk} d_{ijk}\big)\, \boldsymbol{Z}_{ijk} \ \in \mathbb{R}^{p_\beta}. \tag{B.12}$$

Define the per-coordinate merit score

$$G_a(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) := -\frac{\partial \ell}{\partial s_a} = -\sum_{j \ne a} \sum_{k} \psi_{ajk}, \qquad \psi_{ajk} := (c_{ajk} - g_{ajk}) f'(\Delta_{aj}) + \big(a^1_{ajk} - c_{ajk} d_{ajk}\big). \tag{B.13}$$

Collecting the coordinates, the *merit score map* is

$$\boldsymbol{G}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) := \nabla_{\boldsymbol{s}}(-\ell) = (G_0, \ldots, G_n)^\top = -\sum_{i<j} \sum_{k} \Big[(c_{ijk} - g_{ijk}) f'(\Delta_{ij}) + \big(a^1_{ijk} - c_{ijk} d_{ijk}\big) \Big] \boldsymbol{\omega}_{ij}, \tag{B.14}$$

which takes values in $\mathbf{1}^\perp$ because every $\boldsymbol{\omega}_{ij} = \boldsymbol{e}_i - \boldsymbol{e}_j$ is centered. When $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ are held fixed we write $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) := \boldsymbol{G}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta})$, viewed as a map $\mathbb{R}^{n+1} \to \mathbf{1}^\perp \subset \mathbb{R}^{n+1}$. The MLE solves $\boldsymbol{G}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathbf{0}$ on $\mathbf{1}^\perp$ together with (B.11) and (B.12).

B.5.4 HESSIAN

Set $w_{\eta,ijk} := g_{ijk}(1 - g_{ijk})$ and $w_{\theta,ijk} := d_{ijk}(1 - d_{ijk})$. Differentiating once more (θ is linear so $\partial^2 \theta = 0$, while $\frac{\partial^2 \eta}{\partial \Delta^2} = f''$) gives the per-comparison merit curvature

$$W_{ijk} := -\frac{\partial^2 \ell_{ijk}}{\partial \Delta^2} = \underbrace{w_{\eta,ijk}\, f'(\Delta_{ij})^2 + c_{ijk}\, w_{\theta,ijk}}_{\text{PSD part}} \ - \ \underbrace{(c_{ijk} - g_{ijk})\, f''(\Delta_{ij})}_{\text{“saboteur”}}, \tag{B.15}$$

and the full Hessian of $-\ell$ has the block structure (all sums $\sum_{i<j} \sum_k$)

$$\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) := \nabla^2_{\boldsymbol{s}}(-\ell) = \sum W_{ijk}\, \boldsymbol{\omega}_{ij} \boldsymbol{\omega}_{ij}^\top \qquad \text{(full merit block, } (n{+}1) \times (n{+}1)\text{)}, \tag{B.16}$$

$$\boldsymbol{A}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} := \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(-\ell) = \sum w_{\eta,ijk}\, \boldsymbol{Y}_{ijk} \boldsymbol{Y}_{ijk}^\top, \tag{B.17}$$

$$\boldsymbol{A}_{\boldsymbol{\beta}\boldsymbol{\beta}} := \nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}}(-\ell) = \sum c_{ijk}\, w_{\theta,ijk}\, \boldsymbol{Z}_{ijk} \boldsymbol{Z}_{ijk}^\top, \tag{B.18}$$

$$\boldsymbol{B}_{\boldsymbol{\gamma}} := \nabla^2_{\boldsymbol{s}\boldsymbol{\gamma}}(-\ell) = \sum w_{\eta,ijk}\, f'(\Delta_{ij})\, \boldsymbol{\omega}_{ij} \boldsymbol{Y}_{ijk}^\top, \tag{B.19}$$

$$\boldsymbol{B}_{\boldsymbol{\beta}} := \nabla^2_{\boldsymbol{s}\boldsymbol{\beta}}(-\ell) = \sum c_{ijk}\, w_{\theta,ijk}\, \boldsymbol{\omega}_{ij} \boldsymbol{Z}_{ijk}^\top, \tag{B.20}$$

$$\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\beta}}(-\ell) = \mathbf{0}. \tag{B.21}$$

Blocks (B.17)–(B.18) are positive semidefinite; the cross block (B.21) vanishes because, given the predictors, the two channels share no randomness. The merit block (B.16) is *not* PSD pathwise: the saboteur in (B.15) is signed, so a single W_{ijk} can be negative, e.g. on a cited near-tie, where $f' \approx 0$ and $c = 1$, $W \approx w_\theta - (1-g)f''$. This non-convexity is the central difficulty. The two off-diagonal blocks $\boldsymbol{B}_{\boldsymbol{\gamma}}, \boldsymbol{B}_{\boldsymbol{\beta}}$ are named here for later use (Lemma 8 and Corollary 1).

B.6 Notation and external tools

Notation. Following the template conventions, bold lowercase letters denote vectors ($\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{\omega}_{ij}$) and bold uppercase letters denote matrices ($\boldsymbol{J}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{V}, \boldsymbol{\Sigma}$); in keeping with the Bradley–Terry literature we also write in bold uppercase the per-comparison covariate vectors $\boldsymbol{Y}_{ijk} \in \mathbb{R}^{p_\gamma}$, $\boldsymbol{Z}_{ijk} \in \mathbb{R}^{p_\beta}$ and the vector-valued score maps $\boldsymbol{G}, \boldsymbol{D}, \boldsymbol{T}$. Calligraphic letters denote balls ($\mathcal{B}$), events ($\mathcal{E}$), σ-algebras ($\mathcal{F}$) and nets ($\mathcal{N}$). Starred quantities ($\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*, g^*_{ijk}, \dots$) are evaluated at the truth. Universal constants may change from line to line.

Norms. For $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_\infty = \max_a |x_a|$ and $\|\boldsymbol{x}\|_2 = (\sum_a x_a^2)^{1/2}$. For a matrix $\boldsymbol{A} = (A_{ab})$ we write $\|\boldsymbol{A}\|_{\max} = \max_{a,b} |A_{ab}|$ (entrywise) and use triple bars for the induced operator norms: $|||\boldsymbol{A}|||_2$ is the spectral norm (induced by $\|\cdot\|_2$) and $|||\boldsymbol{A}|||_\infty = \max_a \sum_b |A_{ab}|$ is the maximum absolute row sum (induced by $\|\cdot\|_\infty$). For symmetric $\boldsymbol{A}$, $|||\boldsymbol{A}|||_2 \le |||\boldsymbol{A}|||_\infty$. For symmetric $\boldsymbol{A}$ acting on the centered subspace, $\lambda_{\min,\perp}(\boldsymbol{A})$ is the smallest eigenvalue restricted to $\mathbf{1}^\perp = \{\boldsymbol{x} : \mathbf{1}^\top \boldsymbol{x} = 0\}$.

Restricted spectral norm. For a symmetric matrix $\boldsymbol{M}$ that maps $\mathbf{1}^\perp$ into itself, write

$$|||\boldsymbol{M}|||_{2,\perp} := \sup_{\boldsymbol{x} \in \mathbf{1}^\perp,\ \|\boldsymbol{x}\|_2 = 1} \|\boldsymbol{M}\boldsymbol{x}\|_2$$

for its spectral norm restricted to the centered subspace. In particular, if $\boldsymbol{M} = \boldsymbol{I} - \eta \boldsymbol{N}$ with $\boldsymbol{N}$ symmetric and PSD on $\mathbf{1}^\perp$ and $0 \le \eta\lambda \le 1$ for every eigenvalue λ of $\boldsymbol{N}|_{\mathbf{1}^\perp}$, then

$$|||\boldsymbol{I} - \eta \boldsymbol{N}|||_{2,\perp} = 1 - \eta\, \lambda_{\min,\perp}(\boldsymbol{N}). \tag{B.22}$$

Balls. For $\boldsymbol{x} \in \mathbb{R}^d$ and $r > 0$, $\mathcal{B}_2(\boldsymbol{x}, r)$ and $\mathcal{B}_\infty(\boldsymbol{x}, r)$ denote the *open* ℓ_2 and ℓ_∞ balls of radius r centered at $\boldsymbol{x}$, and $\overline{\mathcal{B}_2}, \overline{\mathcal{B}_\infty}$ their closures. For the fixed-dimensional structural parameters we write $\mathcal{B}(\boldsymbol{x}, r)$ for the open ℓ_2 ball.

Sigmoid bounds. For all x: $0 < \sigma'(x) = \sigma(1-\sigma) \le \frac{1}{4}$, $|\sigma''| \le \frac{1}{4}$, and $\frac{d}{dx}[\sigma(1-\sigma)] = \sigma(1-\sigma)(1-2\sigma)$ has absolute value $\le \frac{1}{4}$.

We import the following standard tools.

Fact 1 (Vector Bernstein)**.** *For independent mean-zero $\boldsymbol{x}_\ell \in \mathbb{R}^d$ with $\|\boldsymbol{x}_\ell\|_2 \le R$ and $\sum_\ell \mathbb{E}\,\|\boldsymbol{x}_\ell\|_2^2 \le \nu$,*

$$\mathbb{P}\left(\left\| \sum_\ell \boldsymbol{x}_\ell \right\|_2 \ge t \right) \le 2(d+1) \exp\Big(- \frac{t^2}{2\nu + 2Rt/3} \Big).$$

Fact 2 (Matrix Bernstein)**.** *For independent symmetric mean-zero $\boldsymbol{X}_\ell \in \mathbb{R}^{d\times d}$ with $|||\boldsymbol{X}_\ell|||_2 \le R$ and $\big|\big|\big|\sum_\ell \mathbb{E}\boldsymbol{X}_\ell^2\big|\big|\big|_2 \le \nu$,*

$$\mathbb{P}\left(\left|\left|\left| \sum_\ell \boldsymbol{X}_\ell \right|\right|\right|_2 \ge t \right) \le 2d\, \exp\left(-\frac{t^2}{2\nu + 2Rt/3} \right).$$

Fact 3 (Hoeffding)**.** *For independent $x_\ell \in [\alpha_\ell, \beta_\ell]$,*

$$\mathbb{P}\left(\Big| \sum_\ell (x_\ell - \mathbb{E}x_\ell) \Big| \ge t \right) \le 2 \exp\left(-\frac{2t^2}{\sum_\ell (\beta_\ell - \alpha_\ell)^2} \right).$$

Fact 4 (Degree-theoretic existence; Ortega–Rheinboldt 1970)**.** *If $\Omega \subset \mathbb{R}^p$ is open and bounded, $\boldsymbol{F} : \overline{\Omega} \to \mathbb{R}^p$ is continuous, and there is $\boldsymbol{x}_0 \in \Omega$ with $(\boldsymbol{x} - \boldsymbol{x}_0)^\top \boldsymbol{F}(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \partial\Omega$, then $\boldsymbol{F}$ has a zero in $\overline{\Omega}$.*

Fact 5 (Gershgorin disc theorem)**.** *If $\boldsymbol{A} = (A_{ab})_{a,b=1}^p$ is a $p \times p$ matrix, then*

$$\operatorname{spec}(\boldsymbol{A}) \subseteq \bigcup_{a=1}^{p} \Big\{ z \in \mathbb{C} : \ |z - A_{aa}| \leq \textstyle\sum_{b \neq a} |A_{ab}| \Big\}.$$

In particular, $\lambda_{\max}(\boldsymbol{A}) \leq \max_a \left(A_{aa} + \sum_{b \neq a} |A_{ab}| \right)$ when $\boldsymbol{A}$ is symmetric.

Fact 6 (Schur complement criterion for positive definiteness; Haynsworth 1968)**.** *Let $\boldsymbol{M}$ be a symmetric matrix in block form*

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{Q}^\top & \boldsymbol{R} \end{pmatrix},$$

with $\boldsymbol{P}$ symmetric and invertible, and let $\boldsymbol{M}/\boldsymbol{P} := \boldsymbol{R} - \boldsymbol{Q}^\top \boldsymbol{P}^{-1} \boldsymbol{Q}$ denote the Schur complement of $\boldsymbol{P}$ in $\boldsymbol{M}$. Then the inertias add,

$$\operatorname{In}(\boldsymbol{M}) = \operatorname{In}(\boldsymbol{P}) + \operatorname{In}(\boldsymbol{M}/\boldsymbol{P}),$$

where $\operatorname{In}(\cdot) = (n_+, n_-, n_0)$ counts the positive, negative and zero eigenvalues. In particular, if $\boldsymbol{P} \succ 0$ then

$$\boldsymbol{M} \succ 0 \quad \Longleftrightarrow \quad \boldsymbol{M}/\boldsymbol{P} \succ 0.$$

B.7 Assumptions and the population curvature floor

Assumption 1 (Dense bounded design)**.** $1 \leq m_{ij} \leq m^*$ for all $0 \leq i < j \leq n$, with m^* a fixed constant.

Assumption 2 (Compactness)**.** There exist absolute constants $C_s, \kappa > 0$ such that, for all n,

$$\|\boldsymbol{s}^*\|_\infty \leq C_s, \qquad \sup_{0 \leq i,j \leq n,\ k} \|\boldsymbol{Y}_{ijk}\|_2 \leq \kappa, \qquad \sup_{0 \leq i,j \leq n,\ k} \|\boldsymbol{Z}_{ijk}\|_2 \leq \kappa.$$

Assumption 3 (Smooth even gate)**.** f is even and C^3, so that

$$F_1 := \sup_{|u| \leq 2C_s + 1} |f'(u)|, \qquad F_2 := \sup_{|u| \leq 2C_s + 1} |f''(u)|, \qquad F_3 := \sup_{|u| \leq 2C_s + 1} |f'''(u)|$$

are finite.

Remark 1 (Working region). The supremum range $|u| \leq 2C_s + 1$ in A3 is a fixed interval strictly containing the range $[-2C_s, 2C_s]$ of the population gaps. Throughout, the *working region* is the set $\{\boldsymbol{s} : \|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2}\}$: every merit vector $\boldsymbol{s}$ that arises below lies within $o(1)$ of $\boldsymbol{s}^*$, hence lies in the working region for n large; its gaps then obey $|\Delta_{ij}| \leq 2C_s + 1$, so the bounds F_1, F_2, F_3 apply to every gap encountered. The Hessian-geometry lemmas of Appendix B.10 are stated for $\|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2}$ accordingly.

Under A1–A3 the predictors $\eta^*_{ijk}, \theta^*_{ijk}$ lie in a fixed compact interval, so there are absolute constants $0 < \underline{g} \leq \overline{g} < 1$ with $\underline{g} \leq g^*_{ijk} \leq \overline{g}$, and likewise for d^*_{ijk}; hence $w^*_{\eta,ijk}, w^*_{\theta,ijk} \in [\underline{w}, \frac{1}{4}]$ for an absolute $\underline{w} > 0$. Define the *population merit weight*

$$\bar{w}_{ijk} := w^*_{\eta,ijk}\, f'(\Delta^*_{ij})^2 + g^*_{ijk}\, w^*_{\theta,ijk} \ \ (\geq 0). \tag{B.23}$$

Lemma 1 (Curvature floor)**.** *Under A2–A3 there are absolute constants $0 < b_0 \leq b_1$ with $b_0 \leq \bar{w}_{ijk} \leq b_1$ for all i, j, k. One may take $b_0 = \underline{g}\,\underline{w}$ and $b_1 = \frac{1}{4}(F_1^2 + 1)$.*

Proof. The upper bound is immediate: $w^*_\eta \leq \frac{1}{4}$, $f'(\Delta^*)^2 \leq F_1^2$, $g^* \leq 1$, $w^*_\theta \leq \frac{1}{4}$, so $\bar{w}_{ijk} \leq \frac{1}{4}F_1^2 + \frac{1}{4}$. For the lower bound, both terms in (B.23) are nonnegative, so it suffices to bound the second:

$$\bar{w}_{ijk} \ \geq \ g^*_{ijk}\, w^*_{\theta,ijk} \ \geq \ \underline{g}\,\underline{w} \ =: \ b_0 \ > \ 0.$$

The point is that at a tie ($\Delta^*_{ij} = 0$) the gate channel vanishes, since f even implies $f'(0) = 0$; the floor is then carried entirely by the direction channel $g^* w^*_\theta$, which is positive precisely because tied pairs are still cited with probability $\geq \underline{g}$ under compactness. □

Remark 2. If $\|s^*\|_\infty$ is allowed to grow, $\underline{g} \to 0$ and $b_0 \to 0$, which is the regime where the rates stated below degrade.

Assumption 4 (Gate design)**.** There is a constant $c_0 > 0$ such that

$$\lambda_{\min}\Big(\sum_{i<j}\sum_{k} w^*_{\eta,ijk}\, \boldsymbol{Y}_{ijk}\boldsymbol{Y}_{ijk}^\top\Big) \geq c_0 n^2.$$

Assumption 5 (Direction design)**.** There is a constant $c_0 > 0$ such that

$$\lambda_{\min}\Big(\sum_{i<j}\sum_{k} g^*_{ijk} w^*_{\theta,ijk}\, \boldsymbol{Z}_{ijk}\boldsymbol{Z}_{ijk}^\top\Big) \geq c_0 n^2.$$

Remark 3. A4 and A5 posit constants that may differ; without loss of generality we use a single c_0 for both, namely the smaller of the two. A further assumption, on the efficient information for the structural block, is stated as A6 in Appendix B.12, where the relevant Schur complement is introduced; it implies A4–A5 (Remark 6), which we nevertheless retain for their interpretability.

B.7.1 The population Hessian floor

At the truth, decompose the realized merit Hessian $\boldsymbol{J}^* := \boldsymbol{J}(s^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$ as

$$\boldsymbol{J}^* = \boldsymbol{V} + \boldsymbol{E}, \qquad \boldsymbol{V} := \mathbb{E}[\boldsymbol{J}^*] = \sum_{i<j}\sum_{k} \bar{w}_{ijk}\, \boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{ij}^\top, \qquad \boldsymbol{E} := \sum_{i<j}\sum_{k} \xi_{ijk}\, \boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{ij}^\top, \quad \text{(B.24)}$$

where, using $\mathbb{E}[c_{ijk}] = g^*_{ijk}$ and w^*_θ deterministic given $s^*, \boldsymbol{\beta}^*$,

$$\xi_{ijk} := W^*_{ijk} - \bar{w}_{ijk} = (c_{ijk} - g^*_{ijk})\big(w^*_{\theta,ijk} - f''(\Delta^*_{ij})\big), \qquad \mathbb{E}[\xi_{ijk}] = 0, \quad |\xi_{ijk}| \leq B := \frac{1}{4} + F_2. \quad \text{(B.25)}$$

$\boldsymbol{V}$ is the expected (population) Hessian; $\boldsymbol{E}$ is the mean-zero "saboteur". The symbol ξ_{ijk} denotes this *curvature* fluctuation throughout; the unrelated leave-one-out *score* fluctuation introduced in (B.30) is written χ_{imk}.

Lemma 2 (Population floor)**.** *Under A1–A3, $\lambda_{\min,\perp}(\boldsymbol{V}) \geq n b_0$, with b_0 as in Lemma 1.*

Proof. $\boldsymbol{V}$ is the Laplacian of the comparison graph with edge weights $\sum_k \bar{w}_{ijk}$. For centered $\boldsymbol{x}$ ($\mathbf{1}^\top \boldsymbol{x} = 0$),

$$\boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{x} = \sum_{i<j}\sum_{k} \bar{w}_{ijk}(x_i - x_j)^2 \;\geq\; b_0 \sum_{i<j}(x_i - x_j)^2 = b_0\, \boldsymbol{x}^\top \boldsymbol{L}\, \boldsymbol{x},$$

using $\bar{w}_{ijk} \geq b_0$ (Lemma 1) and $m_{ij} \geq 1$ (A1), where $\boldsymbol{L}$ is the unweighted Laplacian of K_{n+1} (the complete graph on $\{0, \ldots, n\}$). Since $\lambda_{\min,\perp}(\boldsymbol{L}) = \lambda_2(\boldsymbol{L}) = n+1$ for K_{n+1}, $\boldsymbol{x}^\top \boldsymbol{V} \boldsymbol{x} \geq n b_0 \|\boldsymbol{x}\|_2^2$. Minimizing over centered unit $\boldsymbol{x}$ gives the claim. □

B.8 The leave-one-out construction

The ℓ_∞ analysis of Appendix B.11 compares each gradient iterate with a *leave-one-out (LOO)* proxy, defined here once and for all. Fix an item $m \in \{0, 1, \ldots, n\}$, and let $\mathcal{F}_{-m}$ be the σ-algebra generated by the outcomes of all comparisons *not* involving item m,

$$\mathcal{F}_{-m} := \sigma\big(\{c_{ijk}, a^1_{ijk} \;:\; 0 \leq i < j \leq n,\; m \notin \{i, j\},\; k\}\big).$$

The *LOO objective* replaces every comparison involving m by its expectation under the truth, leaving the remaining comparisons untouched:

$$\ell^{(m)}(s, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{\substack{i<j \\ i\neq m,\; j\neq m}} \sum_{k} \ell_{ijk}(s, \boldsymbol{\gamma}, \boldsymbol{\beta}) \;+\; \sum_{i\neq m}\sum_{k} \bar{\ell}_{imk}(s, \boldsymbol{\gamma}, \boldsymbol{\beta}), \quad \text{(B.26)}$$

where

$$\bar{\ell}_{imk}(s, \boldsymbol{\gamma}, \boldsymbol{\beta}) := g^*_{imk} \log g_{imk} + (1 - g^*_{imk}) \log(1 - g_{imk}) \\ + g^*_{imk}\big[d^*_{imk} \log d_{imk} + (1 - d^*_{imk}) \log(1 - d_{imk})\big].$$

Its merit gradient $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}} := \nabla_{\boldsymbol{s}}(-\ell^{(m)})$ is

$$\begin{aligned}\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) = &- \sum_{\substack{i<j \\ i\neq m,\ j\neq m}} \sum_{k} \big[(c_{ijk} - g_{ijk}) f'(\Delta_{ij}) + (a^1_{ijk} - c_{ijk} d_{ijk})\big]\boldsymbol{\omega}_{ij} \\ &- \sum_{i\neq m} \sum_{k} \big[(g^*_{imk} - g_{imk}) f'(\Delta_{im}) + g^*_{imk}(d^*_{imk} - d_{imk})\big]\boldsymbol{\omega}_{im},\end{aligned} \tag{B.27}$$

and its merit Hessian $\boldsymbol{J}^{(m)} := \nabla^2_{\boldsymbol{s}}(-\ell^{(m)})$ is the graph Laplacian

$$\boldsymbol{J}^{(m)}(\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{\beta}) = \sum_{i<j} W^{(m)}_{ij}\, \boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{ij}^\top, \qquad W^{(m)}_{ij} = \sum_k W^{(m)}_{ijk}, \tag{B.28}$$

with edge weights

$$W^{(m)}_{ijk} = \begin{cases} W_{ijk}, & m \notin \{i,j\}, \\ w_{\eta,imk}\, f'(\Delta_{im})^2 + g^*_{imk}\, w_{\theta,imk} - (g^*_{imk} - g_{imk})\, f''(\Delta_{im}), & m \in \{i,j\}. \end{cases} \tag{B.29}$$

Remark 4 (Symmetry of the weights). Every per-comparison curvature weight is symmetric under $i \leftrightarrow j$: indeed $g_{jik} = g_{ijk}$ and $w_{\eta,jik} = w_{\eta,ijk}$ (the gate predictor is even), $w_{\theta,jik} = w_{\theta,ijk}$ (since $d_{jik} = 1 - d_{ijk}$), $f'(\Delta_{ji})^2 = f'(\Delta_{ij})^2$ and $f''(\Delta_{ji}) = f''(\Delta_{ij})$ (f' odd, f'' even), and $c_{jik} = c_{ijk}$; hence $W_{jik} = W_{ijk}$ and likewise $W^{(m)}_{mik} = W^{(m)}_{imk}$ (using $g^*_{mik} = g^*_{imk}$). The score summands are instead antisymmetric, $\psi_{jik} = -\psi_{ijk}$, matching $\boldsymbol{\omega}_{ji} = -\boldsymbol{\omega}_{ij}$. We use the two orderings of a pair interchangeably below.

Like $\boldsymbol{G}$ and $\boldsymbol{J}$, the objects $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ and $\boldsymbol{J}^{(m)}$ take values in / act on $\mathbf{1}^\perp$ (every $\boldsymbol{\omega}_{im}$ is centered). Two structural facts are used repeatedly.

Measurability. The construction (B.26) strips out all of item m's randomness: in the m-edge terms $\bar{\ell}_{imk}$ the outcomes c_{imk}, a^1_{imk} have been replaced by their population expectations g^*_{imk} and $g^*_{imk}d^*_{imk}$, so $\bar{\ell}_{imk}$ is a *non-random* function of $(\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{\beta})$ (the covariates and the population quantities are fixed under A1–A3). The remaining terms depend on the data only through the outcomes of comparisons disjoint from m, i.e. the generators of $\mathcal{F}_{-m}$. Hence $\ell^{(m)}(\cdot,\boldsymbol{\gamma},\boldsymbol{\beta})$, $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$, and $\boldsymbol{J}^{(m)}$ are $\mathcal{F}_{-m}$-measurable as random functions of $\boldsymbol{s}$ (a non-random function being measurable with respect to any σ-algebra). Consequently any iterative scheme driven by $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ and started at the deterministic point $\boldsymbol{s}^*$ produces $\mathcal{F}_{-m}$-measurable iterates; in particular the LOO trajectory $\{\boldsymbol{s}^{t,(m)}\}_t$ of Algorithm 2 below is $\mathcal{F}_{-m}$-measurable at every step.

The score discrepancy. Subtracting (B.14) from (B.27), all pairs disjoint from m cancel and the m-edges combine into

$$\begin{aligned}\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) = \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) &:= \sum_{i\neq m}\sum_k \chi_{imk}\,\boldsymbol{\omega}_{im}, \\ \chi_{imk} &:= (c_{imk} - g^*_{imk})\big(f'(\Delta_{im}) - d_{imk}\big) + \big(a^1_{imk} - g^*_{imk}d^*_{imk}\big).\end{aligned} \tag{B.30}$$

Note that $\boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ depends on the structural parameters only through $d_{imk} = \sigma(\Delta_{im} + \boldsymbol{Z}^\top_{imk}\boldsymbol{\beta})$, i.e. on $\boldsymbol{\beta}$ and not on $\boldsymbol{\gamma}$.

B.9 Concentration

Lemma 3 (Score bounds). *Under A1–A3, for every $c > 0$, with probability at least $1 - O(n^{-c})$,*

$$\textit{(merit)} \quad \max_{0\le a\le n} |G_a(\boldsymbol{s}^*,\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*)| \le C_\psi\sqrt{n\log n}, \qquad C_\psi := 2(F_1+1)\sqrt{(c+1)\,m^*}, \tag{B.31}$$

$$\textit{(structural)} \quad \|\nabla_{\boldsymbol{\gamma}}\ell(\boldsymbol{s}^*,\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*)\|_2 \le C_\kappa\, n\sqrt{\log n}, \qquad \|\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{s}^*,\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*)\|_2 \le C_\kappa\, n\sqrt{\log n}, \tag{B.32}$$

where $C_\kappa := \kappa\sqrt{2\,m^*(c+1)\max\{p_\gamma, p_\beta\}}$.

Proof. Merit. Fix a. By (B.13), $G_a = -\sum_{j\neq a}\sum_k \psi_{ajk}$ at the truth, a sum of $m_a := \sum_{j\neq a} m_{aj} \in [n, nm^*]$ independent terms. Each ψ_{ajk} is a bounded function of (c_{ajk}, a^1_{ajk}) with $|\psi_{ajk}| \le F_1 + 1$ and is mean-zero: $\mathbb{E}[(c-g^*)f'] = 0$ since $\mathbb{E}c = g^*$, and $\mathbb{E}[a^1 - cd^*] = \mathbb{E}[\mathbb{E}[a^1 \mid c] - cd^*] = \mathbb{E}[cd^* - cd^*] = 0$. Hoeffding (Fact 3) with per-term range $2(F_1+1)$ gives

$$\mathbb{P}\big(\,|G_a| \ge t\big) \le 2\exp\!\left(-\frac{2t^2}{m_a(2(F_1+1))^2}\right) = 2\exp\!\left(-\frac{t^2}{2m_a(F_1+1)^2}\right).$$

Take $t = C_\psi\sqrt{n\log n}$ with $C_\psi = 2(F_1+1)\sqrt{(c+1)m^*}$. Since $m_a \le nm^*$, the exponent is

$$-\frac{C_\psi^2\, n\log n}{2m_a(F_1+1)^2} \le -\frac{4(F_1+1)^2(c+1)m^*\, n\log n}{2\,nm^*(F_1+1)^2} = -2(c+1)\log n,$$

so $\mathbb{P}(|G_a| \ge t) \le 2n^{-2(c+1)}$. A union bound over $a = 0,\dots,n$ gives the merit claim with probability $\ge 1 - 2(n+1)n^{-2(c+1)} = 1 - O(n^{-(2c+1)}) = 1 - O(n^{-c})$.

Structural. Each coordinate of $\nabla_{\boldsymbol\gamma}\ell$ (and of $\nabla_{\boldsymbol\beta}\ell$) is a sum of $M := \sum_{i<j} m_{ij} \le \binom{n+1}{2}m^*$ independent mean-zero terms; the l-th coordinate of $\nabla_{\boldsymbol\gamma}\ell$ is $\sum_{i<j,k}(c_{ijk} - g^*_{ijk})Y_{ijk,l}$, with each term bounded in absolute value by κ since $|c - g^*| \le 1$ and $|Y_{ijk,l}| \le \|\boldsymbol{Y}_{ijk}\|_2 \le \kappa$ (and likewise for $\boldsymbol\beta$). Hoeffding with per-term range 2κ and $t = C'_\kappa\sqrt{M\log n}$, $C'_\kappa := \kappa\sqrt{2(c+1)}$, gives, for each fixed coordinate,

$$\mathbb{P}\big(\,|\text{coord}| \ge t\big) \le 2\exp\!\Big(-\frac{2t^2}{M(2\kappa)^2}\Big) = 2\exp\big(-(c+1)\log n\big) = 2n^{-(c+1)}.$$

A union bound over the fixed $p_\gamma + p_\beta$ coordinates yields $\max\{\|\nabla_{\boldsymbol\gamma}\ell\|_\infty, \|\nabla_{\boldsymbol\beta}\ell\|_\infty\} \le C'_\kappa\sqrt{M\log n}$ with probability $1 - O(n^{-(c+1)}) = 1 - O(n^{-c})$. Since p_γ, p_β are fixed and $M \le \binom{n+1}{2}m^* \le n^2 m^*$ (using $\binom{n+1}{2} \le n^2$ for $n \ge 1$), so $\sqrt{M} \le n\sqrt{m^*}$,

$$\|\nabla_{\boldsymbol\gamma}\ell\|_2 \le \sqrt{p_\gamma}\, C'_\kappa\sqrt{M\log n} \le \kappa\sqrt{2p_\gamma m^*(c+1)}\; n\sqrt{\log n} \le C_\kappa\, n\sqrt{\log n},$$

with $C_\kappa := \kappa\sqrt{2m^*(c+1)\max\{p_\gamma, p_\beta\}}$, and identically for $\nabla_{\boldsymbol\beta}\ell$. □

Lemma 4 (Spectral control of the saboteur)**.** *Under A1–A3, for every $c > 0$ there is an absolute constant C_E such that the event*

$$\mathcal{E}_{\text{sab}} := \big\{\, \|\!|\boldsymbol{E}|\!\|_2 \le C_E\sqrt{n\log n}\big\}$$

has probability at least $1 - O(n^{-c})$.

Proof. Write $\boldsymbol{E} = \sum_{i<j,k}\boldsymbol{X}_{ijk}$ with $\boldsymbol{X}_{ijk} = \xi_{ijk}\,\boldsymbol\omega_{ij}\boldsymbol\omega_{ij}^\top$, independent, symmetric, mean-zero ($\mathbb{E}\xi_{ijk} = 0$ by (B.25)). Since $\|\boldsymbol\omega_{ij}\|_2^2 = 2$ and $|\xi_{ijk}| \le B$, $\|\!|\boldsymbol{X}_{ijk}|\!\|_2 \le 2B =: R$. Using $(\boldsymbol\omega_{ij}\boldsymbol\omega_{ij}^\top)^2 = 2\,\boldsymbol\omega_{ij}\boldsymbol\omega_{ij}^\top$ and $\mathbb{E}[\xi_{ijk}^2] \le B^2$,

$$\sum_{i<j,k}\mathbb{E}\boldsymbol{X}_{ijk}^2 = \sum_{i<j,k} 2\,\mathbb{E}[\xi_{ijk}^2]\,\boldsymbol\omega_{ij}\boldsymbol\omega_{ij}^\top \preceq 2B^2m^*\sum_{i<j}\boldsymbol\omega_{ij}\boldsymbol\omega_{ij}^\top,$$

and the right-hand side of the last display is $2B^2m^*$ times the Laplacian of K_{n+1}, of spectral norm $n+1$, so $\Big\|\!\Big|\sum_{i<j,k}\mathbb{E}\boldsymbol{X}_{ijk}^2\Big|\!\Big\|_2 \le 2B^2m^*(n+1) =: \nu$. With $t = C_E\sqrt{n\log n}$, the matrix-Bernstein (Fact 2) exponent satisfies

$$-\frac{t^2}{2\nu + 2Rt/3} \;\underset{n\to\infty}{\sim}\; -\frac{C_E^2}{4B^2m^*}\log n,$$

so $\mathbb{P}(\|\!|\boldsymbol{E}|\!\|_2 \ge t) \le 2(n+1)\,n^{-C_E^2/(4B^2m^*)} = O(n^{-c})$ for C_E large enough. □

Lemma 5 (Leave-one-out perturbation bound)**.** *Assume A1–A3, and fix $m \in \{0, 1, \dots, n\}$, a pair $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, and an $\mathcal{F}_{-m}$-measurable random point $\boldsymbol{s} \in \mathbb{R}^{n+1}$. With $\boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ and χ_{imk} as in* (B.30)*, for every $c > 0$, with $K := 2(F_1 + 2)\sqrt{2(c+1)\, m^*}$, there is $n_0 = n_0(c)$ such that for all $n \geq n_0$,*

$$\mathbb{P}\Big(\Big\{ \Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) \Big\|_2 > K\sqrt{n \log n} \Big\} \cap \Big\{ \|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2} \Big\} \Big) \;=\; O(n^{-c}).$$

In particular, if $\|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2}$ almost surely, then $\Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) \Big\|_2 \leq K\sqrt{n \log n}$ with probability at least $1 - O(n^{-c})$.

Proof. Condition on $\mathcal{F}_{-m}$. By hypothesis $\boldsymbol{s}$ is then deterministic, hence so are $f'(\Delta_{im})$ and d_{imk}, and the confinement event $\{\|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2}\}$ is $\mathcal{F}_{-m}$-measurable. On the $\mathcal{F}_{-m}$-atoms where confinement fails, the conditional probability of the displayed intersection is zero, so it suffices to prove the conditional bound on the atoms where $\|\boldsymbol{s}\|_\infty \leq C_s + \frac{1}{2}$; there, every gap satisfies $|\Delta_{im}| \leq 2C_s + 1$, so the bounds of A3 apply. The pairs $\{(c_{imk}, a^1_{imk})\}_{i \neq m, k}$ are independent across (i, k) with the law (B.8) and independent of $\mathcal{F}_{-m}$. Set $\boldsymbol{X}_{imk} := \chi_{imk}\, \boldsymbol{\omega}_{im} \in \mathbb{R}^{n+1}$ and verify vector Bernstein (Fact 1) for $\{\boldsymbol{X}_{imk}\}_{i \neq m, k}$ under the conditional law.

(a) Mean-zero. $\mathbb{E}[c_{imk} - g^*_{imk}] = 0$ and $\mathbb{E}[a^1_{imk} - g^*_{imk} d^*_{imk}] = 0$. Since $f'(\Delta_{im}) - d_{imk}$ is deterministic under the conditioning, $\mathbb{E}[\chi_{imk} \mid \mathcal{F}_{-m}] = 0$, hence $\mathbb{E}[\boldsymbol{X}_{imk} \mid \mathcal{F}_{-m}] = \boldsymbol{0}$.

(b) Uniform bound. $|c - g^*| \leq 1$, $\big|a^1 - g^* d^*\big| \leq 1$, $|f'| \leq F_1$ (A3, on the confinement atoms), $d \in [0, 1]$, so $|\chi_{imk}| \leq (F_1 + 1) + 1 = F_1 + 2$. As $\|\boldsymbol{\omega}_{im}\|_2 = \sqrt{2}$, $\|\boldsymbol{X}_{imk}\|_2 = \sqrt{2}\, |\chi_{imk}| \leq \sqrt{2}\,(F_1 + 2) =: R$.

(c) Total second moment. There are n indices $i \neq m$ in $\{0, \dots, n\}$, each with $\leq m^*$ comparisons, so

$$\sum_{i \neq m, k} \mathbb{E}\big[\, \|\boldsymbol{X}_{imk}\|_2^2 \mid \mathcal{F}_{-m} \big] = \sum_{i \neq m, k} 2\, \mathbb{E}[\chi^2_{imk} \mid \mathcal{F}_{-m}] \leq 2(F_1 + 2)^2 m^*\, n =: \nu.$$

By vector Bernstein in dimension $d = n + 1$, for any $t \geq 0$,

$$\mathbb{P}\Big(\Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) \Big\|_2 \geq t \;\Big|\; \mathcal{F}_{-m} \Big) \leq 2(n + 2) \exp\Big(-\frac{t^2}{2\nu + 2Rt/3} \Big).$$

Take $t = K\sqrt{n \log n}$ with $K = 2(F_1 + 2)\sqrt{2(c+1)m^*}$. Then

$$\frac{Rt}{3\nu} = \frac{\sqrt{2}(F_1 + 2)\, K\sqrt{n \log n}}{3 \cdot 2(F_1 + 2)^2 m^* n} = \frac{K}{3\sqrt{2}\,(F_1 + 2) m^*} \sqrt{\frac{\log n}{n}} \;\leq\; 1$$

for $n \geq n_0$, so $2\nu + 2Rt/3 \leq 4\nu$ and

$$\mathbb{P}\Big(\Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) \Big\|_2 \geq t \;\Big|\; \mathcal{F}_{-m} \Big) \leq 2(n + 2) \exp\Big(-\frac{t^2}{4\nu} \Big) = 2(n + 2) \exp\Big(-\frac{K^2\, n \log n}{8(F_1 + 2)^2 m^* n} \Big).$$

With $K^2 = 8(c + 1)(F_1 + 2)^2 m^*$ the exponent is $-(c + 1) \log n$, so the conditional probability is at most $2(n + 2)\, n^{-(c+1)} = O(n^{-c})$ on the confinement atoms. Taking expectations over $\mathcal{F}_{-m}$ completes the proof. □

Remark 5 (Why the confinement dichotomy). The uniform bound in step (b) is the only place where the position of $\boldsymbol{s}$ enters, through $|f'(\Delta_{im})| \leq F_1$; outside the working region f' need not be bounded. Lemma 5 is therefore stated in the "dichotomy" form: the exceptional event is small *on the part of the space where $\boldsymbol{s}$ is confined*. In the applications below, $\boldsymbol{s}$ is an LOO iterate whose confinement is exactly what the coupled induction of Lemma 14 establishes on the good event, so the two statements mesh with no circularity: the good events are defined in dichotomy form, and the induction supplies the confinement that activates the score bound at each step.

B.10 HESSIAN ESTIMATES

Lemma 6 (Hessian Lipschitz bound in $\boldsymbol{s}$)**.** *Under A1–A3, for all $\boldsymbol{x}, \boldsymbol{y}$ with $\|\boldsymbol{x}\|_\infty, \|\boldsymbol{y}\|_\infty \le C_s + \frac{1}{2}$ and any fixed $\boldsymbol{\gamma}, \boldsymbol{\beta}$,*

$$\|\!|\boldsymbol{J}(\boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta})|\!\|_\infty \le 4\, L_f\, (\max_a m_a)\ \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \le 4\, L_f\, m^* n\ \|\boldsymbol{x} - \boldsymbol{y}\|_\infty ,$$

where $L_f = \frac{1}{4}F_1^3 + \frac{3}{4}F_1F_2 + F_3 + \frac{1}{4}$ and $m_a := \sum_{b\neq a} m_{ab}$. The same bound holds verbatim for the leave-one-out Hessian $\boldsymbol{J}^{(m)}$ of (B.28)*.*

Proof. Since $\boldsymbol{J}$ depends on $\boldsymbol{s}$ only through the gaps, it suffices to bound the scalar derivative $W'_{ijk}(\Delta) := \frac{dW_{ijk}}{d\Delta}$ from (B.15). Using

$$\frac{dg}{d\Delta} = w_\eta f', \quad \frac{dw_\eta}{d\Delta} = w_\eta(1-2g)f', \quad \frac{dd}{d\Delta} = w_\theta, \quad \frac{dw_\theta}{d\Delta} = w_\theta(1-2d), \quad \frac{dc}{d\Delta} = 0,$$

the three terms of W differentiate to

$$\begin{aligned} \frac{d}{d\Delta}\big[w_\eta f'^2\big] &= w_\eta(1-2g)f'^3 + 2w_\eta f' f'', \\ \frac{d}{d\Delta}\big[c\, w_\theta\big] &= c\, w_\theta(1-2d), \\ \frac{d}{d\Delta}\big[-(c-g)f''\big] &= w_\eta f' f'' - (c-g)f'''. \end{aligned}$$

Adding (the two $w_\eta f' f''$ terms combine to $3w_\eta f' f''$),

$$W'(\Delta) = w_\eta(1-2g)f'^3 + 3w_\eta f' f'' + c\, w_\theta(1-2d) - (c-g)f'''.$$

Bounding with $w_\eta, w_\theta \le \frac{1}{4}$, $|1-2g|, |1-2d| \le 1$, $c \in [0,1]$, $|c-g| \le 1$, and $|f'| \le F_1, |f''| \le F_2, |f'''| \le F_3$ (valid since $\|\boldsymbol{x}\|_\infty, \|\boldsymbol{y}\|_\infty \le C_s + \frac{1}{2}$ gives gaps of magnitude at most $2C_s + 1$, the range of A3) gives $|W'(\Delta)| \le \frac{1}{4}F_1^3 + \frac{3}{4}F_1F_2 + \frac{1}{4} + F_3 = L_f$. To assemble: $\boldsymbol{J}$ is a Laplacian with edge weights $W_{ij} = \sum_k W_{ijk}$; set $\delta W_{ij} = W_{ij}(\boldsymbol{x}) - W_{ij}(\boldsymbol{y})$. Row a of $\boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{J}(\boldsymbol{y})$ has diagonal $\sum_{b\neq a}\delta W_{ab}$ and off-diagonals $-\delta W_{ab}$, so

$$\big\|\big(\boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{J}(\boldsymbol{y})\big)_{a\cdot}\big\|_1 = \Big|\sum_{b\neq a}\delta W_{ab}\Big| + \sum_{b\neq a}|\delta W_{ab}| \le 2\sum_{b\neq a}|\delta W_{ab}| .$$

By the mean value theorem and $|\Delta^{\boldsymbol{x}}_{ab} - \Delta^{\boldsymbol{y}}_{ab}| = |(x_a - y_a) - (x_b - y_b)| \le 2\,\|\boldsymbol{x} - \boldsymbol{y}\|_\infty$, each comparison contributes $|W'_{abk}|\ |\Delta^{\boldsymbol{x}}_{ab} - \Delta^{\boldsymbol{y}}_{ab}| \le 2L_f\,\|\boldsymbol{x} - \boldsymbol{y}\|_\infty$, so $|\delta W_{ab}| \le m_{ab}\, L_f \cdot 2\,\|\boldsymbol{x} - \boldsymbol{y}\|_\infty$. Summing over b,

$$\|(\boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{J}(\boldsymbol{y}))_{a\cdot}\|_1 \le 4L_f\Big(\sum_{b\neq a} m_{ab}\Big)\,\|\boldsymbol{x} - \boldsymbol{y}\|_\infty = 4L_f m_a\,\|\boldsymbol{x} - \boldsymbol{y}\|_\infty ,$$

and maximizing over a and using $m_a \le m^* n$ (A1) gives the claim. For $\boldsymbol{J}^{(m)}$ the m-edge weight $W^{(m)}_{imk}$ has, by the same calculation,

$$\Big|\frac{dW^{(m)}_{imk}}{d\Delta}\Big| = \big|w_\eta(1-2g)f'^3 + 3w_\eta f' f'' + g^* w_\theta(1-2d) - (g^* - g)f'''\big| \le L_f$$

(now $g^* \le 1$ replaces $c \in [0,1]$ and $|g^* - g| \le 1$ replaces $|c-g| \le 1$), so the identical bound holds. □

Lemma 7 (Hessian Lipschitz bound in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$)**.** *Under A1–A3, there is an absolute constant L_ζ such that for all $\boldsymbol{s}$ with $\|\boldsymbol{s}\|_\infty \le C_s + \frac{1}{2}$ and all $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathbb{R}^{p_\gamma}$, $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^{p_\beta}$,*

$$\|\!|\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}', \boldsymbol{\beta}')|\!\|_\infty \le L_\zeta\, n\,\big(\,\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2\,\big).$$

The same bound holds verbatim for the leave-one-out Hessian $\boldsymbol{J}^{(m)}$ of (B.28)*.*

Proof. As in Lemma 6, $\boldsymbol{J}$ is a Laplacian with edge weights $W_{ij} = \sum_k W_{ijk}$, and the row-sum norm is controlled by $\sum_{b\neq a,k} |W_{abk}(\boldsymbol{\gamma},\boldsymbol{\beta}) - W_{abk}(\boldsymbol{\gamma}',\boldsymbol{\beta}')|$ up to the factor 2 from the diagonal-plus-off-diagonal structure. It therefore suffices to bound the per-comparison increment of W_{ijk} in $(\boldsymbol{\gamma},\boldsymbol{\beta})$ at fixed $\boldsymbol{s}$.

Recall $W_{ijk} = w_\eta f'(\Delta_{ij})^2 + c\, w_\theta - (c-g) f''(\Delta_{ij})$ with $g = \sigma(f(\Delta_{ij}) + \boldsymbol{Y}_{ijk}^\top \boldsymbol{\gamma})$, $w_\eta = g(1-g)$, $d = \sigma(\Delta_{ij} + \boldsymbol{Z}_{ijk}^\top \boldsymbol{\beta})$, $w_\theta = d(1-d)$. Differentiating through the predictors $\eta = f + \boldsymbol{Y}^\top \boldsymbol{\gamma}$ and $\theta = \Delta + \boldsymbol{Z}^\top \boldsymbol{\beta}$:

$$\nabla_{\boldsymbol{\gamma}} g = w_\eta\, \boldsymbol{Y}, \quad \nabla_{\boldsymbol{\gamma}} w_\eta = w_\eta(1-2g)\, \boldsymbol{Y}, \quad \nabla_{\boldsymbol{\beta}} d = w_\theta\, \boldsymbol{Z}, \quad \nabla_{\boldsymbol{\beta}} w_\theta = w_\theta(1-2d)\, \boldsymbol{Z},$$

while $\nabla_{\boldsymbol{\gamma}} \Delta = \nabla_{\boldsymbol{\beta}} \Delta = \mathbf{0}$ at fixed $\boldsymbol{s}$ and $\nabla_{\boldsymbol{\beta}} g = \nabla_{\boldsymbol{\gamma}} d = \mathbf{0}$. Hence

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} W_{ijk} &= \big[w_\eta(1-2g) f'(\Delta_{ij})^2 + f''(\Delta_{ij})\, w_\eta \big]\, \boldsymbol{Y}_{ijk}, \\ \nabla_{\boldsymbol{\beta}} W_{ijk} &= c\, w_\theta(1-2d)\, \boldsymbol{Z}_{ijk}. \end{aligned}$$

Using $w_\eta, w_\theta \le \frac{1}{4}$, $|1-2g|, |1-2d| \le 1$, $c \in [0,1]$, $|f'| \le F_1$, $|f''| \le F_2$ on the working region (A3), and $\|\boldsymbol{Y}_{ijk}\|_2, \|\boldsymbol{Z}_{ijk}\|_2 \le \kappa$ (A2),

$$\|\nabla_{\boldsymbol{\gamma}} W_{ijk}\|_2 \le \frac{1}{4}(F_1^2 + F_2)\,\kappa, \qquad \|\nabla_{\boldsymbol{\beta}} W_{ijk}\|_2 \le \frac{1}{4}\,\kappa.$$

By the mean value theorem along the segment from $(\boldsymbol{\gamma}',\boldsymbol{\beta}')$ to $(\boldsymbol{\gamma},\boldsymbol{\beta})$,

$$|W_{abk}(\boldsymbol{\gamma},\boldsymbol{\beta}) - W_{abk}(\boldsymbol{\gamma}',\boldsymbol{\beta}')| \le \frac{1}{4}(F_1^2+F_2)\kappa\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2 + \frac{1}{4}\kappa\, \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 .$$

Summing over the $\le m^*$ comparisons per pair and over $b \neq a$, applying the factor 2 from the Laplacian row structure, and maximizing over a,

$$\begin{aligned} \|\!|\boldsymbol{J}(\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{\beta}) - \boldsymbol{J}(\boldsymbol{s},\boldsymbol{\gamma}',\boldsymbol{\beta}')|\!\|_\infty &\le \frac{1}{2} m^* \kappa\, n \Big((F_1^2+F_2)\, \|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 \Big) \\ &\le L_\zeta\, n \big(\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 \big), \end{aligned}$$

with $L_\zeta := \frac{1}{2} m^* \kappa\, \max\{F_1^2 + F_2,\, 1\}$, an absolute constant under A1–A3.

For $\boldsymbol{J}^{(m)}$, the m-edge weight $W_{imk}^{(m)}$ of (B.29) has, by the same differentiation (the saboteur $-(g_{imk}^* - g_{imk}) f''(\Delta_{im})$ contributing $+\nabla_{\boldsymbol{\gamma}} g_{imk}\, f''(\Delta_{im}) = w_{\eta,imk} f''(\Delta_{im})\, \boldsymbol{Y}_{imk}$, exactly as $-(c_{ijk} - g_{ijk}) f''$ does above),

$$\begin{aligned} \nabla_{\boldsymbol{\gamma}} W_{imk}^{(m)} &= \big[\, w_{\eta,imk}(1-2g_{imk})\, f'(\Delta_{im})^2 + w_{\eta,imk}\, f''(\Delta_{im}) \,\big] \boldsymbol{Y}_{imk}, \\ \nabla_{\boldsymbol{\beta}} W_{imk}^{(m)} &= g_{imk}^*\, w_{\theta,imk}(1-2d_{imk})\, \boldsymbol{Z}_{imk}, \end{aligned}$$

so with $g_{imk}^* \le 1$ replacing $c_{ijk} \in [0,1]$ the bounds $\left\|\nabla_{\boldsymbol{\gamma}} W_{imk}^{(m)}\right\|_2 \le \frac{1}{4}(F_1^2+F_2)\kappa$ and $\left\|\nabla_{\boldsymbol{\beta}} W_{imk}^{(m)}\right\|_2 \le \frac{1}{4}\kappa$ are unchanged, and the row-sum assembly gives the identical bound. □

Lemma 8 (Cross-Jacobian bound)**.** *Assume A1–A3 and $n \ge 2$. There is an absolute constant C_M such that, for all $(\boldsymbol{\gamma},\boldsymbol{\beta}) \in \mathbb{R}^{p_\gamma} \times \mathbb{R}^{p_\beta}$ and all $\boldsymbol{s}$ with $\|\boldsymbol{s}\|_\infty \le C_s + \frac{1}{2}$,*

$$\left\| \frac{\partial \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s})}{\partial(\boldsymbol{\gamma},\boldsymbol{\beta})} \right\|_2 \le C_M\, n^{3/2}.$$

The same bound holds verbatim for the leave-one-out cross-Jacobian $\partial \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s}) / \partial(\boldsymbol{\gamma},\boldsymbol{\beta})$, with $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}$ as in (B.27).

Proof. Since $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}} = \nabla_{\boldsymbol{s}}(-\ell)$ (the merit score (B.14)),

$$\frac{\partial \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}}{\partial(\boldsymbol{\gamma},\boldsymbol{\beta})} = \nabla^2_{\boldsymbol{s},(\boldsymbol{\gamma},\boldsymbol{\beta})}(-\ell) = \big[\, \boldsymbol{B}_{\boldsymbol{\gamma}} \;\big|\; \boldsymbol{B}_{\boldsymbol{\beta}} \,\big],$$

with $\boldsymbol{B}_{\boldsymbol{\gamma}}, \boldsymbol{B}_{\boldsymbol{\beta}}$ as in (B.19)–(B.20). We bound $\|\!|\boldsymbol{B}_{\boldsymbol{\gamma}}|\!\|_2$ and $\|\!|\boldsymbol{B}_{\boldsymbol{\beta}}|\!\|_2$ directly.

For $\boldsymbol{B}_{\boldsymbol{\gamma}}$, take unit $\boldsymbol{u} \in \mathbf{1}^{\perp} \subset \mathbb{R}^{n+1}$ and unit $\boldsymbol{v} \in \mathbb{R}^{p_\gamma}$ (the columns of $\boldsymbol{B}_{\boldsymbol{\gamma}}$ lie in $\mathbf{1}^{\perp}$, so this restriction loses nothing):

$$\boldsymbol{u}^{\top} \boldsymbol{B}_{\boldsymbol{\gamma}} \boldsymbol{v} = \sum_{i<j,k} w_{\eta,ijk} f'(\Delta_{ij}) \, (u_i - u_j) \, (\boldsymbol{Y}_{ijk}^{\top} \boldsymbol{v}).$$

By Cauchy–Schwarz,

$$\big|\boldsymbol{u}^{\top} \boldsymbol{B}_{\boldsymbol{\gamma}} \boldsymbol{v}\big| \le \underbrace{\Big(\sum_{i<j,k} \big(w_{\eta,ijk} f'(\Delta_{ij})\big)^2 (u_i - u_j)^2 \Big)^{1/2}}_{=:T_1} \underbrace{\Big(\sum_{i<j,k} (\boldsymbol{Y}_{ijk}^{\top} \boldsymbol{v})^2 \Big)^{1/2}}_{=:T_2}.$$

For T_1: $|w_{\eta,ijk} f'(\Delta_{ij})| \le \frac{1}{4} F_1$, and summing the $\le m^*$ comparisons per pair, $\sum_{i<j,k} (u_i - u_j)^2 \le m^* \sum_{i<j} (u_i - u_j)^2 \le m^* (n+1)$, so $T_1 \le \frac{1}{4} F_1 \sqrt{m^* (n+1)}$. For T_2: $(\boldsymbol{Y}_{ijk}^{\top} \boldsymbol{v})^2 \le \kappa^2$ and there are $\le \binom{n+1}{2} m^*$ terms, so $T_2 \le \kappa \sqrt{\binom{n+1}{2} m^*} \le \frac{1}{\sqrt{2}} \kappa \sqrt{m^*} (n+1)$. Taking the supremum over $\boldsymbol{u}, \boldsymbol{v}$,

$$\|\!|\boldsymbol{B}_{\boldsymbol{\gamma}}|\!\|_2 \le T_1 T_2 \le \frac{1}{4\sqrt{2}} \, F_1 \kappa \, m^* \, (n+1)^{3/2}.$$

We bound $\boldsymbol{B}_{\boldsymbol{\beta}}$ identically, with $|c_{ijk} w_{\theta,ijk}| \le \frac{1}{4}$ (so F_1 is replaced by 1) and $\boldsymbol{Z}_{ijk}$ in place of $\boldsymbol{Y}_{ijk}$:

$$\|\!|\boldsymbol{B}_{\boldsymbol{\beta}}|\!\|_2 \le \frac{1}{4\sqrt{2}} \, \kappa \, m^* \, (n+1)^{3/2}.$$

Using $\|\!|[\boldsymbol{B}_{\boldsymbol{\gamma}} \mid \boldsymbol{B}_{\boldsymbol{\beta}}]|\!\|_2 \le \|\!|\boldsymbol{B}_{\boldsymbol{\gamma}}|\!\|_2 + \|\!|\boldsymbol{B}_{\boldsymbol{\beta}}|\!\|_2$ and $(n+1)^{3/2} \le 2 n^{3/2}$ for $n \ge 2$,

$$\left\| \frac{\partial \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s})}{\partial (\boldsymbol{\gamma}, \boldsymbol{\beta})} \right\|_2 \le \frac{1}{2\sqrt{2}} \, (F_1 + 1) \, \kappa \, m^* \, n^{3/2} = C_M \, n^{3/2}, \qquad C_M := \frac{1}{2\sqrt{2}} (F_1 + 1) \kappa m^*.$$

For the leave-one-out variant, differentiating the m-edge terms of (B.27) at fixed $\boldsymbol{s}$ (using $\nabla_{\boldsymbol{\gamma}}(g^*_{imk} - g_{imk}) = -w_{\eta,imk} \boldsymbol{Y}_{imk}$ and $\nabla_{\boldsymbol{\beta}}(d^*_{imk} - d_{imk}) = -w_{\theta,imk} \boldsymbol{Z}_{imk}$, the starred quantities being constants) gives

$$\boldsymbol{B}_{\boldsymbol{\gamma}}^{(m)} := \partial_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)} = \sum_{\substack{i<j \\ m \notin \{i,j\}}} \sum_{k} w_{\eta,ijk} f'(\Delta_{ij}) \, \boldsymbol{\omega}_{ij} \boldsymbol{Y}_{ijk}^{\top} + \sum_{i \ne m} \sum_{k} w_{\eta,imk} f'(\Delta_{im}) \, \boldsymbol{\omega}_{im} \boldsymbol{Y}_{imk}^{\top},$$

and $\boldsymbol{B}_{\boldsymbol{\beta}}^{(m)}$ analogously with $g^*_{imk} w_{\theta,imk}$ in place of $c_{ijk} w_{\theta,ijk}$. Every column again lies in $\mathbf{1}^{\perp}$, each coefficient obeys the same bounds ($|w_{\eta,imk} f'(\Delta_{im})| \le \frac{1}{4} F_1$, $|g^*_{imk} w_{\theta,imk}| \le \frac{1}{4}$, with $g^*_{imk} \le 1$ replacing $c_{ijk} \in [0,1]$), and the sums range over each unordered pair exactly once with $\le m^*$ comparisons, so the Cauchy–Schwarz argument applies verbatim and yields the identical bound $C_M n^{3/2}$. □

Lemma 9 (Cross-block Lipschitz bound)**.** *Assume A1–A3 and $n \ge 2$. Write $\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) := [\, \boldsymbol{B}_{\boldsymbol{\gamma}} \mid \boldsymbol{B}_{\boldsymbol{\beta}} \,]$ for the merit–structural cross block of* (B.19)–(B.20)*. There is an absolute constant L_B such that, for all $\boldsymbol{s}, \boldsymbol{s}'$ with $\|\boldsymbol{s}\|_\infty, \|\boldsymbol{s}'\|_\infty \le C_s + \frac{1}{2}$ and all $\boldsymbol{\zeta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$, $\boldsymbol{\zeta}' = (\boldsymbol{\gamma}', \boldsymbol{\beta}')$,*

$$\|\!|\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{\zeta}) - \boldsymbol{B}(\boldsymbol{s}', \boldsymbol{\zeta}')|\!\|_2 \;\le\; L_B \, n^{3/2} \Big(\|\boldsymbol{s} - \boldsymbol{s}'\|_\infty + \|\boldsymbol{\zeta} - \boldsymbol{\zeta}'\|_2 \Big).$$

Proof. Both blocks share the structure $\sum_{i<j} \sum_k \phi_{ijk} \, \boldsymbol{\omega}_{ij} \boldsymbol{v}_{ijk}^{\top}$ with a scalar per-comparison coefficient ϕ_{ijk} and a fixed covariate $\boldsymbol{v}_{ijk}$: for $\boldsymbol{B}_{\boldsymbol{\gamma}}$, $\phi_{ijk} = w_{\eta,ijk} f'(\Delta_{ij})$ and $\boldsymbol{v}_{ijk} = \boldsymbol{Y}_{ijk}$; for $\boldsymbol{B}_{\boldsymbol{\beta}}$, $\phi_{ijk} = c_{ijk} w_{\theta,ijk}$ and $\boldsymbol{v}_{ijk} = \boldsymbol{Z}_{ijk}$. The difference of two such blocks has the same structure with coefficient $\delta_{ijk} := \phi_{ijk}(\boldsymbol{s}, \boldsymbol{\zeta}) - \phi_{ijk}(\boldsymbol{s}', \boldsymbol{\zeta}')$, so the Cauchy–Schwarz factorization of Lemma 8 applies verbatim: for unit $\boldsymbol{u} \in \mathbf{1}^{\perp}$ and unit $\boldsymbol{v}$,

$$\begin{aligned} \big|\boldsymbol{u}^{\top} \big(\boldsymbol{B}_{\boldsymbol{\gamma}}(\boldsymbol{s}, \boldsymbol{\zeta}) - \boldsymbol{B}_{\boldsymbol{\gamma}}(\boldsymbol{s}', \boldsymbol{\zeta}')\big) \boldsymbol{v}\big| &\le \Big(\sum_{i<j,k} \delta_{ijk}^2 (u_i - u_j)^2 \Big)^{1/2} \Big(\sum_{i<j,k} (\boldsymbol{Y}_{ijk}^{\top} \boldsymbol{v})^2 \Big)^{1/2} \\ &\le \max_{i<j,k} |\delta_{ijk}| \cdot \frac{1}{\sqrt{2}} \, \kappa \, m^* \, (n+1)^{3/2}. \end{aligned}$$

It remains to bound the coefficient increments. For $\boldsymbol{B}_{\boldsymbol{\gamma}}$: on the working region,

$$\Big|\frac{\partial(w_\eta f')}{\partial\Delta}\Big| = \big|w_\eta(1-2g)f'^2 + w_\eta f''\big| \le \frac{1}{4}(F_1^2+F_2),$$
$$\|\nabla_{\boldsymbol{\gamma}}(w_\eta f')\|_2 = |w_\eta(1-2g)f'|\ \|\boldsymbol{Y}_{ijk}\|_2 \le \frac{1}{4}F_1\kappa,$$

and $\nabla_{\boldsymbol{\beta}}(w_\eta f') = \mathbf{0}$; with $|\Delta_{ij}(\boldsymbol{s}) - \Delta_{ij}(\boldsymbol{s}')| \le 2\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty$ and the mean value theorem along the segment (which stays in the working region, the ℓ_∞ ball being convex),

$$|\delta_{ijk}| \ \le\ \frac{1}{2}(F_1^2+F_2)\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \frac{1}{4}F_1\kappa\ \|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2\,.$$

For $\boldsymbol{B}_{\boldsymbol{\beta}}$: at fixed data $c_{ijk}\in\{0,1\}$, $|\partial(c\,w_\theta)/\partial\Delta| = |c\,w_\theta(1-2d)| \le \frac{1}{4}$ and $\|\nabla_{\boldsymbol{\beta}}(c\,w_\theta)\|_2 \le \frac{1}{4}\kappa$, with $\nabla_{\boldsymbol{\gamma}}(c\,w_\theta)=\mathbf{0}$, giving $|\delta_{ijk}| \le \frac{1}{2}\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \frac{1}{4}\kappa\,\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2$. Combining the two blocks by $|||[\,\boldsymbol{B}_{\boldsymbol{\gamma}}\mid\boldsymbol{B}_{\boldsymbol{\beta}}\,]|||_2 \le |||\boldsymbol{B}_{\boldsymbol{\gamma}}|||_2 + |||\boldsymbol{B}_{\boldsymbol{\beta}}|||_2$ and using $(n+1)^{3/2}\le 2n^{3/2}$ for $n\ge 2$,

$$|||\boldsymbol{B}(\boldsymbol{s},\boldsymbol{\zeta}) - \boldsymbol{B}(\boldsymbol{s}',\boldsymbol{\zeta}')|||_2 \ \le\ L_B\,n^{3/2}\Big(\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|_2\,\Big),$$

with $L_B := \sqrt{2}\,\kappa\,m^*\big(F_1^2+F_2+F_1\kappa+1+\kappa\big)$, an absolute constant under A1–A3. □

Lemma 10 (Structural-block Lipschitz bound)**.** *Assume A1–A3. Write* $\boldsymbol{A} := \mathrm{diag}(\boldsymbol{A}_{\boldsymbol{\gamma\gamma}}, \boldsymbol{A}_{\boldsymbol{\beta\beta}})$ *for the block-diagonal structural Hessian of* (B.17) *and* (B.18)*. There is an absolute constant* L_A *such that, for all* $\boldsymbol{s},\boldsymbol{s}'$ *with* $\|\boldsymbol{s}\|_\infty, \|\boldsymbol{s}'\|_\infty \le C_s + \frac{1}{2}$ *and all* $\boldsymbol{\zeta}=(\boldsymbol{\gamma},\boldsymbol{\beta})$, $\boldsymbol{\zeta}'=(\boldsymbol{\gamma}',\boldsymbol{\beta}')$,

$$|||\boldsymbol{A}(\boldsymbol{s},\boldsymbol{\zeta}) - \boldsymbol{A}(\boldsymbol{s}',\boldsymbol{\zeta}')|||_2 \ \le\ L_A\,n^2\Big(\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|_2\,\Big).$$

Proof. $\boldsymbol{A}$ is block-diagonal, so it suffices to bound each block. For the gate block, each weight $w_{\eta,ijk} = g(1-g)$ obeys, on the working region,

$$|\partial_\Delta w_{\eta,ijk}| = |w_\eta(1-2g)f'(\Delta_{ij})| \le \frac{1}{4}F_1, \qquad \|\nabla_{\boldsymbol{\gamma}} w_{\eta,ijk}\|_2 = |w_\eta(1-2g)|\ \|\boldsymbol{Y}_{ijk}\|_2 \le \frac{1}{4}\kappa,$$

and $\nabla_{\boldsymbol{\beta}} w_{\eta,ijk} = \mathbf{0}$; with $|\Delta_{ij}(\boldsymbol{s}) - \Delta_{ij}(\boldsymbol{s}')| \le 2\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty$ and the mean value theorem along the segment (contained in the working region by convexity),

$$|w_{\eta,ijk}(\boldsymbol{s},\boldsymbol{\gamma}) - w_{\eta,ijk}(\boldsymbol{s}',\boldsymbol{\gamma}')| \le \frac{1}{2}F_1\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \frac{1}{4}\kappa\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2\,.$$

Since $\Big\|\boldsymbol{Y}_{ijk}\boldsymbol{Y}_{ijk}^\top\Big\|_2 \le \kappa^2$ and there are $\le \binom{n+1}{2}m^* \le n^2 m^*$ summands, the triangle inequality gives

$$|||\boldsymbol{A}_{\boldsymbol{\gamma\gamma}}(\boldsymbol{s},\boldsymbol{\gamma}) - \boldsymbol{A}_{\boldsymbol{\gamma\gamma}}(\boldsymbol{s}',\boldsymbol{\gamma}')|||_2 \le \kappa^2 m^* n^2\big(\frac{1}{2}F_1\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty + \frac{1}{4}\kappa\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2\,\big).$$

For the direction block, at fixed data $c_{ijk}\in\{0,1\}$, $|\partial_\Delta(c\,w_\theta)| \le \frac{1}{4}$, $\|\nabla_{\boldsymbol{\beta}}(c\,w_\theta)\|_2 \le \frac{1}{4}\kappa$, $\nabla_{\boldsymbol{\gamma}}(c\,w_\theta) = \mathbf{0}$, and the same assembly applies with $\boldsymbol{Z}_{ijk}$ in place of $\boldsymbol{Y}_{ijk}$. Summing the two blocks yields the claim with $L_A := \kappa^2 m^*\max\{F_1+1,\ \kappa\}$. □

Lemma 11 (Upper curvature bound)**.** *Assume A1–A3. With* $W_{\max} := \frac{1}{4}F_1^2 + \frac{1}{4} + F_2$ *and* $C_{\max} := 2m^* W_{\max}$*, the merit Hessian satisfies, for* every $\boldsymbol{s}$ *with* $\|\boldsymbol{s}\|_\infty \le C_s + \frac{1}{2}$ *and every fixed* $(\boldsymbol{\gamma},\boldsymbol{\beta})$*,*

$$|||\boldsymbol{J}(\boldsymbol{s},\boldsymbol{\gamma},\boldsymbol{\beta})|||_2 \ \le\ C_{\max}\,n.$$

The same bound holds for the leave-one-out Hessian $\boldsymbol{J}^{(m)}$ *of* (B.28)*.*

Proof. $\boldsymbol{J}$ is the graph Laplacian with edge weights $W_{ij} = \sum_k W_{ijk}$, where, from (B.15), the first two terms of W_{ijk} are nonnegative and the saboteur is bounded by F_2, so

$$-W_{\max} \le\ -|c_{ijk}-g_{ijk}|\,|f''(\Delta_{ij})| \le\ W_{ijk} \le\ \frac{1}{4}F_1^2 + \frac{1}{4} + F_2 = W_{\max}, \qquad \text{i.e.} \quad |W_{ijk}| \le W_{\max}.$$

For a graph Laplacian $\boldsymbol{J} = \sum_{i<j} W_{ij}\, \boldsymbol{\omega}_{ij}\boldsymbol{\omega}_{ij}^\top$ with weights W_{ij}, row a has diagonal $\sum_{b\neq a} W_{ab}$ and off-diagonals $-W_{ab}$, so by Gershgorin (Fact 5) every eigenvalue z lies in a disc of centre $\sum_{b\neq a} W_{ab}$ and radius $\sum_{b\neq a} |W_{ab}|$ for some a; since $\boldsymbol{J}$ is symmetric, its eigenvalues are real and

$$\|\!|\boldsymbol{J}|\!\|_2 \;=\; \max|z| \;\le\; \max_a \Big(\Big|\sum_{b\neq a} W_{ab}\Big| + \sum_{b\neq a} |W_{ab}|\Big) \;\le\; 2\max_a \sum_{b\neq a} |W_{ab}|\,.$$

Each $W_{ab} = \sum_k W_{abk}$ is a sum of $m_{ab} \le m^*$ comparisons, each $|W_{abk}| \le W_{\max}$, with at most n neighbours b, so $\sum_{b\neq a} |W_{ab}| \le m^* n\, W_{\max}$. Hence $\|\!|\boldsymbol{J}|\!\|_2 \le 2m^* W_{\max}\, n = C_{\max}\, n$. For $\boldsymbol{J}^{(m)}$ the m-edge weight also satisfies $\left|W^{(m)}_{imk}\right| = \left|w_\eta f'^2 + g^* w_\theta - (g^* - g) f''\right| \le \frac{1}{4}F_1^2 + \frac{1}{4} + F_2 = W_{\max}$, so the identical bound holds. □

Lemma 12 (Local positive-definiteness of the Hessian)**.** *Assume A1–A3, with b_0 as in Lemma 1. For any sequences $\varepsilon_n, \rho_n = o(1)$ the following holds: for every $c > 0$ there exists n_0 such that for all $n \ge n_0$, on $\mathcal{E}_{\mathrm{sab}}$ as in Lemma 4, for every $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le \rho_n$ and every $\boldsymbol{s} \in \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)}$,*

$$\lambda_{\min,\perp}\big(\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta})\big) \;\ge\; \frac{1}{2}\, n b_0.$$

Proof. We establish the bound at $\boldsymbol{s}^*$, where the only randomness is used, then extend to the ball deterministically.

Step 1: positive definiteness at $\boldsymbol{s}^*$. With $\boldsymbol{J}^* = \boldsymbol{V} + \boldsymbol{E}$ as in (B.24), Lemma 2 gives $\lambda_{\min,\perp}(\boldsymbol{V}) \ge n b_0$ and Lemma 4 gives $\|\!|\boldsymbol{E}|\!\|_2 \le C_E \sqrt{n \log n}$ on an event of probability at least $1 - O(n^{-c})$. Weyl's inequality on the centered subspace yields

$$\lambda_{\min,\perp}(\boldsymbol{J}^*) = \lambda_{\min,\perp}(\boldsymbol{V} + \boldsymbol{E}) \;\ge\; \lambda_{\min,\perp}(\boldsymbol{V}) - \|\!|\boldsymbol{E}|\!\|_2 \;\ge\; n b_0 - C_E \sqrt{n \log n}.$$

Step 2: extension to the ball. Fix $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ in the stated balls and $\boldsymbol{s} \in \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)}$. For any centered unit vector $\boldsymbol{z}$ ($\mathbf{1}^\top \boldsymbol{z} = 0$, $\|\boldsymbol{z}\|_2 = 1$),

$$\boldsymbol{z}^\top \boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta})\, \boldsymbol{z} \;\ge\; \lambda_{\min,\perp}(\boldsymbol{J}^*) - \|\!|\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}^*|\!\|_2\,.$$

Using $\|\!|\cdot|\!\|_2 \le \|\!|\cdot|\!\|_\infty$ for symmetric matrices, the triangle inequality, Lemma 6 (Lipschitz in $\boldsymbol{s}$) and Lemma 7 (Lipschitz in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$),

$$\begin{aligned}
\|\!|\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}^*|\!\|_2 &\le \|\!|\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}(\boldsymbol{s}^*, \boldsymbol{\gamma}, \boldsymbol{\beta})|\!\|_\infty + \|\!|\boldsymbol{J}(\boldsymbol{s}^*, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \boldsymbol{J}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)|\!\|_\infty \\
&\le 4L_f m^* n\, \|\boldsymbol{s} - \boldsymbol{s}^*\|_\infty + L_\varsigma n\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\big) \\
&\le 4L_f m^* n\, \varepsilon_n + 2L_\varsigma n\, \rho_n \;=\; o(n),
\end{aligned}$$

since $\varepsilon_n, \rho_n = o(1)$ (the segment $[\boldsymbol{s}^*, \boldsymbol{s}]$ lies in the working region for n large, so Lemma 6 applies).

Conclusion. On the event above,

$$\lambda_{\min,\perp}\big(\boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta})\big) = \inf_{\boldsymbol{z}:\, \mathbf{1}^\top \boldsymbol{z} = 0,\ \|\boldsymbol{z}\|_2 = 1} \boldsymbol{z}^\top \boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \boldsymbol{z} \;\ge\; n b_0 - C_E \sqrt{n \log n} - o(n) \;\ge\; \frac{1}{2}\, n b_0$$

for $n \ge n_0$, since $\sqrt{n \log n} = o(n)$. □

B.11 Stage 1: existence of a local consistent estimator and ℓ_∞ rates

Recall the merit score map $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}} = \nabla_{\boldsymbol{s}}(-\ell)$ from (B.14), valued in $\mathbf{1}^\perp$. For fixed $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ we write $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ for a zero of $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ in $\mathbf{1}^\perp$; its existence and uniqueness near $\boldsymbol{s}^*$ are the content of Lemma 13.

Lemma 13 (Stage-1 ℓ_∞ rate)**.** *Assume A1–A3. Fix any constant $C_{\gamma\beta} \ge 0$ and put*

$$\rho_n := C_{\gamma\beta}\sqrt{\frac{\log n}{n}}, \qquad \varepsilon_n := C_5\sqrt{\frac{\log n}{n}}, \qquad C_5 := C_4 + 1,$$

with $C_4 = C_4(C_{\gamma\beta})$ the constant of Lemma 14 below. Then for every $c > 0$ there is an integer $n_0 = n_0(c)$ and an event of probability at least $1 - O(n^{-c})$ on which, for all $n \geq n_0$ and every $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with $\max\{\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\} \leq \rho_n$, the merit equation $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}) = \mathbf{0}$ has a solution $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$, unique in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$, and

$$\left\|\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} - \boldsymbol{s}^*\right\|_\infty \;\leq\; C_5\sqrt{\frac{\log n}{n}}.$$

The proof rests on the leave-one-out technique. Fix $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ near the truth and run gradient descent on $-\ell(\cdot, \boldsymbol{\gamma}, \boldsymbol{\beta})$, initialized at $\boldsymbol{s}^*$ (Algorithm 1); for each item m, run the same scheme on the LOO objective $\ell^{(m)}$ of (B.26) (Algorithm 2).

Algorithm 1 Gradient descent on the merit block (fixed $\boldsymbol{\gamma}, \boldsymbol{\beta}$)

Require: step size $\eta = (C_{\max} n)^{-1}$, fixed $(\boldsymbol{\gamma}, \boldsymbol{\beta})$
1: $\boldsymbol{s}^0 \leftarrow \boldsymbol{s}^*$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $\quad \boldsymbol{s}^{t+1} \leftarrow \boldsymbol{s}^t - \eta\, \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^t)$ ▷ $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}} = \nabla_{\boldsymbol{s}}(-\ell)$ is the merit score (B.14)
4: **end for**
5: **return** $\boldsymbol{s}^T$

Algorithm 2 LOO gradient descent on the merit block (fixed $\boldsymbol{\gamma}, \boldsymbol{\beta}$, item m)

Require: step size $\eta = (C_{\max} n)^{-1}$, fixed $(\boldsymbol{\gamma}, \boldsymbol{\beta})$
1: $\boldsymbol{s}^{0,(m)} \leftarrow \boldsymbol{s}^*$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: $\quad \boldsymbol{s}^{t+1,(m)} \leftarrow \boldsymbol{s}^{t,(m)} - \eta\, \boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)})$ ▷ $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ as in (B.27)
4: **end for**
5: **return** $\boldsymbol{s}^{T,(m)}$

The gauge is preserved automatically. Because $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ and $\boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ are valued in $\mathbf{1}^\perp$ (every $\boldsymbol{\omega}_{ij}$ is centered) and both trajectories start at $\boldsymbol{s}^0 = \boldsymbol{s}^{0,(m)} = \boldsymbol{s}^* \in \mathbf{1}^\perp$, every iterate $\boldsymbol{s}^t, \boldsymbol{s}^{t,(m)}$ remains in $\mathbf{1}^\perp$; no explicit projection is required. In particular $\boldsymbol{s}^t - \boldsymbol{s}^* \in \mathbf{1}^\perp$ and $\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)} \in \mathbf{1}^\perp$ for all t.

Two facts drive the argument. First, because the realized merit Hessian is positive definite on a neighborhood of $\boldsymbol{s}^*$ (Lemma 12), the iterates contract toward the local maximizer $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ and, after $\Theta(\log n)$ steps, lie within negligible distance of it; it therefore suffices to control the iterates themselves. Second, the iterates remain close to their starting point $\boldsymbol{s}^*$, which yields consistency.

The difficulty is that this confinement is naturally controlled in ℓ_2, whereas the target is the sharper, per-coordinate ℓ_∞ rate: an ℓ_2 bound permits the entire error to concentrate on a single coordinate. To rule this out we use, for each coordinate m, the leave-one-out proxy of Appendix B.8: by (B.30) the proxy is $\mathcal{F}_{-m}$-measurable, which restores the concentration needed to bound the m-th coordinate individually. The genuine iterate and its proxy are then shown to remain close, and combining the two gives the uniform ℓ_∞ bound.

Concretely, the argument is a single induction carrying four coupled estimates simultaneously along the trajectory: the ℓ_2 confinement, the per-coordinate accuracy of the proxy, the closeness of iterate and proxy, and the ℓ_∞ bound, which is the invariant that keeps the trajectory inside the region where positive definiteness holds (Lemma 12) and where the LOO score bound of Lemma 5 is active (Remark 5).

Lemma 14 (Coupled trajectory induction)**.** *Assume A1–A3. Fix a horizon $T = T(n) \geq 1$, a constant $C_{\gamma\beta} \geq 0$, and a pair $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with*

$$\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2 \leq C_{\gamma\beta}\sqrt{\frac{\log n}{n}}, \qquad \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C_{\gamma\beta}\sqrt{\frac{\log n}{n}}.$$

Run Algorithm 1 from $\boldsymbol{s}^0 = \boldsymbol{s}^*$, and for each $m \in \{0, 1, \dots, n\}$ run Algorithm 2 from $\boldsymbol{s}^{0,(m)} = \boldsymbol{s}^*$, both with step size $\eta = (C_{\max} n)^{-1}$. Define the events

$$\mathcal{E}_{\text{score}} := \Big\{ \max_{0 \le a \le n} \big| G_a(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) \big| \le C_\psi \sqrt{n \log n} \Big\},$$

$$\mathcal{E}_{\text{sab}} := \big\{ \, \|\boldsymbol{E}\|_2 \le C_E \sqrt{n \log n} \, \big\},$$

$$\mathcal{E}_{\text{loo}}(\boldsymbol{\gamma}, \boldsymbol{\beta}) := \bigcap_{m=0}^{n} \bigcap_{t=0}^{T} \left(\Big\{ \Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)}) \Big\|_2 \le 2K \sqrt{n \log n} \Big\} \cup \Big\{ \Big\| \boldsymbol{s}^{t,(m)} - \boldsymbol{s}^* \Big\|_\infty > \frac{1}{2} \Big\} \right),$$

with $\boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ as in (B.30) *and K as in Lemma 5, and set $\mathcal{E} := \mathcal{E}_{\text{score}} \cap \mathcal{E}_{\text{sab}} \cap \mathcal{E}_{\text{loo}}(\boldsymbol{\gamma}, \boldsymbol{\beta})$. Then for every $c > 0$ there exist constants $C_1, C_2, C_3, C_4 > 0$, depending only on c, on the constants of A1–A3, and on $C_{\gamma\beta}$ (not on n or t), and an integer $n_0 = n_0(c)$, such that for all $n \ge n_0$,*

$$\mathbb{P}(\mathcal{E}) \;\ge\; 1 - O\big((T+1)\, n^{1-c}\big),$$

and on $\mathcal{E}$ the following hold simultaneously for all $0 \le t \le T$:

$$\big\| \boldsymbol{s}^t - \boldsymbol{s}^* \big\|_2 \le C_1 \sqrt{\log n}, \tag{B.33a}$$

$$\max_{0 \le m \le n} \big| \, s^{t,(m)}_m - s^*_m \, \big| \le C_2 \sqrt{\frac{\log n}{n}}, \tag{B.33b}$$

$$\max_{0 \le m \le n} \Big\| \boldsymbol{s}^t - \boldsymbol{s}^{t,(m)} \Big\|_2 \le C_3 \sqrt{\frac{\log n}{n}}, \tag{B.33c}$$

$$\big\| \boldsymbol{s}^t - \boldsymbol{s}^* \big\|_\infty \le C_4 \sqrt{\frac{\log n}{n}}. \tag{B.33d}$$

Proof. We proceed by induction on t. The constants are fixed at the outset, in the acyclic order $C_1 \to C_2 \to C_3 \to C_4$ recorded in the closing paragraph; they depend only on the stated quantities and not on n_0, so n_0 may then be chosen large enough for the finitely many $o(1)$ absorptions below (in particular, for every invocation of Lemma 12 at the radii $C_4\sqrt{\log n/n}$, $(C_3 + C_4)\sqrt{\log n/n}$ and ρ_n, all $o(1)$, whose defining event is $\mathcal{E}_{\text{sab}} \supseteq \mathcal{E}$). Let the induction hypothesis $\mathcal{H}(t)$ assert that (B.33a)–(B.33d) hold for all $0 \le \tau \le t$. Since both trajectories start at $\boldsymbol{s}^*$, $\mathcal{H}(0)$ holds trivially. Assume $\mathcal{H}(t)$ for some $0 \le t < T$. Throughout, $\boldsymbol{s}^t, \boldsymbol{s}^{t,(m)} \in \mathbf{1}^\perp$ and the differences $\boldsymbol{s}^t - \boldsymbol{s}^*$, $\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)} \in \mathbf{1}^\perp$ (gauge preservation). Note also that $\mathcal{H}(t)$ activates the LOO score bound: by (B.33c) and (B.33d), $\big\| \boldsymbol{s}^{\tau,(m)} - \boldsymbol{s}^* \big\|_\infty \le (C_3 + C_4)\sqrt{\log n/n} \le \frac{1}{2}$ for all $\tau \le t$ and $n \ge n_0$, so on $\mathcal{E}_{\text{loo}}$ the first alternative of the dichotomy is in force and

$$\Big\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}\big(\boldsymbol{s}^{\tau,(m)}\big) \Big\|_2 \le 2K \sqrt{n \log n}, \qquad 0 \le m \le n, \; 0 \le \tau \le t. \tag{B.34}$$

Step 1: $\big\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^* \big\|_2 \le C_1 \sqrt{\log n}$. By Algorithm 1,

$$\boldsymbol{s}^{t+1} - \boldsymbol{s}^* = \boldsymbol{s}^t - \boldsymbol{s}^* - \eta \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*) - \eta\big(\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^t) - \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*)\big).$$

Writing $\bar{\boldsymbol{J}} := \int_0^1 \boldsymbol{J}(\boldsymbol{s}^* + \tau(\boldsymbol{s}^t - \boldsymbol{s}^*))\, d\tau$, the fundamental theorem of calculus gives $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^t) - \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*) = \bar{\boldsymbol{J}}(\boldsymbol{s}^t - \boldsymbol{s}^*)$, so

$$\boldsymbol{s}^{t+1} - \boldsymbol{s}^* = (\boldsymbol{I} - \eta \bar{\boldsymbol{J}})(\boldsymbol{s}^t - \boldsymbol{s}^*) - \eta \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*).$$

By (B.33d) the segment $[\boldsymbol{s}^*, \boldsymbol{s}^t]$ lies in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, C_4\sqrt{\log n/n})}$, hence in the working region for $n \ge n_0$; Lemma 11 gives $\eta \le 1/\lambda_{\max}(\bar{\boldsymbol{J}})$ and Lemma 12 gives $\lambda_{\min,\perp}(\bar{\boldsymbol{J}}) \ge \frac{1}{2} n b_0$. Since $\boldsymbol{s}^t - \boldsymbol{s}^* \in \mathbf{1}^\perp$, (B.22) gives

$$\big\| \boldsymbol{I} - \eta \bar{\boldsymbol{J}} \big\|_{2,\perp} = 1 - \eta\, \lambda_{\min,\perp}(\bar{\boldsymbol{J}}) \le 1 - \frac{1}{2} \eta n b_0.$$

Hence, with (B.33a),

$$\big\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^* \big\|_2 \le \big(1 - \frac{1}{2}\eta n b_0\big) C_1 \sqrt{\log n} + \eta \, \big\| \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*) \big\|_2 .$$

For the residual, by the fundamental theorem of calculus in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$,

$$\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*) = \boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}(\boldsymbol{s}^*) + \Big(\int_0^1 \frac{\partial \boldsymbol{G}(\boldsymbol{s}^*, \cdot)}{\partial(\boldsymbol{\gamma}, \boldsymbol{\beta})} \big((\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) + \tau((\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*))\big) d\tau \Big) \big((\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\big),$$

so

$$\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*)\|_2 \le \|\boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}(\boldsymbol{s}^*)\|_2 + \sup_{(\boldsymbol{\gamma}',\boldsymbol{\beta}')} \left\| \frac{\partial \boldsymbol{G}(\boldsymbol{s}^*, \cdot)}{\partial(\boldsymbol{\gamma}, \boldsymbol{\beta})}(\boldsymbol{\gamma}', \boldsymbol{\beta}') \right\|_2 \|(\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\|_2 .$$

On $\mathcal{E}_{\text{score}}$, $\|\boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}(\boldsymbol{s}^*)\|_\infty \le C_\psi \sqrt{n \log n}$, hence $\|\boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}(\boldsymbol{s}^*)\|_2 \le \sqrt{n+1}\, C_\psi \sqrt{n \log n} \le \sqrt{2}\, C_\psi\, n\sqrt{\log n}$. By Lemma 8 (applied at $\boldsymbol{s}^*$, which lies in the working region, uniformly over all $(\boldsymbol{\gamma}', \boldsymbol{\beta}')$) the supremum is $\le C_M n^{3/2}$, and $\|(\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\|_2 \le \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le 2C_{\gamma\beta}\sqrt{\log n / n}$. Therefore

$$\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*)\|_2 \le C_6\, n\sqrt{\log n}, \qquad C_6 := \sqrt{2}\, C_\psi + 2C_M C_{\gamma\beta}, \tag{B.35}$$

and, with $\eta = (C_{\max} n)^{-1}$,

$$\big\|\boldsymbol{s}^{t+1} - \boldsymbol{s}^*\big\|_2 \le \big(1 - \frac{1}{2}\eta n b_0\big) C_1 \sqrt{\log n} + \eta\, C_6\, n\sqrt{\log n}.$$

This is $\le C_1\sqrt{\log n}$ provided

$$C_1 \;\ge\; \frac{2C_6}{b_0}.$$

Step 2: $\max_{0 \le m \le n} \big| s_m^{t+1,(m)} - s_m^* \big| \le C_2 \sqrt{\log n / n}$. Fix $m \in \{0, \dots, n\}$. By Algorithm 2, the m-th coordinate updates as $s_m^{t+1,(m)} = s_m^{t,(m)} - \eta\, [\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s}^{t,(m)})]_m$, where the only terms of the LOO score involving coordinate m are

$$[\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s})]_m = -\sum_{i \ne m} \sum_k \big[(g_{mik}^* - g_{mik}) f'(\Delta_{mi}) + g_{mik}^* (d_{mik}^* - d_{mik})\big].$$

Since $[\boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}^{(m)}(\boldsymbol{s}^*)]_m = 0$,

$$s_m^{t+1,(m)} - s_m^* = \big(s_m^{t,(m)} - s_m^*\big) - \eta \underbrace{\big[\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s}^{t,(m)}) - \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s}^*)\big]_m}_{=:S} - \eta \underbrace{\big[\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}(\boldsymbol{s}^*) - \boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}^{(m)}(\boldsymbol{s}^*)\big]_m}_{=:Z}.$$

Control of S. By the mean value theorem applied to the scalar map $[\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}]_m$, there is $\bar{\boldsymbol{s}}$ on $[\boldsymbol{s}^*, \boldsymbol{s}^{t,(m)}]$ with $S = \big(\nabla_{\boldsymbol{s}} [\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}]_m(\bar{\boldsymbol{s}})\big)^\top (\boldsymbol{s}^{t,(m)} - \boldsymbol{s}^*)$. Since $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)} = \nabla_{\boldsymbol{s}}(-\ell^{(m)})$, one has $\nabla_{\boldsymbol{s}} [\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}^{(m)}]_m = (\boldsymbol{J}^{(m)})_{m\cdot}$, the m-th row of (B.28). Hence

$$S = (\boldsymbol{J}^{(m)})_{mm}(\bar{\boldsymbol{s}}) \big(s_m^{t,(m)} - s_m^*\big) + \sum_{i \ne m} (\boldsymbol{J}^{(m)})_{mi}(\bar{\boldsymbol{s}}) \big(s_i^{t,(m)} - s_i^*\big).$$

At the *true* parameters the saboteur in (B.29) vanishes, hence $W_{imk}^{(m)}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) = \bar{w}_{imk} \ge b_0$ (Lemma 1 and Remark 4) and, summing over the $\ge n$ comparisons meeting m,

$$\big(\boldsymbol{J}^{(m)}\big)_{mm}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) \ge n b_0.$$

At the working pair $(\boldsymbol{\gamma}, \boldsymbol{\beta}) \ne (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$ the saboteur does *not* vanish at $\boldsymbol{s}^*$, so we instead transfer the floor by perturbation. By (B.33c) and (B.33d) the segment point $\bar{\boldsymbol{s}}$ lies in $\overline{\mathcal{B}_\infty\big(\boldsymbol{s}^*, (C_3 + C_4)\sqrt{\log n / n}\big)}$ (the ball being convex), hence in the working region for $n \ge n_0$. Since any diagonal entry of a symmetric matrix $\boldsymbol{A}$ satisfies $|A_{mm}| \le \|\boldsymbol{A}\|_\infty$, the leave-one-out variants of Lemma 6 and Lemma 7, applied at $\bar{\boldsymbol{s}}$ with $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le C_{\gamma\beta}\sqrt{\log n / n}$, give

$$\Big| \big(\boldsymbol{J}^{(m)}\big)_{mm}(\bar{\boldsymbol{s}}, \boldsymbol{\gamma}, \boldsymbol{\beta}) - \big(\boldsymbol{J}^{(m)}\big)_{mm}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) \Big|$$

$$\le 4L_f m^* n \, \|\bar{\boldsymbol{s}} - \boldsymbol{s}^*\|_\infty + L_\zeta n\big(\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\big)$$
$$\le \big[\, 4L_f m^*(C_3 + C_4) + 2L_\zeta C_{\gamma\beta} \,\big]\sqrt{n \log n} = o(n),$$

so that

$$\big(\boldsymbol{J}^{(m)}\big)_{mm}(\bar{\boldsymbol{s}}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \ge nb_0 - o(n) \;\ge\; \frac{1}{2}\, nb_0 \qquad \text{for } n \ge n_0.$$

For the off-diagonal part, $\big|(\boldsymbol{J}^{(m)})_{mi}(\bar{\boldsymbol{s}})\big| = \big|W^{(m)}_{mi}(\bar{\boldsymbol{s}})\big| \le m^* W_{\max}$ for each $i \ne m$, so

$$\Big\|(\boldsymbol{J}^{(m)})_{m,-m}(\bar{\boldsymbol{s}})\Big\|_2 := \Big(\sum_{i \ne m} \Big|(\boldsymbol{J}^{(m)})_{mi}(\bar{\boldsymbol{s}})\Big|^2\Big)^{1/2} \le \sqrt{n}\, m^* W_{\max} = \frac{C_{\max}}{2}\sqrt{n}.$$

By the triangle inequality with (B.33a) and (B.33c),

$$\big\|\boldsymbol{s}^{t,(m)} - \boldsymbol{s}^*\big\|_2 \le \Big\|\boldsymbol{s}^{t,(m)} - \boldsymbol{s}^t\Big\|_2 + \big\|\boldsymbol{s}^t - \boldsymbol{s}^*\big\|_2 \le C_3\sqrt{\frac{\log n}{n}} + C_1\sqrt{\log n} = (C_1 + o(1))\sqrt{\log n}.$$

Hence, by Cauchy–Schwarz,

$$\Big|\sum_{i \ne m} (\boldsymbol{J}^{(m)})_{mi}(\bar{\boldsymbol{s}})\,(s_i^{t,(m)} - s_i^*)\Big| \le \Big\|(\boldsymbol{J}^{(m)})_{m,-m}(\bar{\boldsymbol{s}})\Big\|_2 \,\big\|\boldsymbol{s}^{t,(m)} - \boldsymbol{s}^*\big\|_2 \le \frac{C_{\max}}{2}(C_1 + o(1))\sqrt{n \log n}.$$

Control of Z. By the mean value theorem in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ there is $(\boldsymbol{\gamma}', \boldsymbol{\beta}')$ on $[(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*), (\boldsymbol{\gamma}, \boldsymbol{\beta})]$ with $|Z| \le \big\|\partial_{(\boldsymbol{\gamma},\boldsymbol{\beta})}[\boldsymbol{G}^{(m)}]_m(\boldsymbol{\gamma}', \boldsymbol{\beta}')\big\|_2 \, \|(\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\|_2$. The per-coordinate Jacobian rows (evaluated at $\boldsymbol{s}^*$, in the working region) are

$$\partial_{\boldsymbol{\gamma}}[\boldsymbol{G}^{(m)}]_m = -\sum_{i \ne m,k} w_{\eta,mik} f'(\Delta^*_{mi}) \boldsymbol{Y}^\top_{mik}, \qquad \partial_{\boldsymbol{\beta}}[\boldsymbol{G}^{(m)}]_m = -\sum_{i \ne m,k} g^*_{mik} w_{\theta,mik} \boldsymbol{Z}^\top_{mik},$$

each of the $\le m^* n$ summands bounded in ℓ_2 norm by $\frac{1}{4}F_1\kappa$ (resp. $\frac{1}{4}\kappa$), so $\big\|\partial_{(\boldsymbol{\gamma},\boldsymbol{\beta})}[\boldsymbol{G}^{(m)}]_m\big\|_2 \le \frac{1}{4}(F_1 + 1)\kappa m^* n =: C' n$. With $\|(\boldsymbol{\gamma}, \boldsymbol{\beta}) - (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\|_2 \le 2C_{\gamma\beta}\sqrt{\log n / n}$,

$$|Z| \;\le\; 2C' C_{\gamma\beta}\, n\sqrt{\frac{\log n}{n}} = 2C' C_{\gamma\beta}\sqrt{n \log n}.$$

Note also the ceiling: $(\boldsymbol{J}^{(m)})_{mm}(\bar{\boldsymbol{s}}) = \sum_{i \ne m,k} W^{(m)}_{imk}(\bar{\boldsymbol{s}}) \le m^* n\, W_{\max} = \frac{1}{2} C_{\max} n$, so $\eta\,(\boldsymbol{J}^{(m)})_{mm}(\bar{\boldsymbol{s}}) \le \frac{1}{2} < 1$; combined with the floor $(\boldsymbol{J}^{(m)})_{mm}(\bar{\boldsymbol{s}}) \ge \frac{1}{2} nb_0$, the coefficient of the leading term below satisfies $0 < 1 - \eta(\boldsymbol{J}^{(m)})_{mm}(\bar{\boldsymbol{s}}) \le 1 - \frac{1}{2}\eta n b_0$. Collecting the contributions,

$$\big|s_m^{t+1,(m)} - s_m^*\big| \le \big(1 - \frac{1}{2}\eta n b_0\big)\big|s_m^{t,(m)} - s_m^*\big| + \frac{\eta C_{\max}}{2}(C_1 + o(1))\sqrt{n \log n} + \eta\, 2C' C_{\gamma\beta}\sqrt{n \log n}.$$

By (B.33b) and $\eta = (C_{\max} n)^{-1}$, and using $C_1 + o(1) \le C_1 + 1$ for $n \ge n_0$, the right-hand side is $\le C_2\sqrt{\log n / n}$ provided

$$C_2 \ge \frac{C_{\max}(C_1 + 1) + 4C' C_{\gamma\beta}}{b_0}.$$

Step 3: $\max_{0 \le m \le n} \big\|\boldsymbol{s}^{t+1} - \boldsymbol{s}^{t+1,(m)}\big\|_2 \le C_3\sqrt{\log n / n}$. Writing the two updates and inserting $\pm \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)})$,

$$\boldsymbol{s}^{t+1} - \boldsymbol{s}^{t+1,(m)} = \big(\boldsymbol{I} - \eta\bar{\boldsymbol{J}}\big)(\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)}) - \eta\big(\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)}) - \boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)})\big),$$

where now $\bar{\boldsymbol{J}} := \int_0^1 \boldsymbol{J}(\boldsymbol{s}^{t,(m)} + \tau(\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)}))\, d\tau$. By (B.33c) and (B.33d) the whole segment $[\boldsymbol{s}^{t,(m)}, \boldsymbol{s}^t]$ lies in $\mathcal{B}_\infty(\boldsymbol{s}^*, (C_3 + C_4)\sqrt{\log n / n})$, so Lemmas 11 and 12 apply to $\bar{\boldsymbol{J}}$ and, since $\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)} \in \mathbf{1}^\perp$, (B.22) gives $\big\|\boldsymbol{I} - \eta\bar{\boldsymbol{J}}\big\|_{2,\perp} = 1 - \eta\lambda_{\min,\perp}(\bar{\boldsymbol{J}}) \le 1 - \frac{1}{2}\eta n b_0$. For the second term, the identity (B.30) and the activated LOO bound (B.34) give $\Big\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)}) - \boldsymbol{G}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)})\Big\|_2 = \Big\|\boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)})\Big\|_2 \le 2K\sqrt{n \log n}$ on $\mathcal{E}_{\mathrm{loo}}$. Hence

$$\Big\|\boldsymbol{s}^{t+1} - \boldsymbol{s}^{t+1,(m)}\Big\|_2 \le \big(1 - \frac{1}{2}\eta n b_0\big)\,\Big\|\boldsymbol{s}^t - \boldsymbol{s}^{t,(m)}\Big\|_2 + 2\eta K\sqrt{n \log n}.$$

With $\eta = (C_{\max} n)^{-1}$, $\frac{1}{2}\eta n b_0 = \frac{b_0}{2C_{\max}}$ and $2\eta K\sqrt{n \log n} = \frac{2K}{C_{\max}}\sqrt{\log n / n}$; using (B.33c),

$$\left\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^{t+1,(m)} \right\|_2 \le \Big((1 - \frac{b_0}{2C_{\max}}) C_3 + \frac{2K}{C_{\max}} \Big) \sqrt{\frac{\log n}{n}}.$$

The bracket is $\le C_3$ iff $\frac{2K}{C_{\max}} \le \frac{b_0}{2C_{\max}} C_3$, i.e. $C_3 \ge \frac{4K}{b_0}$.

Step 4: $\left\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^* \right\|_\infty \le C_4 \sqrt{\log n / n}$. For each m,

$$\begin{aligned} \left| s_m^{t+1} - s_m^* \right| &\le \left| s_m^{t+1} - s_m^{t+1,(m)} \right| + \left| s_m^{t+1,(m)} - s_m^* \right| \\ &\le \max_{0 \le m \le n} \left\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^{t+1,(m)} \right\|_2 + \max_{0 \le m \le n} \left| s_m^{t+1,(m)} - s_m^* \right|, \end{aligned}$$

so by Steps 2–3, $\left\| \boldsymbol{s}^{t+1} - \boldsymbol{s}^* \right\|_\infty \le (C_2 + C_3)\sqrt{\log n / n} \le C_4 \sqrt{\log n / n}$ provided $C_4 \ge C_2 + C_3$.

Closing the constants and the probability. Choose, in order, $C_1 \ge 2C_6 / b_0$ with $C_6 = \sqrt{2}\, C_\psi + 2C_M C_{\gamma\beta}$, then $C_2 \ge (C_{\max}(C_1 + 1) + 4C' C_{\gamma\beta})/b_0$, then $C_3 \ge 4K / b_0$, then $C_4 \ge C_2 + C_3$; this is the acyclic order $C_1 \to C_2 \to C_3 \to C_4$, so the choices are consistent and depend only on the stated quantities. Each step preserved (B.33a)–(B.33d), completing the induction. Finally, by Lemma 3, $\mathbb{P}(\mathcal{E}_{\text{score}}) \ge 1 - O(n^{-c})$; by Lemma 4, $\mathbb{P}(\mathcal{E}_{\text{sab}}) \ge 1 - O(n^{-c})$; and for $\mathcal{E}_{\text{loo}}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, each iterate $\boldsymbol{s}^{t,(m)}$ is $\mathcal{F}_{-m}$-measurable (gauge plus Appendix B.8), and $\left\| \boldsymbol{s}^{t,(m)} - \boldsymbol{s}^* \right\|_\infty \le \frac{1}{2}$ implies $\left\| \boldsymbol{s}^{t,(m)} \right\|_\infty \le C_s + \frac{1}{2}$ (A2), so Lemma 5 bounds each term of the complement,

$$\mathbb{P}\Big(\Big\{ \left\| \boldsymbol{D}^{(m)}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^{t,(m)}) \right\|_2 > 2K\sqrt{n \log n} \Big\} \cap \Big\{ \left\| \boldsymbol{s}^{t,(m)} - \boldsymbol{s}^* \right\|_\infty \le \frac{1}{2} \Big\} \Big) = O(n^{-c}),$$

a fortiori at threshold $2K \ge K$; a union bound over the $(n+1)(T+1)$ pairs (m, t) gives $\mathbb{P}(\mathcal{E}_{\text{loo}}(\boldsymbol{\gamma}, \boldsymbol{\beta})) \ge 1 - O((T+1)n^{1-c})$. □

Lemma 14 holds at a fixed $(\boldsymbol{\gamma}, \boldsymbol{\beta})$. To obtain Lemma 13 we run Algorithm 1 for $T = \Theta(\log n)$ steps and feed the coupled bounds into a degree-theoretic existence argument. This requires the LOO event to hold uniformly over all $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ in the structural ball, which is the next lemma.

Lemma 15 (Uniformity of $\mathcal{E}_{\text{loo}}$ over the structural ball)**.** *Assume A1–A3, and let $T = T(n)$ be any horizon with $T \le C_T \log n$ for a constant C_T. For $\boldsymbol{\zeta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ in the balls of Lemma 14, let $\{\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta})\}_{t \le T}$ be the LOO trajectories of Algorithm 2, and write $C_4' := C_3 + C_4$ with C_3, C_4 the constants of Lemma 14. Then for every $c > 0$ there is $n_0 = n_0(c)$ such that for all $n \ge n_0$, on a single event $\widetilde{\mathcal{E}}_{\text{loo}}$ with $\mathbb{P}(\widetilde{\mathcal{E}}_{\text{loo}}) \ge 1 - O(n^{-c})$, the following hold simultaneously, with $\Theta := \mathcal{B}(\boldsymbol{\gamma}^*, \rho_n) \times \mathcal{B}(\boldsymbol{\beta}^*, \rho_n)$:*

(a) (uniform LOO bound) $\sup_{\boldsymbol{\zeta} \in \Theta} \max_{0 \le m \le n} \max_{0 \le t \le T} \left\| \boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta})\big) \right\|_2 \le 2K\sqrt{n \log n}$;

(b) (uniform confinement) $\sup_{\boldsymbol{\zeta} \in \Theta} \max_{0 \le m \le n} \max_{0 \le t \le T} \left\| \boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^* \right\|_\infty \le (C_4' + 1)\sqrt{\log n / n}$.

Moreover $\widetilde{\mathcal{E}}_{\text{loo}} \subseteq \mathcal{E}_{\text{score}} \cap \mathcal{E}_{\text{sab}}$, and on $\widetilde{\mathcal{E}}_{\text{loo}}$ the event $\mathcal{E}$ of Lemma 14 holds at every $\boldsymbol{\zeta}$ in the stated balls.

Proof. Write $p := p_\gamma + p_\beta$ (fixed). Throughout, "the working region" means $\{\boldsymbol{s} : \|\boldsymbol{s}\|_\infty \le C_s + \frac{1}{2}\}$, on which the bounds F_1, F_2, F_3 of A3 apply to every gap (Remark 1).

Step A: sensitivity bounds on the working region. Fix m and $\boldsymbol{\zeta}, \boldsymbol{\zeta}' \in \Theta$.

(i) *Point sensitivity of $\boldsymbol{D}^{(m)}$.* For $\boldsymbol{s}, \boldsymbol{s}'$ in the working region, each summand χ_{imk} of (B.30) is, at fixed $\boldsymbol{\zeta}$, a function of the gap Δ_{im} alone, with

$$\left| \frac{\partial \chi_{imk}}{\partial \Delta_{im}} \right| = \left| (c_{imk} - g^*_{imk})\big(f''(\Delta_{im}) - w_{\theta,imk}\big) \right| \le F_2 + \frac{1}{4} =: L_\chi$$

on the region; hence $|\chi_{imk}(\boldsymbol{s}) - \chi_{imk}(\boldsymbol{s}')| \le L_\chi\,|\Delta_{im}(\boldsymbol{s}) - \Delta_{im}(\boldsymbol{s}')| \le 2L_\chi\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty$. The difference $\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}) - \boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}')$ has m-th coordinate a sum of $\le m^* n$ increments and, for each $i \ne m$, an i-th coordinate consisting of one pair's $\le m^*$ increments; summing in ℓ_2 and using $\|\cdot\|_\infty \le \|\cdot\|_2$,

$$\left\|\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}) - \boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}')\right\|_2 \;\le\; 2L_\chi m^*\sqrt{n^2+n}\,\|\boldsymbol{s}-\boldsymbol{s}'\|_\infty \;\le\; C_D\,n\,\|\boldsymbol{s}-\boldsymbol{s}'\|_2$$

with $C_D := 4L_\chi m^*$.

(ii) *Coefficient sensitivity of $\boldsymbol{D}^{(m)}$.* At fixed $\boldsymbol{s}$ in the working region, $\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}$ depends on $\boldsymbol{\zeta}$ only through $d_{imk} = \sigma(\Delta_{im} + \boldsymbol{Z}_{imk}^\top\boldsymbol{\beta})$ (no $\boldsymbol{\gamma}$-dependence, by (B.30)); as σ is $\frac14$-Lipschitz and $\|\boldsymbol{Z}_{imk}\|_2 \le \kappa$, each summand moves by at most $|c_{imk} - g^*_{imk}| \cdot \frac14\kappa\,\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 \le \frac14\kappa\,\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|_2$, and the same ℓ_2 assembly as in (i) gives

$$\left\|\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}) - \boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}'}(\boldsymbol{s})\right\|_2 \;\le\; C_{\mathrm{coef}}\,n\,\|\boldsymbol{\zeta}-\boldsymbol{\zeta}'\|_2\,, \qquad C_{\mathrm{coef}} := \kappa m^*.$$

(iii) *One-step map.* Writing one LOO step as $\boldsymbol{s}^{t+1,(m)} = \Phi^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}^{t,(m)})$ with $\Phi^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s}) = \boldsymbol{s} - \eta\,\boldsymbol{G}^{(m)}_{\boldsymbol{\zeta}}(\boldsymbol{s})$: on the working region, $\partial_{\boldsymbol{s}}\Phi^{(m)}_{\boldsymbol{\zeta}} = \boldsymbol{I} - \eta\boldsymbol{J}^{(m)}$ has $\|\!|\cdot|\!\|_2 \le 2$ (by Lemma 11 and $\eta = (C_{\max}n)^{-1}$, the spectrum of the symmetric matrix $\eta\boldsymbol{J}^{(m)}$ lies in $[-1,1]$, hence that of $\boldsymbol{I}-\eta\boldsymbol{J}^{(m)}$ in $[0,2]$), and $\left\|\!\left|\partial_{\boldsymbol{\zeta}}\Phi^{(m)}_{\boldsymbol{\zeta}}\right|\!\right\|_2 = \eta\,\left\|\!\left|\partial_{\boldsymbol{\zeta}}\boldsymbol{G}^{(m)}_{\boldsymbol{\zeta}}\right|\!\right\|_2 \le \eta\,C_M n^{3/2} = C_M C_{\max}^{-1}\sqrt{n}$ (Lemma 8, leave-one-out variant). All three bounds require the argument(s) to lie in the working region; establishing that they do, along the trajectories at *arbitrary* $\boldsymbol{\zeta}\in\Theta$, is the purpose of Step C.

Step B: net, good event, and behavior at net points. Since $T \le C_T\log n$, we have $2^{\,T+1} \le n^{\kappa_T}$ with $\kappa_T := 2C_T$ for $n \ge n_0$. Let $\mathcal{N}\subset\Theta$ be a δ-net in ℓ_2 with

$$\delta := n^{-(\kappa_T+2)}, \qquad \text{so} \qquad |\mathcal{N}| \le \Big(\frac{3\rho_n}{\delta}\Big)^p = n^{O(1)}, \quad \log|\mathcal{N}| = O(\log n).$$

Define

$$\widetilde{\mathcal{E}}_{\mathrm{loo}} := \mathcal{E}_{\mathrm{score}} \cap \mathcal{E}_{\mathrm{sab}} \cap \bigcap_{\boldsymbol{\zeta}_j\in\mathcal{N}}\bigcap_{m=0}^{n}\bigcap_{t=0}^{T}\Big(\Big\{\Big\|\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}_j}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\big)\Big\|_2 \le K\sqrt{n\log n}\Big\} \cup \Big\{\Big\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j) - \boldsymbol{s}^*\Big\|_\infty > \frac12\Big\}\Big).$$

Each $\boldsymbol{\zeta}_j \in \mathcal{N}$ is deterministic, so $\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)$ is $\mathcal{F}_{-m}$-measurable (Appendix B.8), and Lemma 5 bounds the probability of each complementary event $\{\|\boldsymbol{D}^{(m)}_{\boldsymbol{\zeta}_j}\|_2 > K\sqrt{n\log n}\} \cap \{\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j) - \boldsymbol{s}^*\|_\infty \le \frac12\}$ by $O(n^{-c'})$: running the lemma at grade c' and union-bounding over $\mathcal{N}\times\{0,\dots,n\}\times\{0,\dots,T\}$, a set of cardinality $n^{O(1)}$ (with exponent depending only on p, κ_T), gives $\mathbb{P}(\widetilde{\mathcal{E}}_{\mathrm{loo}}) \ge 1 - O(n^{-c})$ for c' large enough (depending on c, p, C_T), after intersecting with the events of Lemmas 3 and 4.

On $\widetilde{\mathcal{E}}_{\mathrm{loo}}$, for every net point $\boldsymbol{\zeta}_j$ the event $\mathcal{E} = \mathcal{E}_{\mathrm{score}}\cap\mathcal{E}_{\mathrm{sab}}\cap\mathcal{E}_{\mathrm{loo}}(\boldsymbol{\zeta}_j)$ of Lemma 14 holds: the net dichotomy at threshold K implies the defining dichotomy at threshold $2K$ a fortiori (same escape set, weaker bound). Hence (B.33a)–(B.33d) are in force at every $\boldsymbol{\zeta}_j$ for all $t \le T$; combining (B.33c) (with $\|\cdot\|_\infty \le \|\cdot\|_2$) and (B.33d) yields the *net-point confinement of the LOO trajectories*,

$$\left\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j) - \boldsymbol{s}^*\right\|_\infty \le C_4'\sqrt{\log n/n}, \qquad \boldsymbol{\zeta}_j\in\mathcal{N},\ 0\le m\le n,\ 0\le t\le T, \tag{B.36}$$

with $C_4' = C_3 + C_4$ as in the statement.

Step C: confinement induction off the net. For $\boldsymbol{\zeta}\in\Theta$ fix a nearest net point $\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j(\boldsymbol{\zeta})$, $\|\boldsymbol{\zeta}-\boldsymbol{\zeta}_j\|_2 \le \delta$. Set $\Lambda_t := (2^{\,t}-1)\,C_M C_{\max}^{-1}\sqrt{n}$, and note

$$\Lambda_T\delta \le 2^T C_M C_{\max}^{-1}\sqrt{n}\,n^{-(\kappa_T+2)} \le C_M C_{\max}^{-1} n^{-3/2} \le \sqrt{\log n/n} \qquad (n\ge n_0). \tag{B.37}$$

We prove by induction on t the hypothesis

$$H(t): \qquad \left\|\boldsymbol{s}^{\tau,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^{\tau,(m)}(\boldsymbol{\zeta}_j)\right\|_2 \le \Lambda_\tau \delta \quad \text{for all } \boldsymbol{\zeta} \in \Theta,\ 0 \le m \le n,\ 0 \le \tau \le t.$$

$H(t)$ has the following consequence, by (B.36), (B.37) and $\|\cdot\|_\infty \le \|\cdot\|_2$: for all $\tau \le t$,

$$\left\|\boldsymbol{s}^{\tau,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^*\right\|_\infty \le C_4'\sqrt{\log n/n} + \Lambda_\tau\delta \le (C_4'+1)\sqrt{\log n/n}, \tag{B.38}$$

so that every point $\boldsymbol{s}^{\tau,(m)}(\boldsymbol{\zeta})$, every net-point iterate, and (by convexity of the ℓ_∞ ball) every segment between them lies in the working region for $n \ge n_0$.

$H(0)$ holds trivially, both trajectories starting at $\boldsymbol{s}^*$. Assume $H(t)$, $t < T$. Decompose one step as

$$\boldsymbol{s}^{t+1,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^{t+1,(m)}(\boldsymbol{\zeta}_j) = \underbrace{\Phi_{\boldsymbol{\zeta}}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta})\big) - \Phi_{\boldsymbol{\zeta}}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\big)}_{=:P} + \underbrace{\Phi_{\boldsymbol{\zeta}}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\big) - \Phi_{\boldsymbol{\zeta}_j}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\big)}_{=:Q}.$$

For P: by (B.38) the segment between the two arguments lies in the working region, so Step A(iii) and the mean value theorem give $\|P\|_2 \le 2\left\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\right\|_2 \le 2\Lambda_t\delta$. For Q: the argument $\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)$ lies in the working region by (B.36), and the segment $[\boldsymbol{\zeta}_j, \boldsymbol{\zeta}]$ lies in the convex set Θ, so Step A(iii) gives $\|Q\|_2 \le C_M C_{\max}^{-1}\sqrt{n}\,\delta$. Hence

$$\left\|\boldsymbol{s}^{t+1,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^{t+1,(m)}(\boldsymbol{\zeta}_j)\right\|_2 \ \le\ \big(2\Lambda_t + C_M C_{\max}^{-1}\sqrt{n}\big)\delta \ =\ \Lambda_{t+1}\,\delta,$$

by the recursion $\Lambda_{t+1} = 2\Lambda_t + C_M C_{\max}^{-1}\sqrt{n}$. This proves $H(t+1)$, and the induction delivers $H(T)$; consequence (B.38) at $\tau \le T$ is claim (b).

Step D: off-net extension of the LOO bound. Fix $\boldsymbol{\zeta} \in \Theta$, m, t, and $\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j(\boldsymbol{\zeta})$. By (B.36) the net-point iterate is confined, $\left\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j) - \boldsymbol{s}^*\right\|_\infty \le C_4'\sqrt{\log n/n} \le \frac{1}{2}$ for $n \ge n_0$, so on $\widetilde{\mathcal{E}}_{\mathrm{loo}}$ the first alternative of the net dichotomy is active: $\left\|\boldsymbol{D}_{\boldsymbol{\zeta}_j}^{(m)}(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j))\right\|_2 \le K\sqrt{n\log n}$. All remaining points lie in the working region by (B.36) and (B.38), so Steps A(i)–(ii) apply:

$$\left\|\boldsymbol{D}_{\boldsymbol{\zeta}}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta})\big)\right\|_2 \ \le\ \left\|\boldsymbol{D}_{\boldsymbol{\zeta}_j}^{(m)}\big(\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\big)\right\|_2 + C_D\,n\,\left\|\boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}) - \boldsymbol{s}^{t,(m)}(\boldsymbol{\zeta}_j)\right\|_2 + C_{\mathrm{coef}}\,n\,\|\boldsymbol{\zeta} - \boldsymbol{\zeta}_j\|_2.$$

The first term is $\le K\sqrt{n\log n}$; by $H(T)$ and (B.37) the second is $\le C_D\,n\cdot C_M C_{\max}^{-1} n^{-3/2} = O(n^{-1/2})$; the third is $\le C_{\mathrm{coef}}\,n^{1-(\kappa_T+2)} = O(n^{-1})$. For $n \ge n_0$ the last two are together $\le K\sqrt{n\log n}$, whence the total is $\le 2K\sqrt{n\log n}$. Taking suprema over $\boldsymbol{\zeta} \in \Theta$ and maxima over m, t gives claim (a). The final sentence of the statement restates that $\mathcal{E}$ of Lemma 14 holds at every $\boldsymbol{\zeta}$: its first two constituents hold by construction, and $\mathcal{E}_{\mathrm{loo}}(\boldsymbol{\zeta})$ holds because the actual bound of claim (a) implies the dichotomy at threshold $2K$ a fortiori. □

Proof of Lemma 13. Let $D \ge 1$ be a constant and set

$$\varrho := 1 - \frac{b_0}{2C_{\max}} \in (0,1), \qquad T := \left\lceil \frac{(D+2)\log n}{\log(1/\varrho)} \right\rceil = \Theta(\log n),$$

so that $T+1 \le C_T \log n$ with $C_T := (D+3)/\log(1/\varrho)$ for $n \ge n_0$. Set $\widetilde{\mathcal{E}} := \mathcal{E}_{\mathrm{score}} \cap \mathcal{E}_{\mathrm{sab}} \cap \widetilde{\mathcal{E}}_{\mathrm{loo}}$, with $\widetilde{\mathcal{E}}_{\mathrm{loo}}$ from Lemma 15 at horizon $T+1$. As $\mathcal{E}_{\mathrm{score}}, \mathcal{E}_{\mathrm{sab}}$ are evaluated at $(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$, the union bound gives $\mathbb{P}(\widetilde{\mathcal{E}}) \ge 1 - O(n^{-c})$.

Fix any $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with $\max\{\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_2, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\} \le \rho_n$. By the final assertion of Lemma 15, $\widetilde{\mathcal{E}}$ implies the event $\mathcal{E}$ of Lemma 14 *for this* $(\boldsymbol{\gamma}, \boldsymbol{\beta})$; run Algorithm 1 for $T+1$ steps, applying Lemma 14 with horizon $T+1$. All statements below hold on $\mathcal{E}$; the four bounds (B.33a)–(B.33d) are in force for $0 \le t \le T+1$.

Step 1 (near-stationarity). Since $\boldsymbol{s}^{t+1} - \boldsymbol{s}^t = -\eta \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^t)$, consecutive differences satisfy $\boldsymbol{s}^{t+1} - \boldsymbol{s}^t = (\boldsymbol{I} - \eta \bar{\boldsymbol{J}}_t)(\boldsymbol{s}^t - \boldsymbol{s}^{t-1})$ with $\bar{\boldsymbol{J}}_t := \int_0^1 \boldsymbol{J}(\boldsymbol{s}^{t-1} + \tau(\boldsymbol{s}^t - \boldsymbol{s}^{t-1}))\, d\tau$. By (B.33d) the segment $[\boldsymbol{s}^{t-1}, \boldsymbol{s}^t] \subset \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, C_4\sqrt{\log n/n})}$, so Lemma 11 gives $\eta \le 1/\lambda_{\max}(\bar{\boldsymbol{J}}_t)$ and Lemma 12 gives $\lambda_{\min,\perp}(\bar{\boldsymbol{J}}_t) \ge \frac{1}{2} n b_0$; with $\boldsymbol{s}^t - \boldsymbol{s}^{t-1} \in \mathbf{1}^\perp$, (B.22) gives $\big\|\big\|\boldsymbol{I} - \eta \bar{\boldsymbol{J}}_t\big\|\big\|_{2,\perp} \le \varrho$ for $1 \le t \le T+1$. Iterating from $\boldsymbol{s}^0 = \boldsymbol{s}^*$ and using $\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*)\|_2 \le C_6\, n\sqrt{\log n}$ on $\mathcal{E}$ (display (B.35) in the proof of Lemma 14, with $C_6 = \sqrt{2}\, C_\psi + 2 C_M C_{\gamma\beta}$),

$$\big\|\boldsymbol{s}^1 - \boldsymbol{s}^0\big\|_2 = \eta\, \|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^*)\|_2 \le \frac{C_6}{C_{\max}}\sqrt{\log n}, \qquad \big\|\boldsymbol{s}^{t+1} - \boldsymbol{s}^t\big\|_2 \le \varrho^{\,t}\, \frac{C_6}{C_{\max}}\sqrt{\log n}.$$

Therefore, using $\boldsymbol{s}^{T+1} - \boldsymbol{s}^T = -\eta \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T)$ and $\eta = (C_{\max} n)^{-1}$,

$$\big\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T)\big\|_2 = \frac{1}{\eta}\, \big\|\boldsymbol{s}^{T+1} - \boldsymbol{s}^T\big\|_2 \le C_6\, n\sqrt{\log n}\, \varrho^{\,T} \le n^{-D},$$

the last step by the choice of T (for n large).

Step 2 (existence, via degree theory on a small ball). If $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T) = \mathbf{0}$, set $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} := \boldsymbol{s}^T$ and pass directly to Step 3. Otherwise, put

$$r := \frac{4\, \big\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T)\big\|_2}{n b_0} \in \Big(0,\ \frac{4n^{-D}}{b_0}\Big], \qquad \Omega := \mathcal{B}_2(\boldsymbol{s}^T, r) \cap (\boldsymbol{s}^* + \mathbf{1}^\perp).$$

Because $\big\|\boldsymbol{s}^T - \boldsymbol{s}^*\big\|_\infty \le C_4\sqrt{\log n/n}$ (by (B.33d)) and $r = o(\sqrt{\log n/n})$, the whole ball satisfies $\overline{\Omega} \subset \mathcal{B}_\infty(\boldsymbol{s}^*, C_5\sqrt{\log n/n})$ for $n \ge n_0$, since $C_5 = C_4 + 1$. For $\boldsymbol{x} \in \partial\Omega$ write $\boldsymbol{v} := \boldsymbol{x} - \boldsymbol{s}^T \in \mathbf{1}^\perp$ (as $\boldsymbol{s}^T \in \mathbf{1}^\perp$), $\|\boldsymbol{v}\|_2 = r$, and $\bar{\boldsymbol{J}} := \int_0^1 \boldsymbol{J}(\boldsymbol{s}^T + \tau\boldsymbol{v})\, d\tau$. Lemma 12 applies to $\bar{\boldsymbol{J}}$ (segment inside $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)}$), so

$$\begin{aligned}\boldsymbol{v}^\top \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{x}) &= \boldsymbol{v}^\top \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T) + \boldsymbol{v}^\top \bar{\boldsymbol{J}}\, \boldsymbol{v} \ge -r\, \big\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T)\big\|_2 + \frac{1}{2} n b_0 r^2 \\ &= \frac{1}{2} n b_0\, r\Big(r - \frac{2\, \big\|\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{s}^T)\big\|_2}{n b_0}\Big) > 0.\end{aligned}$$

By the degree-theoretic existence criterion (Fact 4, with anchor $\boldsymbol{s}^T$ and $\boldsymbol{F} = \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ restricted to the affine slice $\boldsymbol{s}^* + \mathbf{1}^\perp$, identified with $\mathbb{R}^n$ and on which $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ takes values in $\mathbf{1}^\perp$), $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ has a zero $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} \in \overline{\Omega}$.

Step 3 (rate). Since $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} \in \overline{\mathcal{B}_2(\boldsymbol{s}^T, r)}$,

$$\|\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} - \boldsymbol{s}^*\|_\infty \le \big\|\boldsymbol{s}^T - \boldsymbol{s}^*\big\|_\infty + \big\|\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} - \boldsymbol{s}^T\big\|_\infty \le C_4\sqrt{\frac{\log n}{n}} + r \ \le\ C_5\sqrt{\frac{\log n}{n}}$$

for $n \ge n_0$, as $r = O(n^{-D}) = o(\sqrt{\log n/n})$ for $D \ge 1$ and $C_5 = C_4 + 1$; in particular $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}} \in \mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)$.

Step 4 (uniqueness). The set $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$ is convex, and for distinct $\boldsymbol{x}, \boldsymbol{y}$ in it, with $\bar{\boldsymbol{J}} := \int_0^1 \boldsymbol{J}(\boldsymbol{y} + \tau(\boldsymbol{x} - \boldsymbol{y}))\, d\tau$ and $\boldsymbol{x} - \boldsymbol{y} \in \mathbf{1}^\perp$, Lemma 12 gives

$$\big(\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{x}) - \boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{y})\big)^\top (\boldsymbol{x} - \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^\top \bar{\boldsymbol{J}}\, (\boldsymbol{x} - \boldsymbol{y}) \ge \frac{1}{2} n b_0\, \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 > 0.$$

Hence $\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ is strictly monotone there and has at most one zero; with Step 2, $\widehat{\boldsymbol{s}}_{\boldsymbol{\gamma},\boldsymbol{\beta}}$ is the unique zero in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$. □

B.12 Stage 2: the profiled likelihood and joint consistency

The population information. At the truth, the Fisher information of the model is the expected Hessian $\mathcal{I}_n := \mathbb{E}\big[\nabla^2(-\ell)(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)\big]$, whose blocks are obtained from (B.17)–(B.20) by replacing

c_{ijk} with g^*_{ijk} (the remaining quantities being deterministic at the truth) and whose merit block is $\boldsymbol{V}$ of (B.24). Explicitly, with

$$\boldsymbol{A}^{\mathrm{pop}} := \begin{pmatrix} \boldsymbol{A}_{\boldsymbol{\gamma\gamma}}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*) & 0 \\ 0 & \sum_{i<j} \sum_k g^*_{ijk} w^*_{\theta,ijk}\, \boldsymbol{Z}_{ijk} \boldsymbol{Z}_{ijk}^\top \end{pmatrix},$$
$$\boldsymbol{B}^{\mathrm{pop}} := \Big[\, \boldsymbol{B}_{\boldsymbol{\gamma}}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*) \;\Big|\; \sum_{i<j} \sum_k g^*_{ijk} w^*_{\theta,ijk}\, \boldsymbol{\omega}_{ij} \boldsymbol{Z}_{ijk}^\top \,\Big]$$

(note $\boldsymbol{A}_{\boldsymbol{\gamma\gamma}}$ and $\boldsymbol{B}_{\boldsymbol{\gamma}}$ carry no data, hence equal their population versions at the truth), define the *efficient information for the structural block* as the Schur complement of the merit block in $\mathcal{I}_n$,

$$\boldsymbol{\Sigma}^{\mathrm{pop}} \;:=\; \boldsymbol{A}^{\mathrm{pop}} - (\boldsymbol{B}^{\mathrm{pop}})^\top \boldsymbol{V}^{-1} \boldsymbol{B}^{\mathrm{pop}}, \tag{B.39}$$

with $\boldsymbol{V}^{-1}$ taken on $\mathbf{1}^\perp$ (legitimate by Lemma 2).

Assumption 6 (Efficient information floor)**.** There is a constant $c_\Sigma > 0$ such that $\lambda_{\min}\big(\boldsymbol{\Sigma}^{\mathrm{pop}}\big) \geq c_\Sigma\, n^2$ for all n large enough.

Remark 6. Since $(\boldsymbol{B}^{\mathrm{pop}})^\top \boldsymbol{V}^{-1} \boldsymbol{B}^{\mathrm{pop}} \succeq 0$ (the columns of $\boldsymbol{B}^{\mathrm{pop}}$ lie in $\mathbf{1}^\perp$, where $\boldsymbol{V} \succ 0$), A6 forces $\lambda_{\min}(\boldsymbol{A}^{\mathrm{pop}}) \geq c_\Sigma n^2$; as $\boldsymbol{A}^{\mathrm{pop}}$ is block-diagonal with blocks exactly the matrices of A4 and A5, Assumption A6 implies both with $c_0 = c_\Sigma$. We nevertheless retain A4–A5 as standing conditions for their interpretability. For asymptotic normality, which is beyond the scope of this appendix, A6 would be strengthened to the existence of the limit $\bar{\boldsymbol{\Sigma}} := \lim_n N^{-1} \boldsymbol{\Sigma}^{\mathrm{pop}} \succ 0$, $N = \binom{n+1}{2}$; only the floor is needed for consistency.

Lemma 16 (Information concentration at the truth)**.** *Assume A1–A3. For every $c > 0$ there is an absolute constant C_Σ and an event $\mathcal{E}_\Sigma$ with $\mathbb{P}(\mathcal{E}_\Sigma) \geq 1 - O(n^{-c})$ on which*

$$\Big\| \boldsymbol{A}_{\boldsymbol{\beta\beta}}(\boldsymbol{s}^*, \boldsymbol{\beta}^*) - \boldsymbol{A}^{\mathrm{pop}}_{\boldsymbol{\beta\beta}} \Big\|_2 \;\leq\; C_\Sigma\, n \sqrt{\log n}, \qquad \Big\| \boldsymbol{B}_{\boldsymbol{\beta}}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*, \boldsymbol{\beta}^*) - \boldsymbol{B}^{\mathrm{pop}}_{\boldsymbol{\beta}} \Big\|_2 \;\leq\; C_\Sigma\, n \sqrt{\log n},$$

where $\boldsymbol{A}^{\mathrm{pop}}_{\boldsymbol{\beta\beta}}$ and $\boldsymbol{B}^{\mathrm{pop}}_{\boldsymbol{\beta}}$ denote the direction blocks of $\boldsymbol{A}^{\mathrm{pop}}$ and $\boldsymbol{B}^{\mathrm{pop}}$.

Proof. Direction block of $\boldsymbol{A}$. Write $\boldsymbol{A}_{\boldsymbol{\beta\beta}}(\boldsymbol{s}^*, \boldsymbol{\beta}^*) - \boldsymbol{A}^{\mathrm{pop}}_{\boldsymbol{\beta\beta}} = \sum_{i<j,k} \boldsymbol{X}_{ijk}$ with $\boldsymbol{X}_{ijk} := (c_{ijk} - g^*_{ijk})\, w^*_{\theta,ijk}\, \boldsymbol{Z}_{ijk} \boldsymbol{Z}_{ijk}^\top$, independent, symmetric, mean-zero, of dimension p_β (fixed). Then $\|\boldsymbol{X}_{ijk}\|_2 \leq \frac{1}{4}\kappa^2 =: R$ and $\Big\| \sum \mathbb{E} \boldsymbol{X}_{ijk}^2 \Big\|_2 \leq \sum \Big\| \mathbb{E} \boldsymbol{X}_{ijk}^2 \Big\|_2 \leq \frac{1}{16} \kappa^4\, m^* n^2 =: \nu$, so matrix Bernstein (Fact 2) with $t = C\, n\sqrt{\log n}$ gives exponent $\lesssim -C^2 \log n / (2 \cdot \frac{1}{16}\kappa^4 m^*)$ for n large, hence probability $O(n^{-(c+1)})$ for C large enough.

Direction block of $\boldsymbol{B}$. Fix a column index $l \leq p_\beta$. The l-th column of the difference is $\sum_{i<j,k} \varsigma_{ijk}\, \boldsymbol{\omega}_{ij} \in \mathbf{1}^\perp$ with $\varsigma_{ijk} := (c_{ijk} - g^*_{ijk})\, w^*_{\theta,ijk}\, Z_{ijk,l}$, independent and mean-zero across (i,j,k); $\|\varsigma_{ijk} \boldsymbol{\omega}_{ij}\|_2 \leq \frac{\sqrt{2}}{4}\kappa =: R$ and $\sum \mathbb{E}\, \|\varsigma_{ijk} \boldsymbol{\omega}_{ij}\|_2^2 \leq \frac{1}{8}\kappa^2 m^* n^2 =: \nu$. Vector Bernstein (Fact 1) in dimension $n+1$ with $t = C' n\sqrt{\log n}$ gives, as in Lemma 5, probability $O(n^{-(c+1)})$ for C' large. A union bound over the p_β columns and $\|\boldsymbol{M}\|_2 \leq \sqrt{p_\beta}\, \max_l \|\boldsymbol{M}_{\cdot l}\|_2$ complete the proof, with $C_\Sigma := \max\{C, \sqrt{p_\beta}\, C'\}$. □

Lemma 17 (Uniform information floor on the joint neighborhood)**.** *Assume A1–A3 and A6, and let $\mathcal{K} := \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \times \overline{\mathcal{B}(\boldsymbol{\gamma}^*, \rho_n)} \times \overline{\mathcal{B}(\boldsymbol{\beta}^*, \rho_n)}$ with $\varepsilon_n, \rho_n = O(\sqrt{\log n / n})$. Write, for $(\boldsymbol{s}, \boldsymbol{\zeta}) \in \mathcal{K}$,*

$$\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{\zeta}) \;:=\; \boldsymbol{A}(\boldsymbol{s}, \boldsymbol{\zeta}) - \boldsymbol{B}(\boldsymbol{s}, \boldsymbol{\zeta})^\top\, \boldsymbol{J}(\boldsymbol{s}, \boldsymbol{\zeta})^{-1}\, \boldsymbol{B}(\boldsymbol{s}, \boldsymbol{\zeta}),$$
$$\boldsymbol{A} := \begin{pmatrix} \boldsymbol{A}_{\boldsymbol{\gamma\gamma}} & 0 \\ 0 & \boldsymbol{A}_{\boldsymbol{\beta\beta}} \end{pmatrix}, \qquad \boldsymbol{B} := [\, \boldsymbol{B}_{\boldsymbol{\gamma}} \mid \boldsymbol{B}_{\boldsymbol{\beta}} \,],$$

with $\boldsymbol{J}^{-1}$ taken on $\mathbf{1}^\perp$. For every $c > 0$ there is n_0 such that for all $n \geq n_0$, on $\mathcal{E}_\Sigma \cap \mathcal{E}_{\mathrm{sab}}$,

$$\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{\zeta}) \;\succeq\; \frac{1}{2}\, c_\Sigma\, n^2\, \boldsymbol{I}_p \qquad \textit{for every } (\boldsymbol{s}, \boldsymbol{\zeta}) \in \mathcal{K},$$

where $p = p_\gamma + p_\beta$.

Proof. Throughout, $\boldsymbol{J}^{-1}$ is well defined on $\mathbf{1}^\perp$ with $\big|\!\big|\!\big|\boldsymbol{J}(\boldsymbol{s},\boldsymbol{\zeta})^{-1}\big|\!\big|\!\big|_{2,\perp} \le 2/(nb_0)$ for all $(\boldsymbol{s},\boldsymbol{\zeta}) \in \mathcal{K}$, by Lemma 12 (whose event is $\mathcal{E}_{\mathrm{sab}}$); likewise $\big|\!\big|\!\big|\boldsymbol{V}^{-1}\big|\!\big|\!\big|_{2,\perp} \le 1/(nb_0)$ by Lemma 2. Also $|\!|\!|\boldsymbol{B}(\boldsymbol{s},\boldsymbol{\zeta})|\!|\!|_2 \le C_M n^{3/2}$ on $\mathcal{K}$ by Lemma 8, and $|\!|\!|\boldsymbol{B}^{\mathrm{pop}}|\!|\!|_2 \le C_M n^{3/2}$ by the identical Cauchy–Schwarz argument (with $g^*_{ijk} w^*_{\theta,ijk} \le \frac{1}{4}$).

Step 1: from $\boldsymbol{\Sigma}^{\mathrm{pop}}$ to $\boldsymbol{\Sigma}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)$. Set $\delta\boldsymbol{A} := \boldsymbol{A}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*) - \boldsymbol{A}^{\mathrm{pop}}$ and $\delta\boldsymbol{B} := \boldsymbol{B}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*) - \boldsymbol{B}^{\mathrm{pop}}$; only their direction blocks are nonzero, and on $\mathcal{E}_\Sigma$ both have spectral norm $\le C_\Sigma n\sqrt{\log n}$ (Lemma 16). Telescoping,

$$\boldsymbol{B}^\top \boldsymbol{J}^{-1}\boldsymbol{B} - (\boldsymbol{B}^{\mathrm{pop}})^\top \boldsymbol{V}^{-1}\boldsymbol{B}^{\mathrm{pop}} = \delta\boldsymbol{B}^\top \boldsymbol{J}^{-1}\boldsymbol{B} + (\boldsymbol{B}^{\mathrm{pop}})^\top \boldsymbol{J}^{-1}\delta\boldsymbol{B} + (\boldsymbol{B}^{\mathrm{pop}})^\top \big(\boldsymbol{J}^{-1} - \boldsymbol{V}^{-1}\big)\boldsymbol{B}^{\mathrm{pop}},$$

all matrices evaluated at $(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)$. On $\mathcal{E}_{\mathrm{sab}}$, $\boldsymbol{J}^{-1} - \boldsymbol{V}^{-1} = -\boldsymbol{J}^{-1}\boldsymbol{E}\,\boldsymbol{V}^{-1}$ with $|\!|\!|\boldsymbol{E}|\!|\!|_2 \le C_E\sqrt{n\log n}$ (Lemma 4), so the three terms are bounded in spectral norm by

$$2 \cdot C_M n^{3/2} \cdot \frac{2}{nb_0} \cdot C_\Sigma n\sqrt{\log n} \;+\; C_M^2 n^3 \cdot \frac{2C_E\sqrt{n\log n}}{n^2 b_0^2} \;=\; O\big(n^{3/2}\sqrt{\log n}\big) \;=\; o(n^2),$$

whence, with $|\!|\!|\delta\boldsymbol{A}|\!|\!|_2 = O(n\sqrt{\log n})$ and Weyl, $\boldsymbol{\Sigma}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*) \succeq \boldsymbol{\Sigma}^{\mathrm{pop}} - o(n^2)\,\boldsymbol{I}_p \succeq \frac{3}{4}c_\Sigma n^2\,\boldsymbol{I}_p$ for $n \ge n_0$, by A6.

Step 2: across $\mathcal{K}$. Fix $(\boldsymbol{s},\boldsymbol{\zeta}) \in \mathcal{K}$. For the $\boldsymbol{A}$-block, Lemma 10 gives $|\!|\!|\boldsymbol{A}(\boldsymbol{s},\boldsymbol{\zeta}) - \boldsymbol{A}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)|\!|\!|_2 \le L_A n^2(\varepsilon_n + 2\rho_n) = o(n^2)$. For the cross block, Lemma 9 gives $|\!|\!|\boldsymbol{B}(\boldsymbol{s},\boldsymbol{\zeta}) - \boldsymbol{B}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)|\!|\!|_2 \le L_B n^{3/2}(\varepsilon_n + 2\rho_n) = o(n^{3/2})$. For the inverse,

$$\begin{aligned}\big|\!\big|\!\big|\boldsymbol{J}(\boldsymbol{s},\boldsymbol{\zeta})^{-1} - \boldsymbol{J}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)^{-1}\big|\!\big|\!\big|_{2,\perp} &\le \frac{4}{n^2 b_0^2}\; |\!|\!|\boldsymbol{J}(\boldsymbol{s},\boldsymbol{\zeta}) - \boldsymbol{J}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*)|\!|\!|_2 \\ &\le \frac{4}{n^2 b_0^2}\big(4L_f m^* n\,\varepsilon_n + 2L_\zeta n\,\rho_n\big) = o(n^{-1})\end{aligned}$$

by Lemmas 6 and 7. Telescoping $\boldsymbol{B}^\top\boldsymbol{J}^{-1}\boldsymbol{B}$ between the two points as in Step 1, each of the three terms is $O(n^{3/2}) \cdot O(n^{-1}) \cdot o(n^{3/2})$ or $O(n^3) \cdot o(n^{-1})$, hence $o(n^2)$; with the $\boldsymbol{A}$-increment and Weyl, $\boldsymbol{\Sigma}(\boldsymbol{s},\boldsymbol{\zeta}) \succeq \boldsymbol{\Sigma}(\boldsymbol{s}^*,\boldsymbol{\zeta}^*) - o(n^2)\,\boldsymbol{I}_p \succeq \frac{1}{2}c_\Sigma n^2\,\boldsymbol{I}_p$ for $n \ge n_0$, uniformly over $\mathcal{K}$ since every $o(n^2)$ bound above depends on $(\boldsymbol{s},\boldsymbol{\zeta})$ only through ε_n, ρ_n. □

Structural scores and Hessian blocks. Write the stacked structural score of $-\ell$ as

$$\boldsymbol{T}(\boldsymbol{s},\boldsymbol{\zeta}) := \nabla_{\boldsymbol{\zeta}}(-\ell)(\boldsymbol{s},\boldsymbol{\zeta}) = \big(\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{s},\boldsymbol{\gamma})^\top,\; \boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{s},\boldsymbol{\beta})^\top\big)^\top \in \mathbb{R}^p, \qquad p = p_\gamma + p_\beta,$$

where, from (B.11)–(B.12),

$$\begin{aligned}\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{s},\boldsymbol{\gamma}) &:= \nabla_{\boldsymbol{\gamma}}(-\ell) = \sum_{i<j}\sum_k (g_{ijk} - c_{ijk})\,\boldsymbol{Y}_{ijk}, \\ \boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{s},\boldsymbol{\beta}) &:= \nabla_{\boldsymbol{\beta}}(-\ell) = \sum_{i<j}\sum_k (c_{ijk}d_{ijk} - a^1_{ijk})\,\boldsymbol{Z}_{ijk},\end{aligned}$$

with the corresponding curvature blocks $\boldsymbol{A}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ and $\boldsymbol{A}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ as in (B.17) and (B.18). Since $g_{ijk} = \sigma(f(\Delta_{ij}) + \boldsymbol{Y}_{ijk}^\top\boldsymbol{\gamma})$ carries no $\boldsymbol{\beta}$ and $d_{ijk} = \sigma(\Delta_{ij} + \boldsymbol{Z}_{ijk}^\top\boldsymbol{\beta})$ carries no $\boldsymbol{\gamma}$ (the data c, a^1 being fixed), at a fixed merit $\boldsymbol{s}$ the map $\boldsymbol{G}_{\boldsymbol{\gamma}}$ depends on $\boldsymbol{\gamma}$ alone and $\boldsymbol{G}_{\boldsymbol{\beta}}$ on $\boldsymbol{\beta}$ alone—equivalently $\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\beta}}(-\ell) = 0$ by (B.21)—so the structural block $\boldsymbol{A}$ of the Hessian is block-diagonal. The two channels are nonetheless *not* independent estimation problems: both share the merit vector, and profiling out $\boldsymbol{s}$ re-couples them through the cross blocks, the profile Hessian being the Schur complement $\boldsymbol{\Sigma} = \boldsymbol{A} - \boldsymbol{B}^\top\boldsymbol{J}^{-1}\boldsymbol{B}$ of Lemma 18, whose off-diagonal block $\boldsymbol{B}_{\boldsymbol{\gamma}}^\top\boldsymbol{J}^{-1}\boldsymbol{B}_{\boldsymbol{\beta}}$ is nonzero in general. The joint argument below therefore treats $\boldsymbol{\zeta} = (\boldsymbol{\gamma},\boldsymbol{\beta})$ as a single p-dimensional parameter.

The profiled likelihood. Let $\Theta_2 := \mathcal{B}(\boldsymbol{\gamma}^*, 2\rho_n) \times \mathcal{B}(\boldsymbol{\beta}^*, 2\rho_n)$ denote the doubled structural ball. Lemmas 13 and 15 hold verbatim with ρ_n replaced by $2\rho_n$ (their hypotheses require only a radius of the form $\mathrm{const}\cdot\sqrt{\log n/n}$, now with constant $2C_{\gamma\beta}$); we henceforth work on the corresponding event $\widetilde{\mathcal{E}}_{\mathrm{loo}}$ of Lemma 15 run at radius $2\rho_n$ (and horizon $T+1$ as in the proof of Lemma 13), on which, by the final assertion of Lemma 15, the conclusions of Lemma 13 hold simultaneously for every

$\boldsymbol{\zeta} \in \Theta_2$—including the degenerate radius $C_{\gamma\beta} = 0$ at $\boldsymbol{\zeta} = \boldsymbol{\zeta}^* := (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$. For $\boldsymbol{\zeta} \in \Theta_2$ write $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}$ for the unique zero of $\boldsymbol{G}_{\boldsymbol{\zeta}}$ in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$ furnished by Lemma 13, with $\varepsilon_n = C_5\sqrt{\log n/n}$ and C_5 the Stage-1 constant at the doubled radius. The *profiled negative log-likelihood* is

$$\varphi(\boldsymbol{\zeta}) \;:=\; -\ell\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big), \qquad \boldsymbol{\zeta} \in \Theta_2. \tag{B.40}$$

Since $-\ell(\cdot, \boldsymbol{\zeta})$ is strictly convex on $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$ (Lemma 12) with interior critical point $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}$, one has equivalently $\varphi(\boldsymbol{\zeta}) = \min\{-\ell(\boldsymbol{s}, \boldsymbol{\zeta}) : \boldsymbol{s} \in \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)\}$, so φ is the merit-profiled objective in the usual sense.

Lemma 18 (Profile regularity and envelope identities)**.** *Assume A1–A3. For every $c > 0$ there is n_0 such that for all $n \geq n_0$, on $\widetilde{\mathcal{E}}_{\mathrm{loo}}$, the following hold.*

*(a) (*Regularity of the profile root.*) The map $\boldsymbol{\zeta} \mapsto \widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}$ is C^1 on the interior of Θ_2, with values in $\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n) \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$ and Jacobian*

$$\frac{d\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}}{d\boldsymbol{\zeta}} \;=\; -\,\boldsymbol{J}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)^{-1}\, \boldsymbol{B}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) \;\in\; \mathbb{R}^{(n+1)\times p},$$

where $\boldsymbol{J}^{-1}$ is taken on $\mathbf{1}^\perp$ and $\boldsymbol{B} = [\,\boldsymbol{B}_{\boldsymbol{\gamma}} \mid \boldsymbol{B}_{\boldsymbol{\beta}}\,]$ as in Lemma 9.

*(b) (*Envelope identities.*) φ is C^2 on the interior of Θ_2, with*

$$\begin{aligned}\nabla\varphi(\boldsymbol{\zeta}) \;&=\; \boldsymbol{T}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big),\\ \nabla^2\varphi(\boldsymbol{\zeta}) \;&=\; \boldsymbol{\Sigma}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) \;=\; \boldsymbol{A}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) - \boldsymbol{B}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)^\top \boldsymbol{J}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)^{-1} \boldsymbol{B}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big).\end{aligned}$$

*(c) (*Profile curvature floor.*) If in addition A6 holds then, on $\widetilde{\mathcal{E}}_{\mathrm{loo}} \cap \mathcal{E}_\Sigma$,*

$$\nabla^2\varphi(\boldsymbol{\zeta}) \;\succeq\; \frac{1}{2}\, c_\Sigma\, n^2\, \boldsymbol{I}_p \qquad \textit{for every } \boldsymbol{\zeta} \textit{ in the interior of } \Theta_2.$$

Proof. (a). Parametrize the gauge slice by $\mathbf{1}^\perp$: writing $\boldsymbol{s} = \boldsymbol{s}^* + \boldsymbol{u}$ with $\boldsymbol{u} \in \mathbf{1}^\perp$, the map $\boldsymbol{F}(\boldsymbol{u}, \boldsymbol{\zeta}) := \boldsymbol{G}_{\boldsymbol{\zeta}}(\boldsymbol{s}^* + \boldsymbol{u}) \in \mathbf{1}^\perp$ is C^2 on an open neighborhood of its domain (under A3, $f \in C^3$ makes ℓ three times continuously differentiable in all arguments, and $\boldsymbol{G}_{\boldsymbol{\zeta}} = \nabla_{\boldsymbol{s}}(-\ell)$). Fix $\boldsymbol{\zeta}_0$ in the interior of Θ_2 and set $\boldsymbol{u}_0 := \widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}_0} - \boldsymbol{s}^*$, so $\boldsymbol{F}(\boldsymbol{u}_0, \boldsymbol{\zeta}_0) = \mathbf{0}$ with $\|\boldsymbol{u}_0\|_\infty < \varepsilon_n$ (the root is interior to the ℓ_∞ ball since, by the proof of Lemma 13, $\|\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}} - \boldsymbol{s}^*\|_\infty \leq C_4\sqrt{\log n/n} + o(\sqrt{\log n/n})$ and $C_5 = C_4 + 1$). The partial derivative $\partial_{\boldsymbol{u}}\boldsymbol{F} = \boldsymbol{J}(\boldsymbol{s}^* + \boldsymbol{u}, \boldsymbol{\zeta})\big|_{\mathbf{1}^\perp}$ is invertible at $(\boldsymbol{u}_0, \boldsymbol{\zeta}_0)$ with $\lambda_{\min,\perp} \geq \frac{1}{2} n b_0$ (Lemma 12, whose event $\mathcal{E}_{\mathrm{sab}} \supseteq \widetilde{\mathcal{E}}_{\mathrm{loo}}$ is in force, its radii ε_n, $2\rho_n$ being $o(1)$). By the implicit function theorem on $\mathbf{1}^\perp \times \mathbb{R}^p$, there is a neighborhood of $\boldsymbol{\zeta}_0$ and a C^1 map $\boldsymbol{\zeta} \mapsto \boldsymbol{u}(\boldsymbol{\zeta})$ with $\boldsymbol{F}(\boldsymbol{u}(\boldsymbol{\zeta}), \boldsymbol{\zeta}) = \mathbf{0}$ and $\boldsymbol{u}(\boldsymbol{\zeta}_0) = \boldsymbol{u}_0$; shrinking the neighborhood keeps $\|\boldsymbol{u}(\boldsymbol{\zeta})\|_\infty < \varepsilon_n$, so by the uniqueness in Lemma 13 the branch coincides with $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}} - \boldsymbol{s}^*$ there. Hence $\boldsymbol{\zeta} \mapsto \widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}$ is C^1 at every interior $\boldsymbol{\zeta}_0$. Differentiating $\boldsymbol{G}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}) = \mathbf{0}$ in $\boldsymbol{\zeta}$,

$$\boldsymbol{J}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)\, \frac{d\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}}{d\boldsymbol{\zeta}} + \frac{\partial \boldsymbol{G}_{\boldsymbol{\zeta}}}{\partial \boldsymbol{\zeta}}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}\big) \;=\; \mathbf{0},$$

and $\partial\boldsymbol{G}_{\boldsymbol{\zeta}}/\partial\boldsymbol{\zeta} = \boldsymbol{B}$ (proof of Lemma 8); solving on $\mathbf{1}^\perp$ (both the columns of $\boldsymbol{B}$ and the range of $d\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}/d\boldsymbol{\zeta}$ lie in $\mathbf{1}^\perp$, the root staying in the gauge slice) gives the stated Jacobian.

(b). By the chain rule and part (a),

$$\nabla\varphi(\boldsymbol{\zeta}) = \Big(\frac{d\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}}{d\boldsymbol{\zeta}}\Big)^\top \underbrace{\nabla_{\boldsymbol{s}}(-\ell)\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)}_{=\,\boldsymbol{G}_{\boldsymbol{\zeta}}(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}) = \mathbf{0}} + \nabla_{\boldsymbol{\zeta}}(-\ell)\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) \;=\; \boldsymbol{T}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) :$$

the merit direction contributes exactly zero, by the definition of the profile root. For the Hessian, differentiate once more:

$$\nabla^2\varphi(\boldsymbol{\zeta}) = \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{s}}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)\, \frac{d\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}}{d\boldsymbol{\zeta}} + \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\zeta}}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) = \boldsymbol{B}^\top\big(-\boldsymbol{J}^{-1}\boldsymbol{B}\big) + \boldsymbol{A} = \boldsymbol{\Sigma}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big),$$

using $\partial \boldsymbol{T}/\partial \boldsymbol{s} = \nabla^2_{\boldsymbol{\zeta s}}(-\ell) = \boldsymbol{B}^\top$ and $\partial \boldsymbol{T}/\partial \boldsymbol{\zeta} = \nabla^2_{\boldsymbol{\zeta\zeta}}(-\ell) = \boldsymbol{A}$, the latter block-diagonal by (B.21). In particular $\nabla^2\varphi$ is symmetric, as it must be, and φ is C^2 since all blocks are C^1 compositions of C^1 maps.

(c). By part (a), $\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big) \in \mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n) \times \Theta_2 \subseteq \mathcal{K}$ with the radii of Lemma 17 (there ρ_n is read as $2\rho_n = O(\sqrt{\log n/n})$), so that lemma applies at $\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}\big)$ and gives $\nabla^2\varphi(\boldsymbol{\zeta}) = \boldsymbol{\Sigma}(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}, \boldsymbol{\zeta}) \succeq \frac{1}{2} c_\Sigma n^2 \, \boldsymbol{I}_p$. □

Lemma 19 (Structural score at the profile anchor)**.** *Assume A1–A3 and $n \geq 2$. Let C_5' be the Stage-1 constant of Lemma 13 run with $C_{\gamma\beta} = 0$, i.e. $C_5' = C_4(0) + 1$ with $C_4(0)$ from Lemma 14 at $C_{\gamma\beta} = 0$; in particular C_5' does not depend on $C_{\gamma\beta}$, and on $\widetilde{\mathcal{E}}_{\mathrm{loo}}$ the merit root $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*}$ satisfies $\|\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*} - \boldsymbol{s}^*\|_\infty \leq C_5'\sqrt{\log n/n}$. Then, on $\widetilde{\mathcal{E}}_{\mathrm{loo}}$,*

$$\big\|\boldsymbol{T}\big(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*}, \boldsymbol{\zeta}^*\big)\big\|_2 \;\leq\; C_\star \, n^{3/2}\sqrt{\log n}, \qquad C_\star \;:=\; \sqrt{2}\,\big(C_\kappa + C_M\, C_5'\big).$$

Proof. At the truth, the two blocks of $\boldsymbol{T}(\boldsymbol{s}^*, \boldsymbol{\zeta}^*)$ are $\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{s}^*, \boldsymbol{\gamma}^*)$ and $\boldsymbol{G}_{\boldsymbol{\beta}}(\boldsymbol{s}^*, \boldsymbol{\beta}^*)$, each of ℓ_2 norm $\leq C_\kappa n\sqrt{\log n}$ on $\mathcal{E}_{\mathrm{score}} \supseteq \widetilde{\mathcal{E}}_{\mathrm{loo}}$ (Lemma 3, structural part), so the stacked vector obeys $\|\boldsymbol{T}(\boldsymbol{s}^*, \boldsymbol{\zeta}^*)\|_2 \leq \sqrt{2}\, C_\kappa \, n\sqrt{\log n}$. For the increment, $\partial \boldsymbol{T}/\partial \boldsymbol{s} = \boldsymbol{B}^\top$ (proof of Lemma 18), so the fundamental theorem of calculus along $\boldsymbol{s}^* + \tau(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*} - \boldsymbol{s}^*)$—a segment contained in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, C_5'\sqrt{\log n/n})}$, hence in the working region—and Lemma 8 give

$$\begin{aligned}\|\boldsymbol{T}(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*}, \boldsymbol{\zeta}^*) - \boldsymbol{T}(\boldsymbol{s}^*, \boldsymbol{\zeta}^*)\|_2 &\leq\; C_M n^{3/2}\, \|\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*} - \boldsymbol{s}^*\|_2 \\ &\leq\; C_M n^{3/2}\sqrt{n+1}\; C_5'\sqrt{\frac{\log n}{n}} \;\leq\; \sqrt{2}\, C_M C_5'\, n^{3/2}\sqrt{\log n}.\end{aligned}$$

Adding the two bounds and using $n\sqrt{\log n} \leq n^{3/2}\sqrt{\log n}$ gives the claim. □

Theorem 2 (Joint ℓ_∞/ℓ_2 consistency)**.** *Assume A1–A3 and A6. For every $c > 0$ there is $n_0 = n_0(c)$ and an event of probability at least $1 - O(n^{-c})$ on which, for all $n \geq n_0$, there is a point $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$ with $\nabla_{\boldsymbol{s}}\ell = \nabla_{\boldsymbol{\gamma}}\ell = \nabla_{\boldsymbol{\beta}}\ell = \mathbf{0}$ satisfying*

$$\|\widehat{\boldsymbol{s}} - \boldsymbol{s}^*\|_\infty = O\Big(\sqrt{\frac{\log n}{n}}\Big), \qquad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 = O\Big(\sqrt{\frac{\log n}{n}}\Big), \qquad \Big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\Big\|_2 = O\Big(\sqrt{\frac{\log n}{n}}\Big).$$

Moreover $\widehat{\boldsymbol{\zeta}} = (\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$ is the unique zero of $\nabla\varphi$ in $\overline{\mathcal{B}(\boldsymbol{\zeta}^*, \rho_n)}$.

Proof. Constants, in acyclic order. First run Lemma 14 at the fixed pair $(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$, i.e. with $C_{\gamma\beta} = 0$, and let C_5' be the resulting Stage-1 constant of Lemma 19; C_5' depends only on the constants of A1–A3 and on c. Next set

$$C_\star := \sqrt{2}\,(C_\kappa + C_M C_5'), \qquad C_{\gamma\beta} := \frac{3C_\star}{c_\Sigma}, \qquad \rho_n := C_{\gamma\beta}\sqrt{\frac{\log n}{n}}.$$

Finally run Lemmas 14 and 15 at the doubled radius $2\rho_n$ (constant $2C_{\gamma\beta}$), obtain the Stage-1 constant C_5 ($\geq C_5'$ without loss of generality), and set $\varepsilon_n := C_5\sqrt{\log n/n}$. The chain is $C_5' \to C_\star \to C_{\gamma\beta} \to C_5$, and C_5 is consumed below only through the $o(1)$ radius ε_n (via Lemmas 17 and 18), never through a score bound; there is no cycle.

Event. Work on $\mathcal{E} := \widetilde{\mathcal{E}}_{\mathrm{loo}} \cap \mathcal{E}_\Sigma$, with $\widetilde{\mathcal{E}}_{\mathrm{loo}}$ the event of Lemma 15 at radius $2\rho_n$ (it contains $\mathcal{E}_{\mathrm{score}}$ and $\mathcal{E}_{\mathrm{sab}}$) and $\mathcal{E}_\Sigma$ that of Lemma 16; the union bound gives $\mathbb{P}(\mathcal{E}) \geq 1 - O(n^{-c})$. On $\mathcal{E}$, the profile root $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}}$ and the profile φ of (B.40) are defined on Θ_2, and Lemma 18 is in force. Note that the root $\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*}$ of Lemma 19 coincides with the profile root at $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*$: both are zeros of $\boldsymbol{G}_{\boldsymbol{\gamma}^*,\boldsymbol{\beta}^*}$ in $\overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \cap (\boldsymbol{s}^* + \mathbf{1}^\perp)$ (using $C_5' \leq C_5$), where Lemma 13 guarantees uniqueness.

Existence of a structural root. Apply the degree-theoretic criterion (Fact 4) to $\boldsymbol{F} = \nabla\varphi$ on $\Omega = \mathcal{B}(\boldsymbol{\zeta}^*, \rho_n)$ with anchor $\boldsymbol{\zeta}^*$; $\boldsymbol{F}$ is continuous on $\overline{\Omega}$ since $\overline{\Omega}$ lies in the interior of Θ_2, where φ is C^2 (Lemma 18(b)). For $\boldsymbol{\zeta} \in \partial\Omega$ write $\boldsymbol{u} := \boldsymbol{\zeta} - \boldsymbol{\zeta}^*$, $\|\boldsymbol{u}\|_2 = \rho_n$. By the fundamental theorem of

calculus along $[\boldsymbol{\zeta}^*, \boldsymbol{\zeta}] \subset \overline{\Omega}$, the envelope identity $\nabla\varphi(\boldsymbol{\zeta}^*) = \boldsymbol{T}(\widehat{\boldsymbol{s}}_{\boldsymbol{\zeta}^*}, \boldsymbol{\zeta}^*)$, and the curvature floor of Lemma 18(c),

$$\boldsymbol{u}^\top \nabla\varphi(\boldsymbol{\zeta}) = \boldsymbol{u}^\top \nabla\varphi(\boldsymbol{\zeta}^*) + \boldsymbol{u}^\top \Big(\int_0^1 \nabla^2\varphi(\boldsymbol{\zeta}^* + \tau\boldsymbol{u})\, d\tau \Big)\boldsymbol{u} \;\geq\; -\rho_n\, C_\star n^{3/2}\sqrt{\log n} \;+\; \frac{1}{2}\, c_\Sigma\, n^2 \rho_n^2,$$

using Lemma 19 for the first term. Hence

$$\boldsymbol{u}^\top \nabla\varphi(\boldsymbol{\zeta}) \;\geq\; \frac{1}{2}\, c_\Sigma\, n^2 \rho_n \Big(\rho_n - \frac{2C_\star}{c_\Sigma}\sqrt{\frac{\log n}{n}} \Big) \;>\; 0,$$

the bracket being strictly positive since $C_{\gamma\beta} = 3C_\star/c_\Sigma > 2C_\star/c_\Sigma$. By Fact 4, $\nabla\varphi$ has a zero $\widehat{\boldsymbol{\zeta}} \in \overline{\Omega}$; since $\boldsymbol{u}^\top \nabla\varphi > 0$ on $\partial\Omega$, the zero is interior, $\widehat{\boldsymbol{\zeta}} \in \mathcal{B}(\boldsymbol{\zeta}^*, \rho_n)$.

Uniqueness of the structural root. For distinct $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \in \overline{\Omega}$, with $\bar{\boldsymbol{\Sigma}} := \int_0^1 \nabla^2\varphi(\boldsymbol{\zeta}_2 + \tau(\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2))\, d\tau \succeq \frac{1}{2} c_\Sigma n^2\, \boldsymbol{I}_p$ (Lemma 18(c), the segment lying in $\overline{\Omega}$),

$$\big(\nabla\varphi(\boldsymbol{\zeta}_1) - \nabla\varphi(\boldsymbol{\zeta}_2)\big)^\top (\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2) = (\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2)^\top\, \bar{\boldsymbol{\Sigma}}\, (\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2) \;\geq\; \frac{1}{2}\, c_\Sigma\, n^2\, \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|_2^2 \;>\; 0,$$

so $\nabla\varphi$ is strictly monotone on $\overline{\Omega}$ and $\widehat{\boldsymbol{\zeta}}$ is its unique zero there.

Conclusion. Set $\widehat{\boldsymbol{s}} := \widehat{\boldsymbol{s}}_{\widehat{\boldsymbol{\zeta}}}$ and $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) := \widehat{\boldsymbol{\zeta}}$. Then $\nabla_{\boldsymbol{s}}\ell(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{\zeta}}) = \boldsymbol{0}$ by the definition of the profile root, and $\nabla_{\boldsymbol{\zeta}}\ell(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{\zeta}}) = -\boldsymbol{T}(\widehat{\boldsymbol{s}}_{\widehat{\boldsymbol{\zeta}}}, \widehat{\boldsymbol{\zeta}}) = -\nabla\varphi(\widehat{\boldsymbol{\zeta}}) = \boldsymbol{0}$ by the envelope identity (Lemma 18(b)). The rates hold with the constants built into the radii: $\|\widehat{\boldsymbol{s}} - \boldsymbol{s}^*\|_\infty \leq \varepsilon_n = C_5\sqrt{\log n/n}$ and $\max\{\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2, \big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\big\|_2\} \leq \big\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}^*\big\|_2 \leq \rho_n = C_{\gamma\beta}\sqrt{\log n/n}$. □

Corollary 1 (Uniqueness and the MLE)**.** *On the event and under the assumptions of Theorem 2, the point $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$ is the unique stationary point of ℓ in $\mathcal{K} = \overline{\mathcal{B}_\infty(\boldsymbol{s}^*, \varepsilon_n)} \times \overline{\mathcal{B}(\boldsymbol{\gamma}^*, \rho_n)} \times \overline{\mathcal{B}(\boldsymbol{\beta}^*, \rho_n)}$ intersected with the gauge slice $\boldsymbol{s}^* + \mathbf{1}^\perp$, and it maximizes ℓ over that set; it is in this sense the MLE.*

Proof. On $\mathbf{1}^\perp \times \mathbb{R}^p$ the full Hessian of $-\ell$ is the bordered matrix $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{J} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{A} \end{pmatrix}$. Throughout $\mathcal{K}$, $\boldsymbol{J} \succeq \frac{1}{2} n b_0$ on $\mathbf{1}^\perp$ (Lemma 12) and $\boldsymbol{M}/\boldsymbol{J} = \boldsymbol{\Sigma} \succeq \frac{1}{2} c_\Sigma n^2\, \boldsymbol{I}_p$ (Lemma 17), so by the Schur criterion (Fact 6) $\boldsymbol{M} \succ 0$ throughout $\mathcal{K}$. Hence $-\ell$ is strictly convex on the convex set $\mathcal{K} \cap [(\boldsymbol{s}^* + \mathbf{1}^\perp) \times \mathbb{R}^{p_\gamma} \times \mathbb{R}^{p_\beta}]$ and admits at most one stationary point there, which is its minimizer if interior. Theorem 2 exhibits an interior stationary point; it is therefore unique and maximizes ℓ over $\mathcal{K} \cap [(\boldsymbol{s}^* + \mathbf{1}^\perp) \times \mathbb{R}^{p_\gamma} \times \mathbb{R}^{p_\beta}]$. □

Dial	Type	Levels	Group
Experience (years)	ordered	3 / 5 / 7	primary
Education level	ordered	BS / MS / PhD	primary
School selectivity	ordered	low / mid / high	primary
Employer prestige	ordered	startup / mid / big tech	primary
Languages known	ordered	2 / 5 / 9	primary
Production launches	ordered	2 / 4 / 6	primary
Skill coverage[†]	ordered	2 / 4 / 6	primary
Prose fluency[‡]	ordered	0 / 5 / 12 mistakes	primary
DSA depth	binary	absent / present	per-skill
Python	binary	absent / present	per-skill
AI/agents experience	binary	absent / present	per-skill
Security assessments	binary	absent / present	per-skill
Vuln-triage automation	binary	absent / present	per-skill
Agentic tooling design	binary	absent / present	per-skill
Side projects	niche[*]	absent / toy / hard	niche
Open-source	niche[*]	absent / patches / maintainer	niche
Security certification	niche[*]	absent / entry / advanced	niche
Capture-the-flag (CTF)	niche[*]	absent / participated / strong	niche

[†] Counts the number of skills mentioned; a composite of the per-skill dials, hence not counted as a separate recovery target.
[‡] Deliberately irrelevant to the role; serves as a leakage check. Hence $D = 16$ relevant dials.
[*] Exactly 2 of the 4 niche dials are non-absent in any given letter.

Table 3: The 18 engineered ground-truth dials of the cover-letter corpus.

C Experiment Details on Interpretable Feature Recovery

This appendix collects the full experimental details for the synthetic cover-letter study of Section 4. Appendix C.1 describes the seeded corpus generation: the planted dials, their levels and sampling distributions, and the generation prompts. Appendix C.2 gives the formal definitions of the recovery metrics (Tracked, Stat.R., Sem.R., Attrib., LK), including the matching procedure between discovered features and planted dials. It also specifies the four baseline configurations (Direct LLM, Direct LLM–Multi-score, Local LLMs, Local LLMs–Multi-score), their prompts, and the API budget of each method. Appendix C.3 provides additional results such as the full list of surfaced features with their fitted parameters and diagnostics, and the judge-written rubric. Finally, Appendix C.4 details the paired-bootstrap protocol behind Figure 15 and Table 7, and gathers additional recovery figures (dial-recovery heatmaps and the feature-merging threshold analysis).

C.1 Dataset

Job ad. The real job ad was archived here. The job ad obtained after trimming boilerplate content is shown in Figure 11.

Example of template. Figure 12 shows one of the ten templates used for the cover letter generation.

In what follows, we describe each dial in more detail: the levels it takes and how those levels are realized in the assembled letter. Everything is also gathered in Table 3.

Experience. The experience dial takes one of three values (3, 5 or 7 years) and is stated in the background clause as a tenure at the current employer. It is realized as a single digit and is one of the three numbers on the letter's digit whitelist.

Education. The education dial takes three levels, BS $<$ MS $<$ PhD, rendered in the opening as a "{EDUCATION} in computer science from {SCHOOL}". The three degrees are written as *two-*

Software Engineer, Information Security Engineering
Google — Singapore

minimum qualifications — ground the education and numbered dials

- Bachelor's degree or equivalent practical experience. ⟨EDUCATION⟩
- 2 years of experience testing, maintaining, or launching software products. ⟨EXPERIENCE YEARS⟩, ⟨NUMBER OF LAUNCHES⟩
- 2 years of experience with software development in one or more programming languages. ⟨NUMBER OF LANGUAGES⟩
- Experience programming in Go. fixed slot, no dial

preferred qualifications — ground the education dial and four role-skill dials

- Master's degree or PhD in Computer Science or a related technical field. ⟨EDUCATION⟩
- 3 years of experience with data structures and algorithms. ⟨SKILL: DSA⟩
- Experience in Python. ⟨SKILL: PYTHON⟩
- Experience with Artificial Intelligence. ⟨SKILL: AI⟩
- Ability to perform security assessments. ⟨SKILL: SECURITY ASSESSMENTS⟩

about the job — fixed context

As a Software Engineer, you will leverage your software development expertise to design, build, and implement agentic tooling and systems that will automate and scale security processes, analyze risks, detect vulnerabilities, and proactively assist developer teams across Google in embedding security best practices throughout the entire development lifecycle. You will be crucial in enabling Google to respond to incoming vulnerability reports and remediate security issues.

responsibilities — ground the two remaining role-skill dials

- Collaborate closely with information security engineers to develop new tools for product security consulting, response and research, especially (but not exclusively) related to AI agents.
- Steer the strategic development of agents and infrastructure to automate vulnerability handling, triage, and conduct variant analysis to reduce manual workloads. ⟨SKILL: VULN-TRIAGE AUTOMATION⟩
- Design and implementation of Go back-end systems and Python AI agents. ⟨SKILL: AGENTIC TOOLING⟩

⟨DIAL⟩ grounds a categorical/numbered dial ⟨SKILL⟩ grounds a role-skill dial

Figure 11: **The job advertisement used as the fixed evaluation context.** Colored tags mark which requirement grounds which dial, using the same visual vocabulary as the letter template (Figure 12): the qualifications ground the education and numbered dials, and the preferred qualifications and responsibilities ground the six role-skill dials. Go is a mandatory requirement realized identically in every letter, not a dial. The remaining dials of Table 3 (school and employer prestige, prose fluency, and the four niche dials) have no anchor in the ad and are described in Appendix C.1.

word phrases: "Bachelor's degree", "Master's degree", "Doctoral degree", so that the dial adds no length signal (the bare token "PhD" would make higher education correlate with shorter letters).

School prestige. The school dial fills the {SCHOOL} slot and takes three ascending selectivity bands: *less selective* (Cal State East Bay, University of Toledo, Kennesaw State University, Central Michigan University), *moderately selective* (strong general publics: Penn State, Ohio State, Arizona State, Virginia Tech), and *highly selective* (MIT, Stanford, UC Berkeley, Princeton). All names are *real* institutions, including the bottom band, whose intended reading is "an ordinary accredited school". Band membership is anchored to a published selectivity classification.

Employer prestige. The employer dial fills the {EMPLOYER} slot and takes three ascending recognizability bands: an *unknown-startup* band of fictional, unrecognizable firms (Northlake, Bluegrove, Quarrypoint, Milldam), a *medium* band of real mid-recognizability companies (Zillow, Etsy, Expedia, Wayfair), and a *big-tech* band (Meta, Amazon, Microsoft, Apple). Unlike the school dial, the bottom band is fictional on purpose: invented names give a clean "no prestige prior" floor, whereas a real but obscure firm would carry its own idiosyncratic associations.

Dear Google Hiring Team, *fixed greeting*

fixed opening prose

Security work is the rare kind of engineering where the best outcome is that nothing happens. That inversion is what pulls me toward the Software Engineer role on the Information Security Engineering team: the chance to build systems whose success is measured in incidents that never occur.

background clause — categorical and numbered dials

My foundation is a ⟨EDUCATION⟩ in computer science from ⟨SCHOOL⟩ and ⟨EXPERIENCE YEARS⟩ years of professional software development at ⟨EMPLOYER⟩, where I have tested, maintained, and launched ⟨NUMBER OF LAUNCHES⟩ production services, writing Go alongside ⟨NUMBER OF LANGUAGES⟩ other languages.

body of six core units: k skill sentences $+(6-k)$ fillers, paragraphed $[2, 2, 2]$

⟨CORE UNIT 1⟩ ⟨CORE UNIT 2⟩
⟨CORE UNIT 3⟩ ⟨CORE UNIT 4⟩
⟨CORE UNIT 5⟩ ⟨CORE UNIT 6⟩

niche paragraph: exactly two active niche dials

⟨NICHE UNIT 1⟩ ⟨NICHE UNIT 2⟩

fixed closing, ending with verbatim authorization sentence

I would be glad to talk in more depth about the team's challenges and where I could contribute from day one. I am fully authorized to work in Singapore and do not require visa sponsorship.

Sincerely,
⟨NAME⟩

⟨DIAL⟩ ground-truth dial value ⟨CORE⟩ skill/filler unit ⟨NICHE⟩ niche unit

Figure 12: **Example cover-letter template.** Roman text is fixed template prose; angle-bracketed slots are filled from the ground-truth dial vector. Skill coverage sets how many core units are capability sentences; paragraph count is constant. Fluency is imposed afterward by injecting $0/5/12$ surface errors.

Programming languages. This numbered dial takes values $\{2, 5, 9\}$ and is stated in the background clause as "writing Go alongside {NUMBER OF LANGUAGES} other languages." Go is a fixed, mandatory slot (a minimum qualification); only the count varies. The count is realized as a single whitelisted digit in a clause present in every letter, so it grades a numbered attribute without introducing any length variation across its levels.

Production launches. This numbered dial takes values $\{2, 4, 6\}$ and appears in the background clause as "tested, maintained, and launched {NUMBER OF LAUNCHES} production services." Like the language count it is a self-contained count (never a duration), realized as a whitelisted digit in a fixed clause. It is assigned orthogonally to the experience dial so that it cannot re-enter the model as an experience proxy.

Skill coverage. The skill-coverage dial sets *how many* of the six role-skills the letter demonstrates, taking values $k \in \{2, 4, 6\}$. Its effect is to determine how many of the six core body units are capability sentences: exactly k core units render a present skill, and the remaining $6 - k$ are *fillers*: role/domain-motivation sentences that never claim a skill, a niche, or communication ability. Because the body always contains six core units laid out in a fixed paragraph plan regardless of k, coverage changes the *content* of the body but not its length or paragraph count, so it cannot ride the verbosity bias.

Skill identity. Which skills are present (as opposed to how many) is a constrained set rather than an ordered dial. The pool has six letter-addressable items derived from the ad's preferred qualifications and responsibilities: data-structures depth, Python, applied-AI experience, security assessments, vulnerability-triage automation, and agentic-tooling design. Each present skill is rendered as one

concrete capability sentence drawn from the current template's phrasing, with no numbers and no durations.

Niche dials. There are four niche dials: side project, open-source involvement, security certification, and capture-the-flag (CTF). Each is a three-level ordered factor: *absent*, a middle level, and a strong level (side project: toy $<$ hard; open source: patches $<$ maintainer; certification: entry (Security+) $<$ advanced (OSCP); CTF: participated $<$ strong). They are subject to a presence constraint: exactly two of the four are non-absent in every letter, so "number of extras mentioned" is held constant and cannot become a richness or length axis. The two active dials are realized, at their assigned level, as evidence clauses in a single dedicated niche paragraph placed at a fixed position after the body.

Fluency. The fluency dial controls surface prose quality and takes three levels anchored to an exact injected-error count: clean (0), mid (5), and degraded (12). Fluency is not a slot but a deterministic post-assembly pass that injects the target number of surface spelling errors into the finished letter; the word count is re-validated in band after injection, so the only thing that varies across fluency levels is the density of errors.

Nuisance factors. Two factors are balanced across the corpus but are not part of the ground truth and are not scored. Each letter is assembled from one of ten templates, which vary the greeting, opening, closing, and the exact wording of every skill, niche, and filler unit while holding the underlying dial content fixed. The applicant's name, drawn from a fixed pool, appears only in the signature and carries no signal.

Dataset creation. The dataset comprises $n = 200$ cover letters. Of these, 184 form a balanced pool: every level of every dial appears an equal number of times, and the joint assignment of levels to letters is chosen as the most orthogonal of 256 independently drawn candidate designs (the candidate minimizing the largest non-structural pairwise $|r|$). The remaining 16 are deliberate anchors: 8 polished-but-thin and 8 rough-gem profiles that anti-correlate the "prestige" dials with the substance dials, included so the judge sees the full range of the design, and excluded from the balance and orthogonality gates below.

Each letter is then assembled deterministically from one of 10 templates keyed to its levels. For example, experience, education, and school/employer prestige are stated in the opening; the $k \in \{2, 4, 6\}$ present role-skills are rendered as capability sentences drawn from a pool (one pool per template), the identity of the k present skills being chosen by a greedy usage-balancing rule so that skill identity is orthogonal to coverage k; production launches is stated in the background clause; the two active niche dials appear as evidence clauses; and prose fluency is imposed by injecting a target number of surface spelling errors (0/5/12).

On the balanced pool the realized design satisfies $|r| \leq 0.15$ over all non-structural dial pairs (max 0.146) and $|r| \leq 0.22$ between each dial and letter length (max 0.202); over the full corpus, where the anchors enter, these are 0.195 and 0.219. Letter length is not part of the design objective but it is verified after assembly. Generation is seeded and fully reproducible.

C.2 Methods Details

C.2.1 Metrics

Each method yields a set of discovered features that we will need to match to the dials.

Matching the discovered features to the dials. For each method, we perform a Hungarian matching assigning each dial to a distinct feature so as to maximize total cosine similarity in TEXT-EMBEDDING-3-LARGE's embedding space. The pairs it rejects (features and dials sharing domain vocabulary but, in the assignment's judgment, not describing the same construct) form an empirical background of unrelated pairs, and we define a threshold η as the 95th percentile of their alignment ($\eta = 0.406$): the level that unrelated pairs almost never attain. A matched pair clearing η is therefore semantically exceptional relative to background similarity, not merely the least-bad assignment; co-

sine similarities carry no absolute scale, so the reference distribution, rather than a fixed constant, is what gives the threshold meaning.

Having defined this threshold, we map a method's set of discovered features to the dial set by performing semantic Hungarian matching, but only keeping matches whose cosine similarity is greater than η.

In what follows, we also say that a feature k *tracks* a dial d if the absolute Pearson correlation $|r_{k,d}|$ between the two is greater than a threshold $\tau = 0.4$ fixed a priori and well above the permutation floor (≈ 0.15 under letter-shuffled nulls; results are stable for $\tau \in [0.3, 0.5]$). The metrics mentioned in Section 4 are mathematically defined as follows:

Statistical recovery. For each dial, does *some* feature track it? We report the fraction of dials with $\max_k |r_{k,d}| > \tau$ (Tracked), and the mean of $\max_k |r_{k,d}|$ over all D dials, misses included (Stat.R.).

Semantic recovery. For each dial, is there a feature that is both *about* it and tracks it? Let M_m assign features to dials by Hungarian matching on semantic alignment, and $c_{m,d}$ the alignment of dial d with its match $M_m(d)$:

$$\text{Sem. R.}(m) = \frac{1}{D} \sum_{d=1}^{D} \mathbf{1}_{\{c_{m,d} \geq \eta\}} \mathbf{1}_{\{|r_{M_m(d),\, d}| > \tau\}}. \tag{C.1}$$

Attributable features. Analogously, what fraction of a method's K_m discovered features measure a single dial? A feature is *attributable* if it tracks some dial and that dial dominates its attribution, i.e. carries a majority of its squared correlation, $\pi_k = \max_d r_{k,d}^2 / \sum_d r_{k,d}^2 > \frac{1}{2}$:

$$\text{Attrib.}(m) = \frac{1}{K_m} \sum_{k=1}^{K_m} \mathbf{1}_{\{\max_d |r_{k,d}| > \tau\}} \mathbf{1}_{\left\{\pi_k > \frac{1}{2}\right\}} \tag{C.2}$$

Leakage. Do scores leak across dials? Since dials are near-orthogonal by construction, off-target correlations of a valid feature set should be small. With $\mathcal{A}_m$ the attributable features and $k_m(d) = \arg\max_k |r_{k,d}|$ the tracker of dial d,

$$\text{LK}(m) = \frac{1}{D} \sum_{d=1}^{D} \underset{k \in \mathcal{A}_m \setminus \{k_m(d)\}}{\text{mean}} |r_{k,d}| \tag{C.3}$$

(lower is better). Restricting to attributable features prevents noise features from diluting the average.

C.2.2 Methods Description

Our approach. For our method, the comparison graph on the $n = 200$ cover letters is a union of $d = 40$ balanced subgraphs (called *matchings*, each a perfect matching in which every item is compared exactly once). The number $d = 40$ was chosen to guarantee high precision.

Prompt 4 was used for the pairwise comparisons (GPT-5.4 MINI), and Prompt 5 was used for the feature taxonomy (GPT-5.4 (HIGH REASONING)). The cost was estimated using the API logs.

Other methods. Other methods feed the entirety of the dataset (Direct LLM)/a sample of it (Local LLMs) to the judge (GPT-5.4 (HIGH REASONING)) and have it draw up a rubric tailored to the samples seen (Prompt 6). This rubric is then used to have GPT-5.4 MINI score each cover letter pointwise on each feature (Prompt 8).

For Local LLMs, only 40 letters of the dataset are given to the judge during the feature elicitation step, enabling the LLM to focus its attention on a smaller sample of data. To ensure that each cover letter is seen at least once, we repeat the elicitation with $E = 27$ different samples, which is the average number of times one would have to repeat the step so that each cover letter is seen at least once. We then consider the union of the features, performed with GPT-5.4 (HIGH REASONING) and Prompt 7 (see Remark 7 for why we used LLM-based taxonomy instead of semantic hierarchical clustering).

To match the cost of our approach, the scoring is repeated 20 times for Direct LLM–Multi-score, and 6 times for Local LLMs–Multi-score. Each scoring is performed by slightly modifying the prompt, changing the instruction variant (Prompt 9) and permuting the features inside the rubric.

C.2.3 PROMPTS

Prompt 4: Comparative method (ours) – pairwise comparison prompt.

```
You are helping build an interpretable scoring framework by examining how job applicants'
cover letters differ. The framework works by collecting many pairwise comparisons: for each
pair, the ways the two letters differ are articulated in plain terms, and aggregated across
many pairs these reveal the distinct dimensions along which the applications vary. There is
no fixed rubric — the dimensions are recovered from comparisons like this one, not
specified in advance.

The letters below are applications for the same software-engineering role. The full job
posting they respond to is:
=== JOB POSTING ===
⟨full job posting⟩
=== END JOB POSTING ===

Your task. You will be shown two cover letters, Letter A and Letter B, both written for the
role above. This pair is one comparison in a larger collection.

Identify the dimensions on which the two letters meaningfully differ as applications for
this role. Report the differences that genuinely stand out to you on reading this pair —
bring up whatever you find salient, and do not work from a predetermined list. Favor real,
noticeable differences over minor ones, but do not force the list to be short: report as
many as actually distinguish these two letters.

For each dimension, first name the concrete, specific feature you actually saw in the
letters — a short phrase pointing to the specific evidence. Then, grounded in that
evidence, name the dimension, describe in one sentence what it captures, and indicate which
letter shows more of it.

Judge each dimension on its own terms, from its own evidence. Applications have uneven
profiles: it is entirely normal for one letter to be stronger on some dimensions and the
other letter stronger on others, so let the letter that shows more of each dimension fall
wherever the evidence for that dimension points.

Out of scope — do not treat these as dimensions of difference:
- the applicant's name or any personal identifier;
- the closing sentence about work authorization / visa sponsorship, which is identical in
every letter;
- length in itself.

How to report. List each dimension as exactly one item — do not combine dimensions with
"and", "/", "&", or a comma; split related dimensions into separate items. For the evidence
field, give a short phrase naming the concrete feature you saw. For the dimension field,
give a concise general noun phrase naming the dimension (the kind of short label you might
use as a rubric column). For the description field, give one sentence saying what that
dimension captures. Indicate which letter shows more of it.

Output only JSON of the form:
{"differences": [{"evidence": "<short phrase>", "dimension": "<concise noun phrase>",
"description": "<one sentence>", "winner": "A" | "B"}, ...]}
```

Prompt 5: Comparative method (ours) – feature taxonomy.

```
You are organizing a flat list of evaluation dimensions into a set of canonical features.

These dimensions were produced by judges comparing job applicants' cover letters for the
same software-engineering role. Each judge, on each comparison, named the dimensions on
which two letters differed. Aggregated across many comparisons, the same underlying
construct was named in many slightly different ways. Your job is to group the different
phrasings of the SAME underlying construct together.
```

Below is the full list. Each entry has an INDEX in square brackets, a short NAME, a one-sentence DESCRIPTION of what it captures, and CITATIONS (how many times it was named — for your awareness only).

⟨*indexed dimension list*⟩

Instructions:
- Group entries that denote the SAME underlying evaluative construct, judging by MEANING, not by surface wording. Different words for the same idea belong in one group; the same word used for different ideas belongs in different groups.
- Do NOT force everything into a small number of groups, but do NOT split one construct into its facets either. A dimension that does not clearly share a construct with any other does NOT become its own feature: place all such genuinely-apart, idiosyncratic dimensions together in a single catch-all group whose canonical name is exactly "Other". Do not create many single-member groups, and do not absorb a genuinely-apart dimension into a loosely-related group to tidy up — send it to "Other".
- Do NOT merge two constructs just because they often co-occur or because strong applicants tend to score well on both. Group only by shared MEANING, never by how applicants might score.
- Citations are context only. Do not merge a frequently-named dimension into a rarely-named one or vice versa on the basis of counts; counts do not determine grouping.

For each group, give:
- a canonical NAME (a concise noun phrase suitable as a rubric column; use exactly "Other" for the catch-all group),
- a one-sentence DEFINITION of the construct the group captures,
- the list of members, each as its INDEX and its NAME copied verbatim from that index.

Rules on membership (these will be checked programmatically):
- Every input index must appear in exactly one group, exactly once. Place all of them; drop none; duplicate none.
- Do not invent indices that are not in the input.
- Each member's "name" must be the exact input NAME at that index — this is a cross-check, so copy it from the bracketed entry, do not paraphrase.

Output only JSON:
{"features": [{"name": "<canonical noun phrase>", "definition": "<one sentence>", "members": [{"index": <int>, "name": "<verbatim input name at that index>"}, ...]}, ...]}

Prompt 6: Pointwise baselines (Direct LLM/Local LLMs) – rubric discovery prompt.

You are helping build an interpretable scoring framework by examining how job applicants' cover letters differ. There is no fixed rubric — the dimensions are recovered from what actually distinguishes these letters, not specified in advance.
=== JOB POSTING ===
⟨*full job posting*⟩
=== END JOB POSTING ===
Below is a sample of cover letters submitted for this role, selected to span the full range of applicant quality.
=== COVER LETTERS ===
⟨*cover letters*⟩
Reading across them, attend to the ways the letters differ from one another — the axes along which you can see one letter showing more, less, or something different from another. Identify the dimensions on which these letters meaningfully differ as applications for this role. Report whatever genuinely stands out to you; do not work from a predetermined list, and do not force the list to be short — report as many dimensions as actually distinguish these letters.
A dimension may be something a letter shows more or less of along a range, or something a letter either shows or does not.
Keep genuinely distinct dimensions separate even when they are related or tend to co-occur — do not merge two dimensions because strong letters tend to show both. List each dimension

```
as exactly one item — do not combine dimensions with "and", "/", "&", or a comma; split
related dimensions into separate items.
For each dimension, first name the concrete, specific feature you actually saw in the
letters — a short phrase pointing to the evidence. Then, grounded in that evidence, name
the dimension and describe in one sentence what it captures. Then write a short assessment
guideline a future reader will use to score a new letter on this dimension on a scale you
need to specify: say what to look for, and describe what distinguishes each score of the
scale you come up with.
Out of scope — do not treat these as dimensions of difference:
- the applicant's name or any personal identifier;
- the closing sentence about work authorization / visa sponsorship, which is identical in
every letter;
- length in itself.
Output only JSON:
{"features": [{"evidence": "<short phrase>", "name": "<concise noun phrase>",
"description": "<one sentence>", "levels": [{"score": 1, "descriptor": "<a letter at the
weakest point of your scale>"}, {"score": 2, "descriptor": "<...>"}, ... up to the top
score of the scale you chose]}, ...]}
— for each dimension choose your own scale (score 1 = weakest, up to a top point N you
decide) and give one descriptor per score point, from 1 to N.
```

Prompt 7: Local LLMs – feature fusion prompt.

```
You are organizing a flat list of evaluation dimensions into a set of canonical features.

These dimensions were produced by independent rubric elicitations over job applicants'
cover letters for the same software-engineering role. Each elicitation read its own random
sample of letters and, working from scratch, named the dimensions on which those letters
meaningfully differed. Aggregated across many independent elicitations, the same underlying
construct was named in many slightly different ways. Your job is to group the different
phrasings of the SAME underlying construct together.

Below is the full list. Each entry has an INDEX in square brackets, a short NAME, a
one-sentence DESCRIPTION of what it captures, and RUN (which elicitation produced it — for
your awareness only).
```

⟨*indexed dimension list*⟩

```
Instructions:
- Group entries that denote the SAME underlying evaluative construct, judging by MEANING,
not by surface wording. Different words for the same idea belong in one group; the same
word used for different ideas belongs in different groups.
- Do NOT force everything into a small number of groups, but do NOT split one construct
into its facets either. A dimension that does not clearly share a construct with any other
does NOT become its own feature: place all such genuinely-apart, idiosyncratic dimensions
together in a single catch-all group whose canonical name is exactly "Other". Do not create
many single-member groups, and do not absorb a genuinely-apart dimension into a
loosely-related group to tidy up — send it to "Other".
- Do NOT merge two constructs just because they often co-occur or because strong applicants
tend to score well on both. Group only by shared MEANING, never by how applicants might
score.
- Run labels are context only. Do not merge a dimension named by many runs into one named
by few, or vice versa, on the basis of how often it recurs; recurrence does not determine
grouping. Two entries from the SAME run were named as distinct dimensions by that run, so
they are unlikely to be the same construct — but judge by MEANING, and merge them if they
plainly are.

For each group, give:
- a canonical NAME (a concise noun phrase suitable as a rubric column; use exactly "Other"
for the catch-all group),
- a one-sentence DEFINITION of the construct the group captures,
- the list of members, each as its INDEX and its NAME copied verbatim from that index.
```

```
Rules on membership (these will be checked programmatically):
- Every input index must appear in exactly one group, exactly once. Place all of them; drop
none; duplicate none.
- Do not invent indices that are not in the input.
- Each member's "name" must be the exact input NAME at that index — this is a cross-check,
so copy it from the bracketed entry, do not paraphrase.

Output only JSON:
{"features": [{"name": "<canonical noun phrase>", "definition": "<one sentence>",
"members": [{"index": <int>, "name": "<verbatim input name at that index>"}, ...]}, ...]}
```

Prompt 8: Pointwise baselines (Direct LLM/Local LLMs) – scoring prompt template.

```
You are scoring a cover letter against a fixed rubric, for the software-engineering role
below.
=== JOB POSTING ===
⟨full job posting⟩
=== END JOB POSTING ===
=== COVER LETTER ===
⟨letter⟩
⟨instruction variant⟩
A dimension you cannot fully assess from this letter should still receive your best
estimate, not a default middle score. Score each dimension independently, and do not let
the score on one dimension affect the scores on others. Applications have uneven profiles:
it is possible for a letter to score low on some dimensions and high on others.
=== RUBRIC DIMENSIONS (score in the order listed; each lists the descriptors for its own
score scale) ===
⟨rubric, in scoring order⟩
Out of scope — do not let these affect any score:
- the applicant's name or any personal identifier;
- the closing sentence about work authorization / visa sponsorship, which is identical in
every letter;
- length in itself.
⟨output specification⟩
```

Prompt 9: Pointwise scoring – the four reasoning-path instruction variants.

[direct]

```
Score the letter on each rubric dimension below independently. Each dimension lists a
descriptor for every point of its own score scale; assign the integer score whose
descriptor best matches the letter on that dimension. Judge each dimension on its own.
```

[evidence_first]

```
For each rubric dimension below, briefly note what in the letter bears on that dimension,
then assign the integer score, from that dimension's own listed scale, whose descriptor
best matches the letter on that dimension. Treat each dimension independently.
```

[standard_referenced]

```
For each rubric dimension below, compare the letter against the score descriptors listed
for that dimension, then assign the integer score whose descriptor it matches most closely.
Judge each dimension on its own.
```

[independent]

```
Go through the rubric dimensions one at a time, in the order listed. For each dimension,
consider only what the letter shows on that specific dimension and assign the integer
score, from that dimension's own listed scale, whose descriptor best matches; do not form
an overall impression of the letter and do not let the other dimensions influence the score.
```

C.3 Additional Results

C.3.1 Model calibration

In this subsection, we study whether or not the statistical model is well-calibrated by computing its Expected Calibration Error (ECE) and its Brier skill score against various baselines.

The fitted model assigns to each (comparison, feature) cell two probability forecasts of observable binary events: the citation gate $\hat{g}_{p,k} = \sigma\big(f(\hat{\Delta}_{p,k}) + \hat{\gamma}_k\big)$, the probability that feature k is cited on comparison p, defined on all $84{,}000$ cells ($K = 21$ features), and the direction gate $\hat{d}_{p,k} = \sigma\big(\hat{\Delta}_{p,k} + \hat{\beta}\big)$, the probability that the first-shown letter is favored given citation, defined on the $29{,}030$ cited cells. We score both as forecasts, out of fold: 5-fold cross-validation over comparisons, so that every cell is predicted by parameters estimated without it, with 95% bootstrap intervals over comparisons throughout; in-sample counterparts quantify the optimism of the fit.

Figure 13 reports reliability: cells are grouped into 10 equal-width bins of predicted probability, the observed event rate in each bin is compared against the diagonal of perfect calibration, and the expected calibration error (ECE) is the occupancy-weighted mean absolute gap. The direction gate is calibrated essentially everywhere: its out-of-fold ECE is 0.010 $[0.008, 0.014]$ (in sample: 0.019), and its predictions populate the full unit interval including the extreme bins, so the diagnostic is exercised where it is most demanding. The citation gate's out-of-fold ECE is 0.046 $[0.044, 0.049]$ against 0.030 in sample, and its reliability curve is flatter than the diagonal (observed rates exceed predictions in the low bins and fall short in the high ones) so the fitted gate is too extreme at both ends, consistent with the amplitude of f slightly overshooting out of fold rather than the gap dependence being absent.

Figure 14 reports discrimination through the Brier skill score, $1 - \mathrm{Brier(model)}/\mathrm{Brier(reference)}$, against two references estimated on the training folds: chance ($p = 1/2$, Brier 0.25) and the empirical rate (the per-feature citation rate for the citation gate, the per-feature rate of favoring the first-shown letter among cited cells for the direction gate). The second reference is the demanding one. For the citation gate it is the maximum-likelihood fit of the gap-independent submodel $f = 0$, whose gate $\sigma(\gamma_k)$ is constant within each feature, so skill over it is a direct out-of-fold test of whether the merit gap carries information about citation. It does: the gate scores 0.069 $[0.063, 0.076]$ against this submodel, rejecting $f = 0$, down from 0.168 in sample; its 0.266 $[0.260, 0.272]$ against chance partly rewards knowing the citation base rate rather than per-cell discrimination. The direction gate scores 0.548 $[0.538, 0.557]$ against chance and 0.545 $[0.536, 0.554]$ against the empirical rate (the two coincide because slot balance keeps the first-shown rates near $1/2$) and retains most of its in-sample skill (0.623). The channel that carries the scores is thus both calibrated and strongly predictive out of sample; the gap effect on citation is significant but second-order, estimated with an amplitude that overshoots out of fold.

C.3.2 Discovered Features

Tables 4 and 5 report the features discovered by each method described in Table 1. More precisely, Table 4 presents each feature with its reliability coefficients and its fitted γ coefficients, Table 5 shows the detailed rubric created by GPT-5.4 (HIGH REASONING).

As we can see in Table 4, our approach yields fine-grained features, as the judge distinguishes the existence of side projects ("Side-project initiative") from their magnitude ("Side-project ambition", which almost passed the reliability thresholds). Moreover, the reliability diagnostics correctly filter features such as "Narrative framing and tone", which is a well-identified feature ($\kappa = 4.9$) but whose citations are inconsistent with the estimated scores ($\rho = 0.72 < \rho_{\min}$): a sign of a "boilerplate" feature that the judge cites independently from the items it is comparing. This is partly captured by the γ coefficient. For example, "Employer prestige" has a smaller γ than "Narrative framing and tone" but is reliable: a sign that the former is discriminative while the latter is a boilerplate feature.

We also notice in Table 5 that the rubric correctly captures some of the dials' levels. For example, the judge correctly identifies that for the Education dial, BS $<$ MS $<$ PhD: the rubric is correctly tailored to the samples that the judge sees.

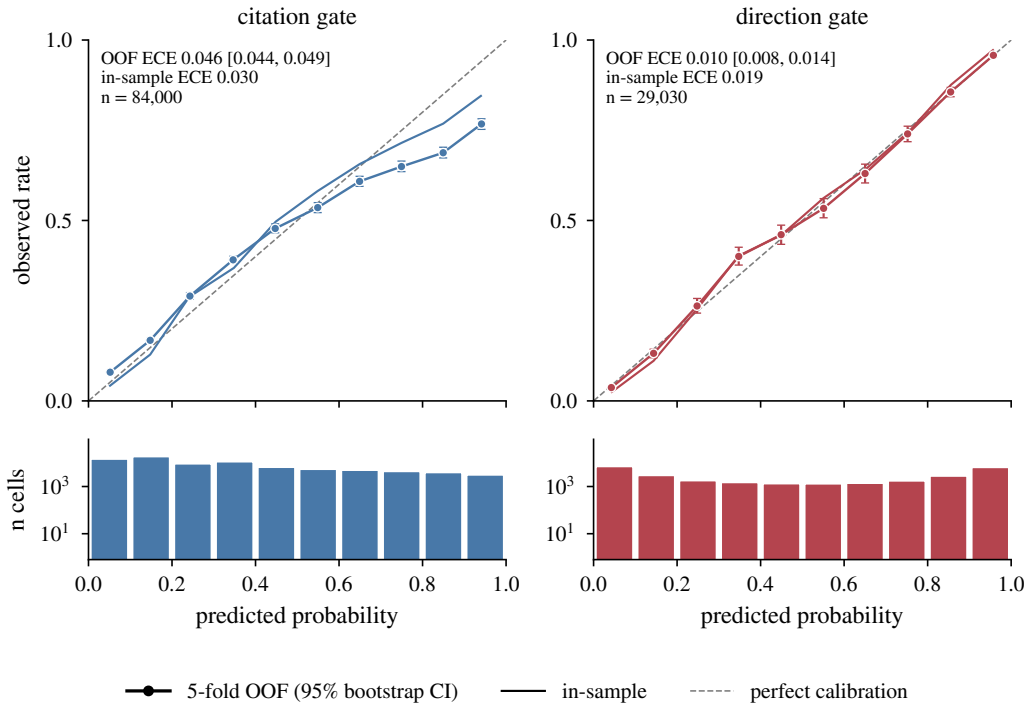

Figure 13: Out-of-fold reliability of the citation gate (left, all cells) and direction gate (right, cited cells) of the feature-discovery model. Points: observed rate per bin of predicted probability, with 95% bootstrap intervals over comparisons; faint line: in-sample; dashed: perfect calibration; bottom: cells per bin (log scale).

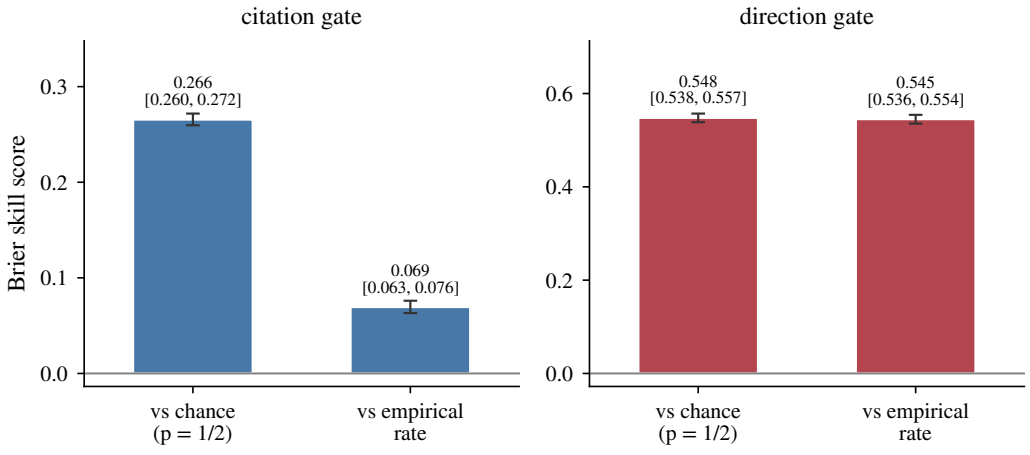

Figure 14: Out-of-fold Brier skill of the two gates against chance and against the training-fold empirical rates; intervals as in Figure 13.

feature	Cite rate	γ_k	κ	ρ	Kept
AI experience	0.613	-1.36 $[-1.45, -1.27]$	3.6	0.937	✓
Security assessment experience	0.596	-1.37 $[-1.46, -1.28]$	2.9	0.935	✓
Security automation experience	0.562	-1.31 $[-1.40, -1.22]$	2.4	0.917	✓
Agentic systems experience	0.544	-1.52 $[-1.61, -1.43]$	2.7	0.924	✓
Professional experience	0.531	-2.43 $[-2.55, -2.31]$	3.8	0.960	✓
Algorithmic strength	0.524	-1.48 $[-1.57, -1.39]$	2.7	0.917	✓
Security certification	0.445	-2.54 $[-2.66, -2.43]$	4.0	0.942	✓
Education level	0.414	-2.75 $[-2.86, -2.64]$	3.3	0.943	✓
Open-source contribution	0.392	-2.45 $[-2.55, -2.34]$	3.9	0.924	✓
Off-hours security practice	0.382	-2.11 $[-2.21, -2.02]$	3.3	0.891	✓
Python experience	0.365	-2.05 $[-2.14, -1.95]$	2.7	0.873	✓
Programming language breadth	0.293	-3.13 $[-3.26, -3.01]$	2.6	0.925	✓
Prestige signaling	0.271	-2.57 $[-2.68, -2.46]$	3.8	0.855	✓
Production delivery scope	0.264	-2.84 $[-2.96, -2.72]$	2.8	0.884	✓
Mission and company motivation	0.258	-2.37 $[-2.48, -2.27]$	3.7	0.770	✓
Platform and developer enablement orientation	0.192	-2.96 $[-3.08, -2.83]$	5.8	0.771	✓
Go experience	0.191	-2.98 $[-3.11, -2.85]$	5.0	0.799	✓
Security philosophy	0.133	-3.79 $[-3.97, -3.61]$	9.4	0.801	✓
Motivation narrative	0.121	-4.16 $[-4.37, -3.95]$	8.8	0.806	✓
Side-project initiative	0.094	-4.74 $[-5.02, -4.47]$	8.1	0.850	✓
Employer prestige	0.074	-5.13 $[-5.45, -4.80]$	8.3	0.820	✓
Side-project ambition	0.194	-3.40 $[-3.55, -3.24]$	10.5	0.866	✗
Narrative framing and tone	0.159	-3.10 $[-3.24, -2.97]$	4.9	0.721	✗
Systems and technical depth	0.112	-3.72 $[-3.89, -3.54]$	11.6	0.699	✗
General security grounding	0.093	-3.88 $[-4.07, -3.69]$	7.8	0.652	✗
Role alignment	0.091	-3.80 $[-3.98, -3.62]$	10.3	0.580	✗
Impact quantification	0.068	-4.74 $[-5.02, -4.46]$	12.8	0.731	✗
Maintenance orientation	0.053	-8.55 $[-9.87, -7.23]$	198.5	0.862	✗
Closing posture	0.050	-4.67 $[-4.95, -4.38]$	10.5	0.564	✗
Learning posture	0.047	-6.61 $[-7.22, -6.00]$	56.1	0.762	✗
Engineering rigor	0.028	-8.86 $[-10.20, -7.51]$	228.0	0.831	✗
Other	0.012	-8.54 $[-9.60, -7.49]$	92.0	0.036	✗
Writing polish	0.009	-9.51 $[-11.26, -7.76]$	156.6	0.440	✗

Table 4: Features surfaced by our comparative method in the point estimate of Table 1. Of 32 surfaced features, we keep the features that have $\rho \geq \rho_{\min} := 0.75$ and $\kappa \leq \kappa_{\max} := 10$, leaving 21 features. Fitted $\hat{\beta} = +0.005$ (95% CI $[-0.033, +0.043]$). CIs for $\hat{\beta}$ and γ coefficients were computed using (B.3).

Feature	Description	Scoring guidelines
Academic degree	Captures the formal level of academic preparation or equivalent technical grounding the applicant presents.	**1** No clear qualifying degree or equivalent preparation is shown. **2** Shows a bachelor's degree or equivalent practical preparation. **3** Shows a master's degree in a relevant technical field. **4** Shows a doctoral degree in a relevant technical field.
Adversarial practice	Captures ongoing offensive-security practice through CTF participation and competitive results.	**1** No ongoing offensive-security practice is shown. **2** Mentions occasional CTF participation or light adversarial practice. **3** Shows regular or serious CTF participation as continuing skill development. **4** Shows strong competitive results, such as near-top or top placements in substantial CTF fields.
Agent boundary design	Captures experience designing agentic systems with explicit control boundaries between Python agents and Go services.	**1** No agentic tooling or agent architecture is described. **2** Mentions agents or agentic systems only at a high level. **3** Describes building agentic systems with Python agents and Go back-end support. **4** Describes explicit guardrails such as bounded actions, permissions, auditing, or enforced boundaries between agents and services.
Algorithmic depth	Captures depth of data-structure and algorithm reasoning as evidenced by concrete redesign or optimization examples.	**1** Does not provide evidence of algorithmic or data-structure skill. **2** Claims strong algorithms knowledge, but with little concrete illustration. **3** Provides a clear example of using better algorithms or data structures to improve a system. **4** Provides a strong example with technical specificity and visible impact, such as removing a bottleneck or making a previously hard problem tractable.
Applied AI deployment	Captures whether the applicant has shipped AI or LLM features in production and understands reliability or evaluation concerns.	**1** No meaningful AI or LLM experience is shown. **2** Mentions AI interest or exposure without a clear shipped production example. **3** Shows direct experience shipping AI or LLM-backed functionality in production. **4** Shows shipped AI or LLM systems plus concrete attention to evaluation, reliability, or failure modes.
Developer enablement mindset	Captures whether the applicant frames security tooling as a way to help developers do the right thing with less friction.	**1** Does not frame security work in terms of helping other engineers. **2** Mentions helping teams only in broad or generic language. **3** Clearly emphasizes making secure behavior easier or more natural for developers. **4** Explicitly focuses on reducing friction, building good defaults, or serving busy developer teams through usable tooling.
Evidence specificity	Captures how consistently the letter backs claims with specific examples, mechanisms, or project details.	**1** Relies mostly on generic assertions and unsubstantiated claims. **2** Mixes some specifics with many broad claims. **3** Uses several concrete examples, technologies, or project details to support claims. **4** Consistently grounds claims in precise examples, mechanisms, or scenario-level detail throughout the letter.
Impact articulation	Captures whether achievements are described in terms of concrete effect on team capacity, risk reduction, or manual workload.	**1** Mostly describes activities performed, with little indication of why they mattered. **2** Hints at benefits, but leaves the concrete effect largely implicit. **3** Explains clear outcomes such as reduced manual work, better reliability, or improved throughput. **4** Makes impact vivid and credible through concrete consequences, strong before-after framing, or quantified value.
Maintainer mindset	Captures whether the applicant brings a maintainer's perspective shaped by testing, inherited systems, failure modes, and long-term operability.	**1** Does not show a maintenance or reliability perspective. **2** Mentions testing, maintenance, or inherited systems without drawing lessons from them. **3** Shows that maintenance or failure-mode experience informs how the applicant designs systems. **4** Strongly presents a maintainer identity centered on operability, inheritance, failure handling, and long-term resilience.
Mechanical polish	Captures proofreading quality and whether spelling or typing errors distract from credibility.	**1** Contains frequent spelling, typing, or proofreading errors that materially distract. **2** Contains several noticeable errors that reduce polish. **3** Contains minor occasional errors, but remains generally professional. **4** Is mechanically clean, polished, and free of distracting errors.
Open-source stewardship	Captures contribution level to open-source software, from occasional patches to maintaining a community-used project.	**1** No open-source contribution is mentioned. **2** Mentions occasional fixes or patches to open-source dependencies. **3** Shows regular stewardship, such as maintaining a small project or handling releases. **4** Shows active maintenance of a community-used project, library, or framework with reviews, releases, and external users.
Personal systems initiative	Captures self-directed engineering initiative shown through side projects, especially ambitious distributed-systems or reliability-heavy builds.	**1** No meaningful side project or independent build is described. **2** Shows a modest personal project, utility, or app that demonstrates follow-through. **3** Shows a substantial self-directed engineering project with nontrivial design or reliability concerns. **4** Shows an ambitious systems project, especially distributed or failure-heavy work chosen to tackle hard engineering problems.

continued on next page

Table 5 continued

Feature	Description	Scoring guidelines
Platform leverage orientation	Captures whether the letter explicitly values foundational platform work with compounding downstream impact.	**1** Does not articulate interest in platform or org-wide leverage. **2** Mentions scale or impact in generic terms. **3** Clearly values foundational work whose benefits reach many teams or systems. **4** Repeatedly frames the role in terms of compounding downstream impact across a large organization like Google.
Production ownership	Captures the breadth of concrete production service launch and maintenance ownership the letter presents.	**1** Mentions production work vaguely or only limited ownership without clear launches. **2** Shows ownership of a small number of production services, with some testing, maintenance, or launch responsibility. **3** Shows clear responsibility across several production services, indicating repeated launch and maintenance experience. **4** Shows broad ownership across many production services, suggesting repeated end-to-end delivery in production.
Python fluency	Captures how strongly the letter shows Python as a real working language for services, automation, or tooling.	**1** Does not show meaningful Python experience. **2** Mentions Python, but only briefly or without clear production use. **3** Shows regular Python use for services, automation, analysis, or internal tooling. **4** Shows strong production-grade Python ownership, with concrete services or tooling others rely on.
Security assessment practice	Captures hands-on experience performing security reviews, threat modeling, code review, or validating remediation.	**1** No clear security assessment work is described. **2** Mentions security reviews generically, without concrete assessment methods. **3** Shows concrete assessment work such as threat modeling, code review, attack-surface mapping, or service reviews. **4** Shows repeated or formal assessment practice, including documented findings and validation of remediation or fixes.
Security credentialing	Captures external validation of security knowledge through named certifications and their rigor.	**1** No security certification is mentioned. **2** Shows a baseline security certification such as Security+. **3** Shows a stronger hands-on certification such as OSCP. **4** Shows multiple or especially rigorous certifications that materially strengthen the security case.
Security operations systems-thinking	Captures whether the letter sees security response as a scalable systems-design problem rather than ad hoc manual toil.	**1** Treats security interest generically, without a clear systems view of the problem. **2** Mentions automation or scale, but without a crisp model of the underlying bottleneck. **3** Frames vulnerability response or security toil as a systems-design or capacity problem. **4** Articulates a sharp leverage model, such as scaling scarce expert judgment, variant analysis, or turning heroic work into routine automation.
Security rigor ethos	Captures whether the letter articulates a clear high-standard philosophy about building security-critical software.	**1** Shows generic enthusiasm for security, but no clear quality philosophy. **2** Signals seriousness about correctness or standards in broad terms. **3** Articulates a distinct philosophy about the bar security tooling should meet. **4** Presents a memorable, role-relevant engineering creed that clearly matches high-assurance security work.
Software tenure	Captures how much relevant professional software experience in years the letter claims.	**1** Less than the role minimum or unclear how much professional software experience the applicant has. **2** Shows roughly entry-level relevant experience, such as about 2 to 3 years in professional software development. **3** Shows solid mid-level experience, such as about 4 to 5 years building and maintaining software. **4** Shows substantial experience, such as 6 or more years of directly relevant software development work.
Vulnerability workflow automation	Captures direct experience automating vulnerability intake, triage, deduplication, routing, or prioritization workflows.	**1** Shows no direct vulnerability-handling automation experience. **2** Expresses interest in automating vulnerability work, but without direct evidence of having done it. **3** Shows hands-on automation of vulnerability intake or triage, such as routing or queue handling. **4** Shows deep workflow automation with concrete functions like classification, deduplication, prioritization, routing, or major manual-load reduction.

Table 5: The rubric elicited by Direct LLM. Descriptions and the per-score guidelines are reproduced verbatim as the judge produced them; the open elicitation lets the judge choose its own scale, and it selected a 1–4 scale for every feature.

C.4 Run-variance analysis

C.4.1 Per-method run-variance bands

Table 6 repeats the point estimates of Table 1 together with each method's run-variance band. The gray bands are per-method run-variance summaries; each arm resamples the unit its own pipeline

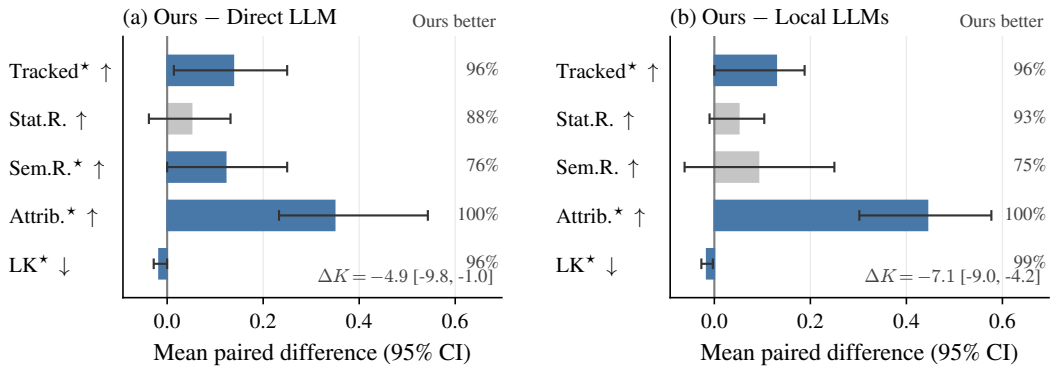

Figure 15: Paired comparisons over $B = 50$ bootstrap draws (with replacement) of the 200-letter corpus; every arm sees the same draws, and differences are taken within replicate. Each row shows the mean paired difference with its 95% percentile CI ($\star$: excludes zero); the right column gives *Ours better*, the share of replicates our method wins (ΔK, the difference of discovered features, is annotated below each panel and carries no share). **The panels bracket discovery variance from opposite sides and are not comparable to one another**: (a) re-runs discovery per replicate on both sides, (b) freezes discovery on both sides, pooled over fusion and corpus draws. **Both panels score our method at a reduced budget** (average degree 25.3 against the designed 40); it leads on several metrics regardless. Exact values: Table 7.

Method	Cost	K	Tracked	Stat.R.	Sem.R.	Attrib.	LK↓
Ours	3.45	21	**100.0%**	**0.850**	**100.0%**	**76.2%**	**0.090**
resampling matchings		[16.2, 22.0]	[87.5, 93.8]	[0.737, 0.805]	[81.2, 93.8]	[71.4, 87.5]	[0.087, 0.095]
Direct LLM	**0.26**	21	81.2%	0.732	81.2%	61.9%	0.109
resampling letters		[17.2, 25.0]	[68.8, 87.5]	[0.649, 0.825]	[62.5, 87.5]	[40.6, 69.6]	[0.109, 0.135]
Direct LLM–Multi-score	3.42	21	81.2%	0.787	81.2%	66.7%	0.118
resampling letters		[21.0, 21.0]	[81.2, 81.2]	[0.788, 0.807]	[81.2, 81.2]	[57.1, 66.7]	[0.119, 0.145]
Local LLMs	2.35	23	75.0%	0.702	68.8%	47.8%	0.110
resampling elicitation subsets		[22.2, 27.8]	[75.0, 87.5]	[0.696, 0.789]	[62.5, 87.5]	[44.0, 55.7]	[0.103, 0.120]
Local LLMs–Multi-score	3.27	23	75.0%	0.737	68.8%	47.8%	0.113
resampling letters		[23.0, 23.0]	[75.0, 75.0]	[0.737, 0.761]	[68.8, 68.8]	[39.1, 52.2]	[0.115, 0.136]

Table 6: Per-method point estimates with run-variance bands. Tracked, Sem.R., Attrib. in %; Stat.R. and LK are correlations; bands are in the same units as the column. Below each method's point estimates, the gray row gives that method's **own** run-variance band ($[2.5, 97.5]$ percentiles across replicates), obtained by resampling the unit named at the left of the row. These bands are **not** confidence intervals for the point estimates above them – the point estimates come from the single full-data run, the bands from re-runs on resampled data, so a point estimate may fall outside its band – and they **must not be compared across methods**, since the arms resample different units. Paired, comparable contrasts: Table 7.

randomizes over, so band widths reflect the design as much as the method and carry no cross-method meaning.

Ours resamples the 40 matchings that make up the comparison graph, and then re-runs the taxonomy on every replicate (the judging of the pairs is not rerun as it is too costly); the K band covers the reliable count after the κ/ρ filter, from 28.9 discovered features on average. Direct LLM resamples the letters and re-elicits its rubric per replicate. Direct LLM–Multi-score shares Direct LLM's rubric but its discovery step is deterministic, so its band reflects letter resampling only; the degenerate entries (K, Tracked, Sem.R.) are exact. Local LLMs redraws which E=27 elicitations are pooled, so its band carries fusion variance but holds the corpus fixed – unlike Direct LLM, whose band varies both. Local LLMs–Multi-score is instead the direct counterpart of Direct LLM–Multi-score: a single

	Difference	95% CI	Ours better
(a) Ours − Direct LLM (discovery re-run per replicate on *both* sides)			
ΔK	-4.90	$[-9.78, -1.00]$	—
Δ **Tracked***	$+0.138$	$[+0.014, +0.250]$	96%
Δ Stat.R.	$+0.051$	$[-0.038, +0.132]$	88%
Δ **Sem.R.***	$+0.122$	$[+0.000, +0.250]$	76%
Δ **Attrib.***	$+0.349$	$[+0.233, +0.543]$	100%
Δ **LK***	-0.017	$[-0.028, -0.000]$	96%
(b) Ours − Local LLMs (discovery frozen on *both* sides; pooled over fusion *and* corpus draws)			
ΔK	-7.08	$[-9.00, -4.23]$	—
Δ **Tracked***	$+0.129$	$[+0.000, +0.188]$	96%
Δ Stat.R.	$+0.051$	$[-0.010, +0.104]$	93%
Δ Sem.R.	$+0.092$	$[-0.062, +0.250]$	75%
Δ **Attrib.***	$+0.444$	$[+0.302, +0.577]$	100%
Δ **LK***	-0.016	$[-0.027, -0.003]$	99%

Table 7: Paired comparisons over $B = 50$ bootstrap draws (with replacement) of the 200-letter corpus; every arm sees the same draws, and differences are taken within replicate. Each row reports the mean difference, its 95% percentile CI (*: excludes zero), and *Ours better*: the share of replicates in which our method wins in the metric's preferred direction (ΔK is descriptive and carries no share). **The panels bracket discovery variance from opposite sides and are not comparable to one another**. **Both panels score our method at a reduced budget (average degree 25.3 against the designed 40)**; it leads on several metrics regardless. Design rationale, the conditional-fusion alternative, and budget details: Appendix C.4.2.

fusion with deterministic discovery, scored with 6 repeated passes averaged before resampling, so its band too reflects letter resampling only, with K, Tracked and Sem.R. again exact.

C.4.2 Design of the paired comparisons

Panel (b) of Figure 15 cannot mirror panel (a). Local LLMs discovery replicates are indexed by the choice of E=27 pooled elicitations rather than by letters, so they admit no pairing with corpus draws, and re-eliciting on each resampled corpus would be prohibitively costly. In (a) both arms re-run discovery on every replicate, so the intervals include discovery variance on both sides; in (b) our taxonomy is held at the canonical fit while Local LLMs's fusion is varied over all 50 draws, the 50×50 (fusion, corpus) differences being pooled; the intervals therefore carry corpus variance for both arms and fusion variance for Local LLMs, but no re-discovery for ours.

Conditioning instead on the canonical fusion (the first E=27 elicitations) gives uniformly larger, all-significant gaps (Tracked $+0.177$, Stat.R. $+0.080$, Sem.R. $+0.177$, Attrib. $+0.466$, LK -0.016), that fusion being among the weaker draws; the pooled figures in Table 7 are the **conservative** choice. Across fusions, the mean differences span Tracked $[+0.052, +0.177]$, Stat.R. $[-0.003, +0.093]$, Sem.R. $[-0.010, +0.240]$, Attrib. $[+0.362, +0.505]$, LK $[-0.023, -0.007]$. Resampling the corpus also reduces our method's comparison budget: a replicate retains only judged pairs with both endpoints drawn (1,603 of 4,000 on average, consistent with the $\approx 0.632^2$ both-endpoints survival rate of an n-out-of-n bootstrap), giving an average degree of 25.3 against the designed 40, on which our method reports ≈ 16 features on average rather than 21. **The advantage does not rest on the larger budget**: at degree 25 the comparative arm costs ≈\$2.24, against \$3.45 at degree 40 and \$2.35 for Local LLMs.

C.4.3 Dials Recovered

Figures 16 and 17 compare the dial-recovery heatmap between our method and Direct LLM. As we can see, our method yields a strong diagonal recovery with only a few discovered features tracking no known dial, while Direct LLM clearly misses 3 dials and has dials tracked by several discovered features.

Remark 7. As we can see in Figure 18, after $E = 27$ rubric elicitations, recovering every dial demands a high acceptance threshold, which leaves an unusable number of near-duplicate features.

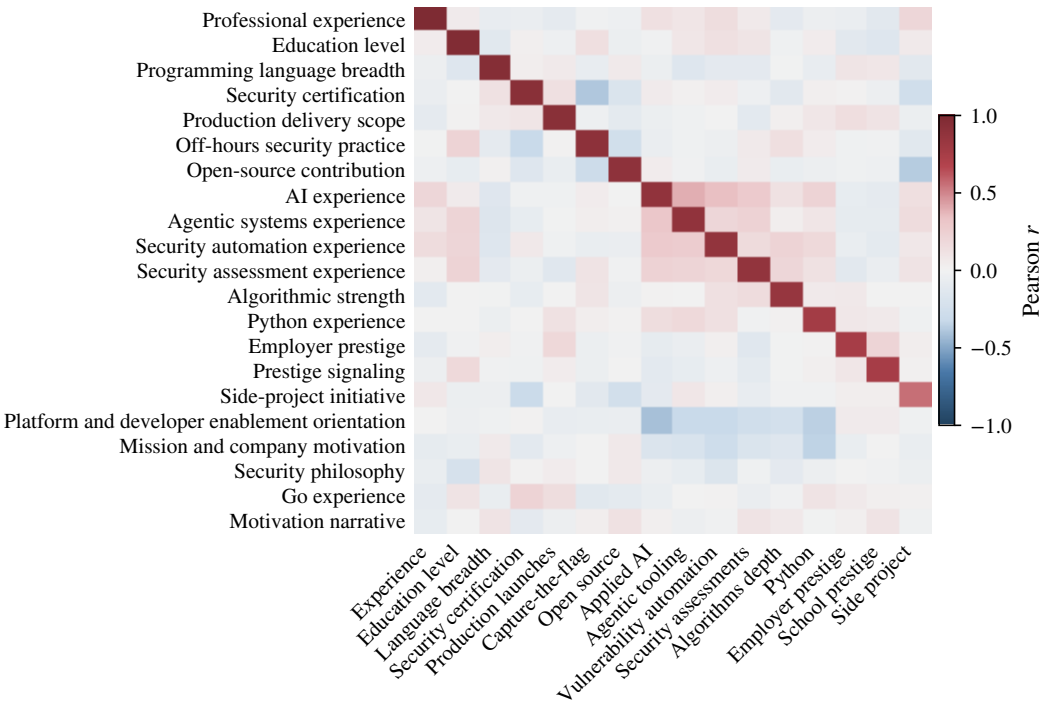

Figure 16: Dial-recovery heatmap yielded by the features discovered by our method. The rows are the discovered features and the columns are the dials.

It is the same problem that led us to use LLM taxonomy instead of semantic clustering in our own standardization step: the judge naturally merges duplicate features and excludes idiosyncratic ones.

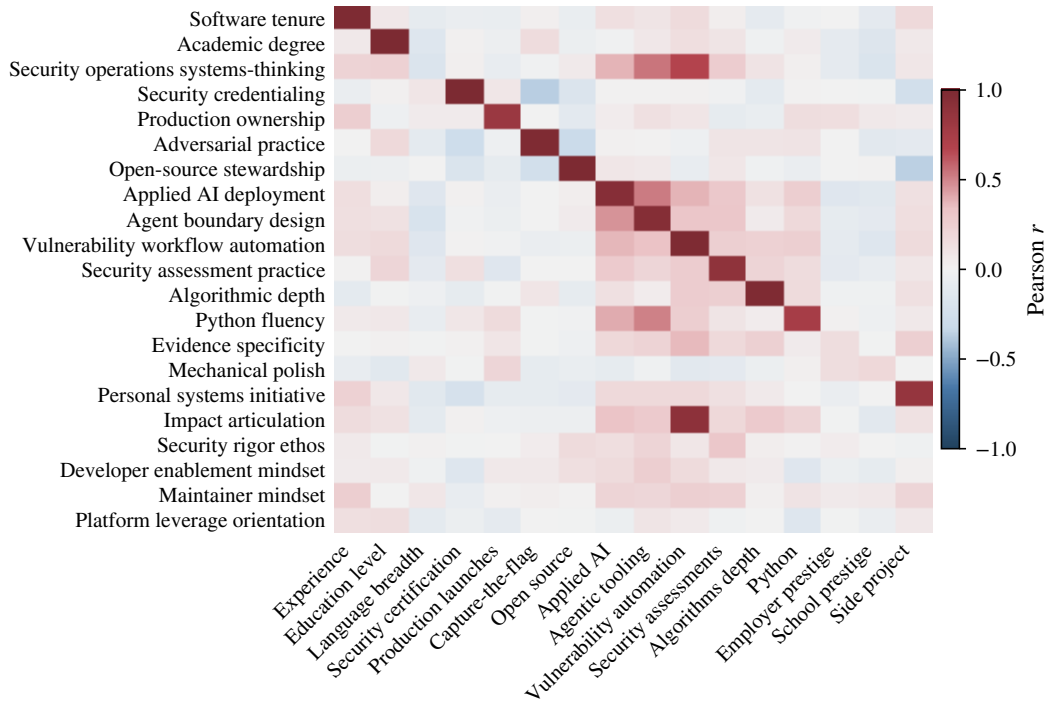

Figure 17: Dial-recovery heatmap yielded by the features discovered by Direct LLM. The rows are the discovered features and the columns the dials. The scores leak more between one another and the dial recovery is less clean than in Figure 16 as some dials are not recovered.

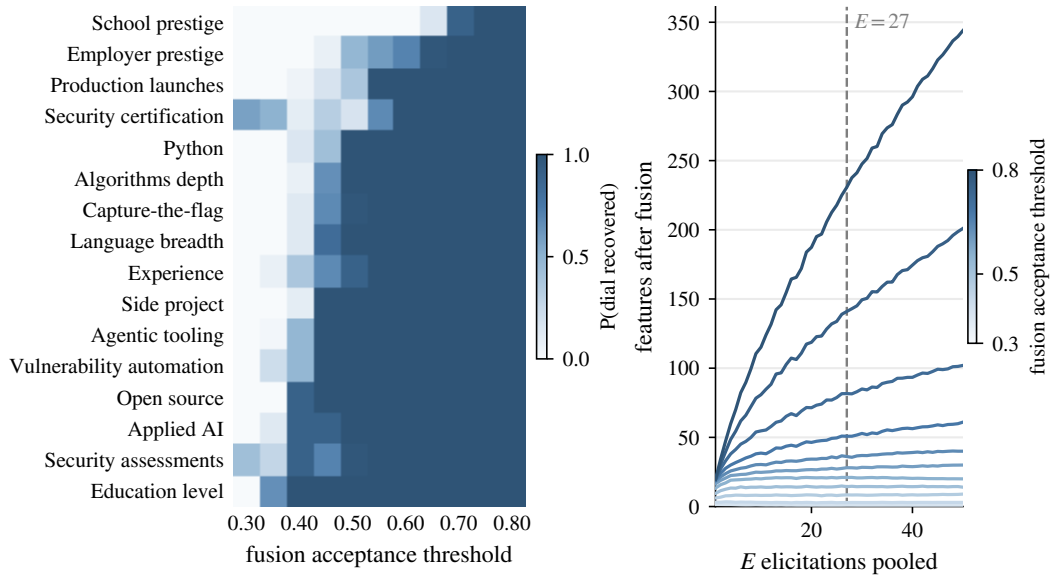

Figure 18: Left: probability that a dial is recovered as a function of the acceptance threshold used to merge similar features across the $E = 27$ elicitations of Local LLMs (dials are matched to the merged features by thresholded Hungarian assignment at the calibrated η). Right: number of features surviving the merge, against the number of pooled elicitations, one curve per acceptance threshold. At $E = 27$, recovering every dial demands a high threshold, which leaves an unusable number of near-duplicate features, hence the need for an LLM taxonomy.

Framing	Accuracy
P1 (unnamed, scrubbed)	53.6% [51.2, 56.1]
P2 (named)	53.6% [51.1, 56.1]
Lift P2 − P1	−0.1 pp [−2.4, +2.3]

Table 8: Memorization probe on 1,600 CMV exchanges pinned to the data of Table 2 (judge: GPT-5.4 MINI; chance = 50%). The lift is assessed by a paired McNemar test on the same exchanges ($b/c = 185/184$, $p = 1.000$).

D EXPERIMENT DETAILS ON THE CHANGEMYVIEW DATASET

D.1 DATASET

D.1.1 SAMPLES USED

Within each discussion tree, Tan et al. (2016) matched each Δ-winning argument to the non-winning argument from the same tree whose opening reply is most topically similar, yielding a balanced task (3,456 training and 807 held-out *pairs* of Δ-winning/non Δ-winning answers) in which topic is held fixed and the goal is to identify the persuasive member of each pair.

To mitigate length bias, we restricted to pairs whose two root replies both fall within $[20, 1200]$ words, leaving 3,411 pairs for the training set and 800 eligible held-out pairs. For our study, we used a sub-sample of 400 pairs from the training set.

D.1.2 MEMORIZATION PROBE

Because the CMV corpus predates the judge's training cutoff, a skeptical reading of Table 2 is that the judge recalls which replies earned a Δ rather than inferring it. To test this, we use 1,600 exchanges in four cells of 400 ($\{\Delta, \text{no-}\Delta\} \times \{\text{train}, \text{held-out}\}$, base rate 50% by construction). Each exchange is judged under two framings on identical de-identified text (Prompts 10 and 11): an *unnamed* framing (P1) asking whether the reply changed the poster's view, with platform identifiers scrubbed, and a *named* framing (P2) that presents the same scrubbed text but identifies the source as r/ChangeMyView and explains the Δ mechanic. A memorized outcome is easier to retrieve when the source is named, so the lift P2 − P1 on constant text is the recall signal.

Naming the source buys the judge nothing: the lift is −0.1 pp with a 95% confidence interval of $[-2.4, +2.3]$ pp (Table 8). Finally, the residual above-chance accuracy of the unnamed framing (53.6%, CI $[51.2, 56.1]$) is not recall-shaped: it is driven by a strong asymmetry across the outcome split (69.4% on no-Δ exchanges versus 37.9% on Δ exchanges under P1), a systematic prior toward predicting "no change". Consistently, naming the source (P2) does not sharpen either cell but shifts the threshold, raising Δ-cell accuracy to 43–46% while lowering no-Δ accuracy to 62–64%.

We conclude that the predictive result of Table 2 is not explained by outcome recall: on the very exchanges it is measured on, the judge shows no recall signal under source naming.

Prompt 10: Contamination probe – shared task framing (P1; unnamed).

```
You are shown one online discussion. It begins with a person's stated
VIEW (a position they hold and are willing to reconsider), followed by
ONE reply from someone arguing against that view. Your job: decide
whether that reply actually changed the original person's view. Answer
strictly yes or no.
=== VIEW ===
⟨OP view, de-identified⟩
=== REPLY ===
⟨root reply, de-identified⟩
⟨output specification⟩
```

Prompt 11: Contamination probe – named-source preamble (P2).

```
This is a post from the r/ChangeMyView (CMV) forum on Reddit. There, an
original poster states a view and awards a "delta" (Δ) to
any reply that succeeds in changing their view.
```
[followed by the shared task framing ofPrompt 10, on identical de-identified text]

D.2 METHODS DETAILS

This subsection gives all the details of each method.

Shared evaluation protocol. All arms score the same $2{,}400$ exchanges (each an (OP view, root reply) pair), which come from $1{,}200$ of Tan's matched pairs, each contributing the delta-winning reply and its matched loser: 400 pairs (800 exchanges) drawn from the training split and all 800 eligible held-out pairs ($1{,}600$ exchanges), with replies of 20–$1{,}200$ words. Given a feature set's per-exchange scores, the paired task is an ℓ_1 logistic regression on the within-pair score difference $\text{sign} \cdot (s_{\text{winner}} - s_{\text{loser}})$ with no intercept, the sign dealt from a balanced random sign per pair so chance is exactly 50%. The penalty λ in the ℓ_1 model is selected by 5-fold GROUPKFOLD cross-validation grouped by original poster over 13 log-spaced values $\lambda \in \{10^{-3}, \ldots, 10^{3}\}$, fitted on the 400 training pairs and evaluated once on the 800 held-out pairs. Selected values are $\lambda = 10$ (D + BT), 10 (D + PW), 31.6 (D_r + BT), 31.6 (D_r + PW), 31.6 (P + BT), 3.2 (P + PW), 3.2 (S + PW), 31.6 (Tan), 31.6 (Emb.) and 31.6 (#words).

D: feature discovery (shared by D + BT, D + PW, D_r + BT, D_r + PW). Features are mined from 150 training pairs (300 exchanges) drawn from the training split only. The 300 exchanges are arranged in a 40-regular comparison graph built as a union of 40 random perfect matchings, giving $6{,}000$ cross-topic comparisons performed by GPT-5.4 MINI using Prompt 12; a thread's own winner–loser edge is forbidden, so the judge never compares two replies to the same post. For each comparison the judge names, in free text, the dimensions on which the two exchanges differ in ways bearing on persuasion, with an evidence phrase and which side shows more of it (7.65 dimensions named per comparison; $45{,}894$ citations over $17{,}724$ distinct phrases). Phrases cited at least 10 times are grouped into canonical features by one GPT-5.4 (HIGH REASONING) call (Prompt 13), yielding 45 features. Excluding OP-side features (they describe the shared post and so cancel within a matched pair), the catch-all "Other", and one singleton label leaves the **29 challenger-side features** used in the (D) feature set. Fitting the statistical model of Section 3 and keeping features having $\rho \geq \rho_{\min} := 0.75$ and $\kappa \leq \kappa_{\max} := 50$ yields the feature set D_r of 16 features. Feature set: Table 10.

P: prior-elicited rubric (shared by P + BT and P + PW). A 16-feature set is elicited from the model's prior, with no data. GPT-5.4 (HIGH REASONING) is given a two-sentence setting (that people state views online and others reply to change them) and asked to name dimensions for assessing persuasiveness, each a matter of degree and distinct from the others, with a name and a one-sentence definition (Prompt 15). Asked instead to choose the count itself, the model produces 15.65 ± 1.42 features over two independent runs of 10 draws, so for a canonical run, we asked it to produce 16 dimensions, which is hence not a binding constraint and matches the size of the D_r set. Feature set: Table 11.

S: sample-elicited features (S + PW). A feature set is elicited from the same model after it has seen data. In a single call, GPT-5.4 (HIGH REASONING) reads 752 pairs ($1{,}504$ replies, about 860k tokens of input), each presented as an OP view followed by two replies written independently to it, and is asked to name the dimensions along which the replies differ as attempts to change the poster's view, giving for each a short evidence phrase, a name and a one-sentence definition (Prompt 16). Features must be properties of a reply rather than of the poster's view, so that they can vary within a matched pair. The number of features is left to the model: with the same open-count instruction but no data, it names about 16 (see P); with data, it names 42. The exchanges keep one pair per distinct original poster, cover the range of reply lengths, and are drawn from both the training and the held-out pairs, without their labels. Feature set: Table 12.

D + BT and P + BT: pairwise Bradley–Terry scoring. Each feature is scored by its own Bradley–Terry model on a frozen comparison graph (Prompt 14). The 800 training exchanges form a 16-regular anchor graph (6,400 comparisons); each of the 1,600 test exchanges is then compared against 16 anchors (25,600 comparisons), for 32,000 judge calls. Every call presents two exchanges with the full feature list and forces a winner on *each* feature, with a short evidence phrase. Anchor scores are fitted per feature by plain Bradley–Terry with a position-bias term β (ridge 10^{-3}); each test exchange is then scored by a one-dimensional convex solve against the *frozen* anchor scores, so test items never influence each other. P + BT uses a graph verified identical to D + BT's, so the two arms differ only in the features named in the prompt.

D + PW, D_r + PW and P + PW: pointwise scoring. The same features are instead rated absolutely (Prompt 17): one call per exchange rates it 0–10 on every feature so the feature set costs 2,400 calls rather than 32,000. The prompt is a deliberate near-copy of the pairwise one; two differences are unavoidable and consequential. A scale must be defined, anchored on the population of replies, because ratings must be comparable across exchanges rated in separate calls — pairwise never needs this, which is precisely the property this arm gives up; and the prompt must push against central tendency, since without it ratings pile up at 5–7 and the arm loses to compression rather than to anything substantive. Because a pointwise score is a single draw, each feature set was additionally rated a *second* time with the features reshuffled into different order, which supplies the independent replicate the noise correction of Figure 2 requires; the same replicate exists for the BT arms for free, by refitting on disjoint halves of the 32,000 comparisons. The accuracy reported in Table 2 is the one attained by the average of these two scores.

D.3 Additional results

D.3.1 Additional measures

This subsection gathers additional results, notably accuracy measures (Table 9) and dimensionality measures (Figure 20).

Other runs. Table 9 shows results about other runs that were made. Methods that specify "full" were fitted on the whole training set. **BOW** and **POS** are two surface baselines of Tan et al. (2016), built identically except for what a reply is reduced to: BOW represents each reply by its bag of lowercased word tokens (vocabulary fitted on the training replies, tokens occurring more than five times kept), while POS discards the words entirely and represents the reply by the sequence of its fine-grained Penn part-of-speech tags as assigned by spaCy, so that only syntactic style remains (42 tags).

As we can see, D + PW and D_r + PW rank among the best methods despite their smaller training and feature set size, beaten only by the opaque embeddings trained on the whole training set and by Tan et al. (2016)'s features, while being interpretable. Additionally, all four panels of Figure 19 show that the reliable discovered features are more predictive, at matched feature count, than either the full discovered set or the sample-elicited set: the reliability filter of the statistical model selects the features that carry the signal.

Measuring effective dimensionality. We ask how many independent directions of variation a feature set resolves, using the participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the eigenvalues λ of the feature set's score correlation matrix over the 1,600 held-out exchanges. $\mathrm{PR} \in [1, k]$, equalling 1 when all k features are one construct and k when they are mutually independent.

Three choices make the comparison meaningful. **a) Rank correlations.** Bradley–Terry scores live on an arbitrary logit scale and pointwise ratings on a coarse 0–10 one, and a covariance-based participation ratio weights a feature by its variance; we therefore use the Spearman correlation matrix, which is invariant to scale and to any monotone reparameterization.

b) Matched feature set size. A feature set with more features trivially has more room for independent variation, so for each k we average the participation ratio over k-subsets of the feature set — enumerated exactly while $\binom{K}{k} \leq 3{,}000$ and sampled (3,000 draws) otherwise — and read comparisons at matched k.

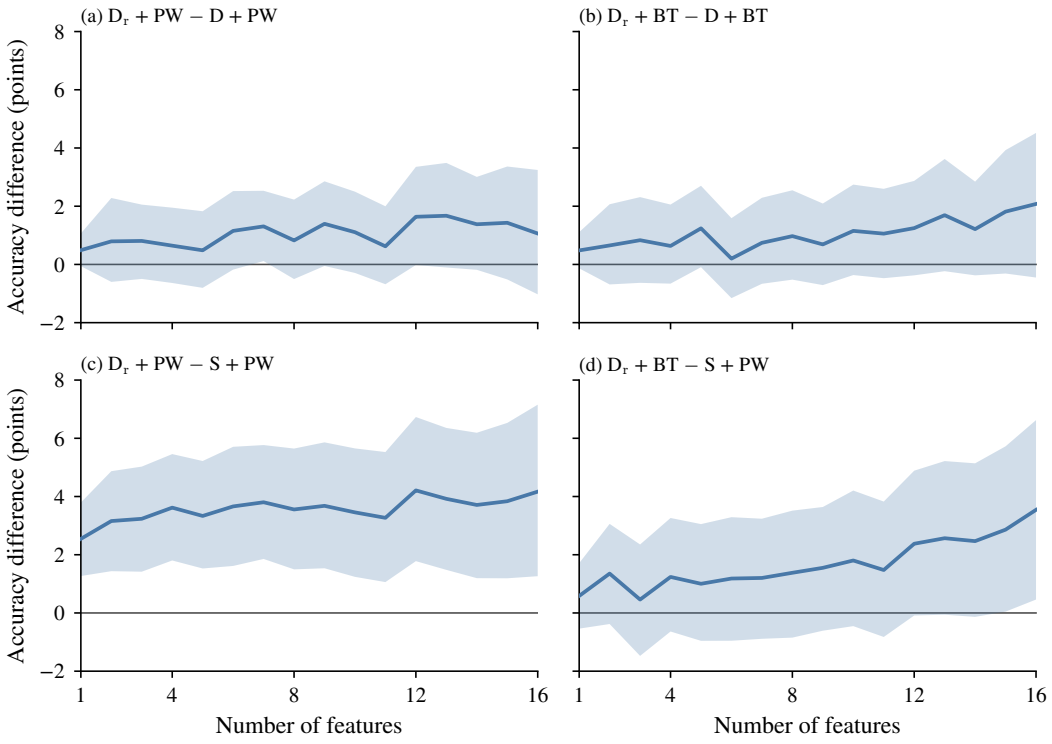

Figure 19: **The reliable discovered features are more predictive, feature for feature, than the full discovered set and than the sample-elicited set.** Each panel shows, as a function of the number of features k, the difference in held-out accuracy between D_r (the 16 reliable discovered features) and another feature set. At each k, an ℓ_1-regularized logistic regression is fitted on up to 50 random k-subsets of each set (all subsets when there are fewer), and accuracy is averaged over subsets; PW scores are averaged over two scoring runs. Shaded: 95% bootstrap interval at each k (4,000 resamples of the 800 held-out pairs and of the drawn subsets). (a, b) Against the full discovered set D, with the same scoring: D_r is ahead at every k, by 1.0 point on average over k under both scorings (95% CIs $[-0.1, 2.2]$ for PW, $[0.0, 2.1]$ for BT). (c, d) Against the sample-elicited set S, scored pointwise only: D_r + PW is ahead at every k, by 3.6 points on average ($[1.8, 5.4]$); D_r + BT's advantage grows with k, reaching 3.5 points at $k = 16$ ($[0.5, 6.6]$).

c) Measurement noise correction. The participation ratio counts independent directions and cannot distinguish signal from independent measurement error, so a noisier scorer is credited with extra dimensions; since a Bradley–Terry score aggregates 16 comparisons while a pointwise score is a single coarse draw, the uncorrected comparison is biased in favor of pointwise. We therefore measure each feature set twice, independently: for the pairwise arms by refitting on disjoint halves of the 32,000 comparisons (the anchor graph split in two, and each test exchange's 16 anchor comparisons split 8/8, so the halves share no data), and for the pointwise arms by a second scoring run with the features reshuffled into different bundles and order.

Writing $C_{ij} = \mathrm{corr}(A_i, B_j)$ for the cross-measurement correlation, the independent errors cancel in expectation, so that C_{ii} estimates feature i's reliability and $R_{ij} = S_{ij}/\sqrt{S_{ii}S_{jj}}$ with $S = (C + C^\top)/2$ disattenuates the correlation between the latent features. The participation ratio is then computed on R. The correction is substantial and asymmetric for the full feature sets: it removes 15–20% of the raw participation ratio for the pairwise arms and 25–28% for the pointwise ones; for the reliable subset D_r it is smaller and nearly symmetric (11% under BT, 13% under PW). Reliabilities are high enough for the division to be safe (median 0.83 and 0.77 for the D and P pointwise score sets; 0.76 and 0.71 for the half-data Bradley–Terry fits, i.e. 0.86 and 0.83 for the full score sets by the Spearman–Brown correction), and no disattenuated matrix carried any negative eigenvalue

Arm	Dims	Accuracy	95% CI
Emb. (full)	3,072	64.88%	[61.50, 68.00]
Tan	3,205	63.75%	[60.38, 67.00]
D + PW	29	62.63%	[59.38, 66.00]
D_r + PW	16	62.13%	[58.88, 65.50]
Emb.	3,072	61.63%	[58.25, 65.00]
D_r + BT	16	61.50%	[58.13, 64.88]
D + BT	29	61.38%	[58.00, 64.75]
BOW (full)	11,348	61.25%	[57.88, 64.63]
P + PW	16	61.25%	[58.00, 64.63]
P + BT	16	61.13%	[57.75, 64.38]
POS (full)	42	60.50%	[57.00, 63.75]
S + PW	42	59.88%	[56.50, 63.25]
#words	1	59.25%	[55.88, 62.75]
Zero-shot	–	56.94%	[54.19, 59.88]

Table 9: **Held-out accuracy of each method.** D_r: the reliable subset of the discovered features D. Emb. denotes TEXT-EMBEDDING-3-LARGE embeddings. Arms marked (full) are trained on all 3,411 training pairs; all others on 400. Accuracies for PW arms are obtained from the average of two scoring runs. Zero-shot asks the judge directly to predict the Δ-winning reply, with the query issued in both presentation orders and the two predictions averaged. Chance $= 50\%$; 95% bootstrap percentile intervals (4,000 resamples of the 800 held-out pairs, shared across arms).

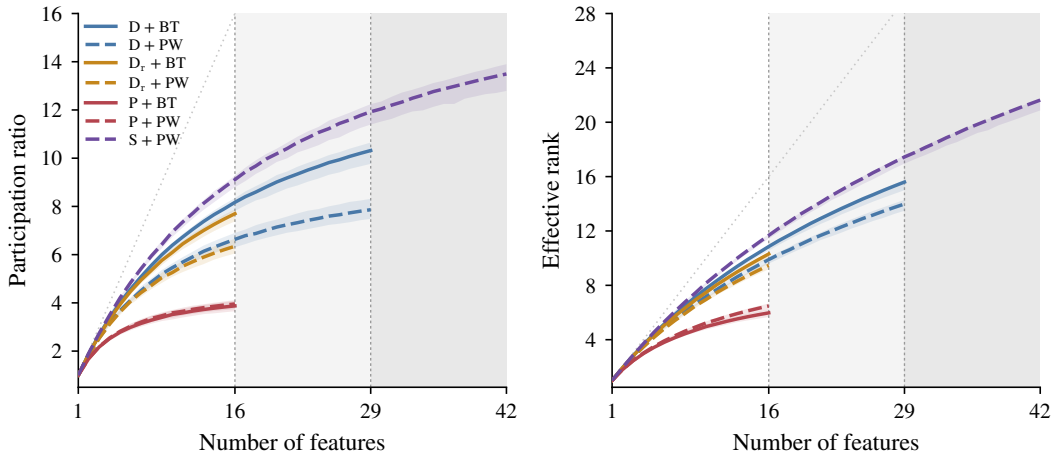

Figure 20: **Both dimensionality measures agree on the ordering of the feature sets.** Noise-corrected effective dimensionality against feature set size k, averaged over k-subsets of each feature set. *Left:* participation ratio, as in Figure 2. *Right:* effective rank (Roy and Vetterli, 2007), $\exp H(p)$ with $p_i = \sigma_i / \sum_j \sigma_j$ the normalized singular values. Both are computed on the disattenuated correlation matrix. The order holds under both measures. Bands, shading and dotted line as in Figure 2.

mass. Bands are 95% percentile bootstrap over the 1,600 exchanges, with the cross-measurement correlation rebuilt inside each of 200 replicates.

D.3.2 Discovered Features and Winning Argumentation Techniques

This subsection is devoted to additional results on the discovered features: Figure 21 completes Figure 3, while Table 10 lists each discovered feature with its description, together with its effectiveness.

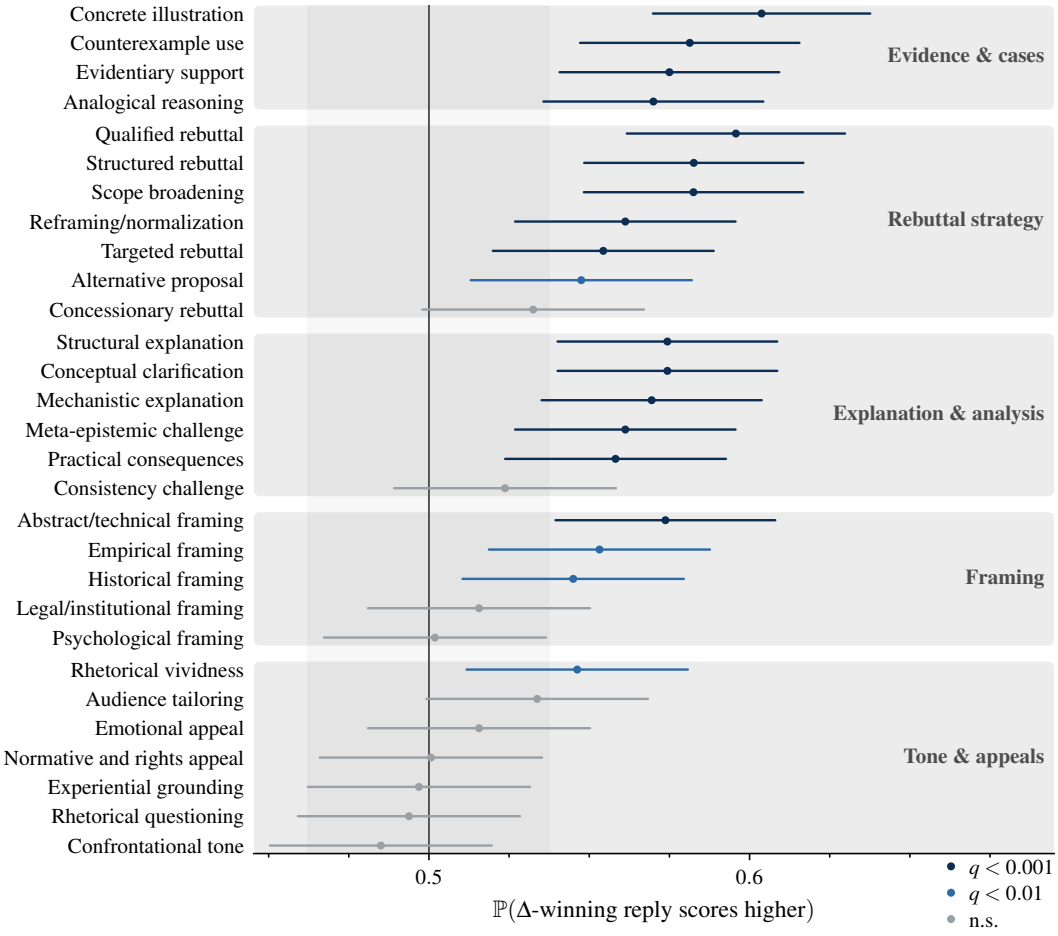

Figure 21: **Effectiveness of the discovered argumentation techniques (D + PW).** The figure is the complete version of Figure 3 and shows each discovered feature's *effectiveness*, i.e. the fraction of the test pairs in which the reply that earned the Δ scores higher on that feature. 0.5 is chance. Brackets give 95% Wald intervals and q is the Benjamini–Hochberg false-discovery rate across the 29 tests. The discovered features are gathered by category. Exact values of the BH q and each feature's description can be found in Table 10.

Feature	Description	Win rate (95% CI)	q
Concrete illustration*	The argument makes its case through specific, vivid, or worked-through examples rather than abstraction alone.	0.604 [0.570, 0.638]	<0.001
Qualified rebuttal*	The challenger narrows the claim, marks exceptions, or emphasizes contextual dependence instead of arguing absolutely.	0.596 [0.562, 0.630]	<0.001
Structured rebuttal development*	The challenger builds an organized, multi-step, or multi-pronged case rather than a single bare point.	0.583 [0.548, 0.617]	<0.001
Scope broadening	The exchange expands or contracts the discussion by moving between narrow cases and wider principles or issue areas.	0.583 [0.548, 0.617]	<0.001
Counterexample use*	The challenger undermines the OP with exceptions, alternative cases, or edge-case scenarios that break the generalization.	0.581 [0.547, 0.616]	<0.001
Challenger evidentiary support*	The challenger backs the rebuttal with external evidence, citations, factual corrections, or expert authority.	0.575 [0.541, 0.609]	<0.001
Structural explanation	The challenger explains the issue through broader social, cultural, or systemic forces rather than isolated choices.	0.574 [0.540, 0.609]	<0.001
Conceptual clarification*	The challenger disputes definitions, categories, meanings, or criteria in order to reframe the disagreement.	0.574 [0.540, 0.609]	<0.001
Abstract or technical framing*	The exchange reasons at a more theoretical, philosophical, scientific, or specialized level rather than in everyday concrete terms.	0.574 [0.539, 0.608]	<0.001
Analogical reasoning*	The argument persuades by comparing the case to parallel real or hypothetical cases, including extended analogies and thought experiments.	0.570 [0.536, 0.604]	<0.001
Mechanistic explanation*	The challenger explains or critiques the causal mechanism by which the disputed outcome would arise.	0.569 [0.535, 0.604]	<0.001
Meta-level epistemic challenge	The challenger contests the framing, method, evidence standards, or burden of proof rather than only the first-order claim.	0.561 [0.527, 0.596]	<0.001
Reframing and normalization	The challenger tries to shift the evaluative lens by redescribing the issue, normalizing it, or offering an alternative way to view it.	0.561 [0.527, 0.596]	<0.001
Practical consequences and feasibility*	The challenger evaluates the view in terms of real-world effects, costs, risks, tradeoffs, or implementability.	0.558 [0.524, 0.593]	<0.001
Targeted rebuttal	The challenger engages the OP's specific claims, premises, or subpoints directly and often one by one.	0.554 [0.520, 0.589]	<0.001
Exchange empirical framing*	The exchange is framed in empirical or quantitative terms, emphasizing measurement, data, or evidentiary specificity.	0.553 [0.519, 0.588]	0.002
Alternative proposal and guidance*	The challenger offers concrete substitute policies, remedies, or practical steps instead of only criticizing.	0.547 [0.513, 0.582]	0.006
Rhetorical vividness	The argument uses especially vivid phrasing, storytelling, or memorable imagery to make the point stick.	0.546 [0.512, 0.581]	0.007
Historical framing and legacy	The argument appeals to historical context, precedent, innovation, or long-run cultural legacy.	0.545 [0.510, 0.580]	0.003
Audience tailoring	The challenger tailors the argument to particular subgroups, audiences, or the OP's specific situation.	0.534 [0.499, 0.568]	0.0502
Concessionary rebuttal	The challenger softens resistance by explicitly granting part of the OP's point before disagreeing.	0.532 [0.498, 0.567]	0.070
Consistency challenge*	The reply pressures the OP by showing that their own standard, principle, or rule yields inconsistency or self-undermining conclusions.	0.524 [0.489, 0.558]	0.195
Legal and institutional framing	The challenger grounds the rebuttal in law, institutions, official roles, or procedural distinctions.	0.516 [0.481, 0.550]	0.369
Emotional appeal and empathy	The reply persuades through feelings, empathy, reassurance, or existential/personal significance.	0.516 [0.481, 0.550]	0.412
Psychological framing and diagnosis	The challenger interprets views or behavior in terms of motives, traits, psychology, or developmental state.	0.502 [0.467, 0.537]	0.966
Normative and rights appeal*	The challenger argues from moral principle, fairness, rights, duty, or harm rather than only from descriptive claims.	0.501 [0.466, 0.535]	1.000
Challenger experiential grounding*	The challenger or exchange relies on firsthand experience, self-disclosure, or identity-based testimony as support.	0.497 [0.462, 0.532]	0.907
Rhetorical questioning	The challenger presses the point through pointed questions or requests for clarification rather than only direct assertion.	0.494 [0.459, 0.528]	0.773
Confrontational tone*	The reply relies on combative, loaded, mocking, or emotionally aggressive rhetoric.	0.485 [0.450, 0.520]	0.427

Table 10: **Which argumentation techniques separate winners from losers (D + PW).** This table details the results shown in Figures 3 and 21. It presents each discovered feature (its name and description) together with its *effectiveness*, i.e. the fraction of the test pairs in which the reply that earned the Δ scores higher on that feature. 0.5 is chance. Brackets give 95% Wald intervals and q is the Benjamini–Hochberg false-discovery rate across the 29 tests; features in bold satisfy $q < 0.05$. 19 of the 29 features separate winners, all of them positively. Features marked * form the reliable subset $\mathrm{D_r}$ (16 features, $\kappa \leq 50$ and $\rho \geq 0.75$).

Feature	Description
Core Claim Relevance	It measures the degree to which the reply focuses on the central claim under dispute rather than drifting to side issues.
Reason Engagement	It measures the degree to which the reply directly addresses the original poster's stated reasons, premises, or assumptions.
View Understanding	It measures the degree to which the reply accurately represents the other person's position before challenging it.
Logical Coherence	It measures the degree to which the reply's conclusions follow clearly and consistently from its premises.
Evidence Strength	It measures the degree to which the reply supports its claims with substantial, pertinent evidence or examples.
Source Trustworthiness	It measures the degree to which the reply relies on credible, authoritative, and verifiable sources when offering support.
Concrete Specificity	It measures the degree to which the reply uses precise details or examples instead of vague generalities.
Counterargument Handling	It measures the degree to which the reply anticipates and answers plausible objections to its own case.
Nuance and Balance	It measures the degree to which the reply acknowledges complexity, uncertainty, or limits rather than overstating its case.
Novel Insight	It measures the degree to which the reply offers a fresh framing, distinction, or information the reader is unlikely to have already considered.
Value Alignment	It measures the degree to which the reply connects its argument to the other person's stated values, goals, or concerns.
Consequence Salience	It measures the degree to which the reply makes the practical or moral implications of the issue vivid and important.
Emotional Resonance	It measures the degree to which the reply evokes fitting emotions that make its point more compelling.
Respectful Tone	It measures the degree to which the reply treats the other person charitably and without dismissiveness or contempt.
Clarity of Expression	It measures the degree to which the reply is easy to follow because its wording and structure are clear.
Economy of Argument	It measures the degree to which the reply makes its case efficiently without unnecessary length or repetition.

Table 11: **Prior-elicited features (P).** The 16 features named by GPT-5.4 (HIGH REASONING) for assessing the persuasiveness of a reply, with no data shown (Prompt 15), in the order produced.

Feature	Description
Topical responsiveness	How directly a reply addresses the poster's actual view rather than pivoting to meta-issues or adjacent topics.
Premise targeting	Whether a reply attacks the core assumption or inference supporting the view instead of focusing on side details.
Clarification demand	How much a reply advances the discussion by asking the poster to specify vague claims, terms, or examples.
Conceptual reframing	Whether a reply changes the frame of the debate by redefining the key concept or explaining that the dispute is really about something else.
Thought experiment	Whether a reply uses an invented scenario to test the implications or consistency of the poster's view.
Analogy use	How much a reply explains its point by mapping the issue onto a different familiar domain.
Consistency test	Whether a reply probes the poster's logic by applying it to parallel cases and checking whether the result is acceptable.
Reductio extension	Whether a reply undermines a view by extending it to an absurd or unacceptable conclusion.
Counterexample case	Whether a reply primarily challenges a general claim by presenting a concrete case that does not fit it.
Firsthand testimony	How much a reply relies on the writer's own experiences or preferences as evidence or illustration.
Stakeholder perspective	Whether a reply argues from direct membership in the affected group or a role that gives special practical exposure to the issue.
Empirical citation	How much a reply supports its claims with studies, data, or directly linked factual evidence.
Expert consensus appeal	Whether a reply leans on the agreement of qualified experts as a reason to update the view.
Legal mechanics	How much a reply explains the governing law, procedure, or institutional rules behind the issue.
Implementation realism	Whether a reply evaluates how workable a proposal would be once logistics, administration, and real-world constraints are considered.
Economic reasoning	How much a reply argues in terms of prices, incentives, resources, employment, or market effects.
Quantified argument	Whether a reply strengthens its case by attaching numerical magnitude or rates to the claims it makes.
Causal mechanism	Whether a reply explains the process by which one thing is supposed to produce another, rather than merely asserting a correlation or outcome.
Function reframing	Whether a reply challenges the view by arguing that the institution or behavior serves a different purpose than the poster assumes.
Use-case focus	Whether a reply argues from the specific practical situations in which the disputed thing works especially well or badly.
Institutional purpose	How much a reply justifies or criticizes something by appealing to the purpose of the larger system it belongs to.
Common-ground signaling	Whether a reply explicitly agrees with part of the poster's position before presenting disagreement.
Empathetic tone	Whether a reply validates the poster's feelings or perspective in a supportive way while disagreeing.
Motive attack	Whether a reply questions the poster's motives, sincerity, or openness instead of mainly addressing the substance of the claim.
Confrontational tone	How much a reply uses insult, scorn, or aggressive language toward the poster or target.
Humor deployment	Whether a reply uses jokes, playful phrasing, or irony as part of its persuasive approach.
Declared stance framing	Whether a reply explicitly frames itself as devil's advocate, partial agreement, or some other interpretive stance before arguing.
Epistemic caution	How much a reply qualifies its claims, notes uncertainty, or avoids overstating what it can prove.
Conditional reasoning	Whether a reply limits its claim to specific conditions or exceptions instead of treating the issue as all-or-nothing.
Abstraction level	Whether a reply operates mainly at a high-level philosophical or theoretical plane rather than a concrete one.
Explanatory correction	Whether a reply mainly persuades by supplying overlooked background information or correcting a misunderstanding.
Personal remedy suggestion	Whether a reply addresses the issue by recommending a practical step the poster could take in their own life.
Audience reframing	Whether a reply changes the evaluation of the thing in question by showing it was made for a different audience or goal than the poster assumes.
Scope challenge	Whether a reply argues that the poster's principle is too narrow to handle the full range of relevant cases.
Temporal horizon	Whether a reply evaluates the claim by emphasizing long-term effects rather than immediate ones, or vice versa.
Norm distinction	Whether a reply distinguishes between what is logically valid, socially useful, morally justified, or legally allowed rather than treating them as the same question.
Tradition appeal	Whether a reply supports or critiques a position by invoking established custom, continuity, or inherited social practice.
Privacy sensitivity	Whether a reply foregrounds the need to protect personal vulnerability, anonymity, or intimate information when evaluating the issue.
Collective agency framing	Whether a reply explains outcomes as the aggregate result of many individuals' choices rather than the intent of a single actor.
Resource referral	Whether a reply mainly tries to persuade by handing off a longer external resource instead of making the argument in the reply itself.
Form substance split	Whether a reply distinguishes between objection to labels, rhetoric, or presentation and objection to the underlying idea itself.
Meta-argumentation	Whether a reply persuades by analyzing the quality or structure of the argument itself rather than primarily debating the topic.

Table 12: **Sample-elicited features (S).** The 42 features named by GPT-5.4 (HIGH REASONING) after reading a sample of 752 exchange pairs, without their outcomes.

D.3.3 PROMPTS

Prompt 12: Comparative method feature-discovery, CMV corpus.

```
You are helping build an interpretable framework for what drives persuasion in debate, by
examining how two attempts to change someone's mind differ. The framework works by
collecting many pairwise comparisons: for each pair, the ways the two exchanges differ are
articulated in plain terms, and aggregated across many pairs these reveal the distinct
dimensions along which persuasion-relevant material varies. There is no fixed rubric — the
dimensions are recovered from comparisons like this one, not specified in advance.
Background. Each item is an online discussion in which an original poster (the "OP") states
a view they currently hold and invites others to change it; a challenger then replies with
an argument aimed at changing that specific view. You will be shown two such exchanges,
Exchange A and Exchange B, each presented as a block that begins with the OP's view (marked
[OP VIEW]) followed by the challenger's reply (marked [CHALLENGE]). Crucially, the two
exchanges concern different posters holding different views on different topics.

Your task. Identify the dimensions on which the two exchanges meaningfully differ in ways
that could bear on whether the OP's view gets changed. Consider the whole exchange: both
the challenger's argument and the OP's post itself. Report the differences that genuinely
stand out to you on reading this pair — bring up whatever you find salient, and do not work
from a predetermined list. Favor real, noticeable differences over minor ones, but do not
force the list to be short: report as many as actually distinguish these two exchanges.
For each dimension, first name the concrete, specific feature you actually saw — a short
phrase pointing to the specific evidence. Then, grounded in that evidence, name the
dimension, describe in one sentence what it captures, and indicate which exchange shows
more of it.
Judge each dimension on its own terms, from its own evidence. Exchanges have uneven
profiles: it is entirely normal for one to be stronger on some dimensions and the other
stronger on others, so let the exchange that shows more of each dimension fall wherever the
evidence for that dimension points.
Out of scope — do not treat these as dimensions of difference:
- the topic or subject matter either exchange is about, and how interesting, important,
difficult, or agreeable either OP's view is: the two exchanges concern different topics by
design, so topic or subject-matter difference is never a dimension here;
- which side you personally agree with, or whether you think either OP's view is correct;
- the username or any other identifier of either poster or challenger;
- length in itself
How to report. List each dimension as exactly one item — do not combine dimensions with
"and", "/", "&", or a comma; split related dimensions into separate items. For the evidence
field, give a short phrase naming the concrete feature you saw. For the dimension field,
give a concise general noun phrase naming the dimension (the kind of short label you might
use as a rubric column). For the description field, give one sentence saying what that
dimension captures. Indicate which exchange shows more of it.
Output only JSON of the form: {"differences": [{"evidence": "<short phrase>", "dimension":
"<concise noun phrase>", "description": "<one sentence>", "winner": "A" | "B"}, ...]}

Exchange A: {arg_a}
Exchange B: {arg_b}
Following the standard above, list the dimensions on which these two exchanges differ in
ways bearing on persuasion — whether in the challenger's argument or in the OP's post — for
each giving the concrete feature you saw, the dimension, a one-sentence description, and
which exchange shows more of it. Compare on persuasion-relevant qualities, not topic.
Respond with the JSON object only.
```

Prompt 13: Comparative method, CMV corpus – feature canonicalization prompt.

```
You are organizing a flat list of evaluation dimensions into a set of canonical features.

These dimensions were produced by judges comparing pairs of online exchanges — in each, an
original poster (the "OP") states a view and a challenger replies with an argument aimed at
changing it. On each comparison a judge named the dimensions on which two exchanges
```

differed in ways bearing on whether the OP's view gets changed; these span both the challenger's argument and the OP's own post. Aggregated across many comparisons, the same underlying construct was named in many slightly different ways. Your job is to group the different phrasings of the SAME underlying construct together.

Below is the full list. Each entry has an INDEX in square brackets, a short NAME, a one-sentence DESCRIPTION of what it captures, and CITATIONS (how many times it was named — for your awareness only).

⟨*indexed dimension list*⟩

Instructions:
- Group entries that denote the SAME underlying construct, judging by MEANING, not by surface wording. Different words for the same idea belong in one group; the same word used for different ideas belongs in different groups.
- A dimension describing the CHALLENGER's argument and a dimension describing the OP's POST are DIFFERENT constructs even when they share a word — group by the specific feature and whose it is, keeping OP-side and argument-side constructs in separate groups.
- Do NOT force everything into a small number of groups, but do NOT split one construct into its facets either. A dimension that does not clearly share a construct with any other — including a framing or appeal specific to one debate's subject matter that does not recur as a general persuasion feature — does NOT become its own feature: place all such genuinely-apart, idiosyncratic dimensions together in a single catch-all group whose canonical name is exactly "Other". Do not create many single-member groups, and do not absorb a genuinely-apart dimension into a loosely-related group to tidy up — send it to "Other".
- Do NOT merge two constructs just because they often co-occur or because persuasive exchanges tend to show both. Group only by shared MEANING, never by how exchanges might score.
- Citations are context only. Do not merge a frequently-named dimension into a rarely-named one or vice versa on the basis of counts; counts do not determine grouping.

For each group, give:
- a canonical NAME (a concise noun phrase suitable as a rubric column; use exactly "Other" for the catch-all group),
- a one-sentence DEFINITION of the construct the group captures,
- the list of members, each as its INDEX and its NAME copied verbatim from that index.

Rules on membership (these will be checked programmatically):
- Every input index must appear in exactly one group, exactly once. Place all of them; drop none; duplicate none.
- Do not invent indices that are not in the input.
- Each member's "name" must be the exact input NAME at that index — this is a cross-check, so copy it from the bracketed entry, do not paraphrase.

Output only JSON:
{"features": [{"name": "<canonical noun phrase>", "definition": "<one sentence>", "members": [{"index": <int>, "name": "<verbatim input name at that index>"}, ...]}, ...]}

Prompt 14: BT feature scoring, CMV corpus.

You are helping apply an interpretable framework for what drives persuasion in debate. The framework scores how two attempts to change someone's mind differ along a fixed set of dimensions, listed below.
Background. Each item is an online discussion in which an original poster (the "OP") states a view they hold and invites others to change it; a challenger then replies with an argument aimed at changing that specific view. You will be shown two such exchanges, Exchange A and Exchange B, each presented as a block that begins with the OP's view (marked [OP VIEW]) followed by the challenger's reply (marked [CHALLENGE]). The two exchanges concern different posters holding different views on different topics.
The dimensions:
[[features]]

```
Your task. For EACH listed dimension, in order, give a short evidence phrase (the concrete
feature you saw) and then indicate which exchange shows more of it — 'A' or 'B'. Some
dimensions describe the challenger's argument, others describe the OP's post — assess each
exactly as its definition states. Judge each dimension on its own terms: exchanges have
uneven profiles, so it is entirely normal for A to show more of some dimensions and B more
of others. Even when the two are close on a dimension, pick the exchange that shows more of
it.
Out of scope — do not let these drive any judgment: the topic or subject matter and how
interesting, important, difficult, or agreeable either OP's view is (the two differ by
design); which side you personally agree with, or whether an OP's view is correct;
usernames or identifiers; length in itself.
Return one item per listed dimension, in the same order, naming the dimension exactly as
given, each with a short evidence phrase and the winner ('A' or 'B').
Output only JSON of the form: {"features": [{"name": "<dimension>", "evidence": "<short
phrase>", "winner": "A" | "B"}, ...]}

Exchange A: {exchange_a}
Exchange B: {exchange_b}
Following the standard above, go through each listed dimension in order: give the concrete
evidence you saw and then say which exchange shows more of it. Compare on the listed
dimensions, not on topic. Respond with the JSON object only.
```

Prompt 15: Prior feature elicitation, CMV corpus.

```
You are helping build a framework for assessing how persuasive an argument is.
In online debates, a person states a view they hold and invites others to change it. Others
reply with arguments aimed at changing that view.
Your task. Name exactly 16 dimensions for assessing how persuasive such a reply is.
Each dimension will be used as a measurement scale: a judge will be shown two replies and
asked which of the two shows more of it. So each dimension must be a matter of degree
rather than a yes/no fact, and the 16 must be distinct from one another.
For each dimension give:
- "name": a short label, 2-5 words;
- "definition": one sentence saying what it measures.
Output only JSON of the form:
{"features": [{"name": "<short label>", "definition": "<one sentence>"}, ...]}

Name the 16 dimensions now, as a single JSON object. Respond with the JSON only.
```

Prompt 16: Sample-based feature elicitation, CMV corpus.

```
You design an interpretable scoring rubric and output only JSON.

You are helping build an interpretable scoring framework by examining how attempts to
change someone's mind differ from one another. There is no fixed set of dimensions — they
are recovered from what actually distinguishes these exchanges, not specified in advance.
In online debates, a person states a view they hold and invites others to change it. Others
reply with arguments aimed at changing that view.
Below is a sample of such exchanges. Each begins with the poster's view (marked [OP VIEW]),
followed by two separate replies to it, written independently by different people (marked
[REPLY A] and [REPLY B]). The two replies answer the same view; they are not responses to
each other, and their order carries no meaning.
=== EXCHANGES ===
⟨sample of exchanges⟩
=== END EXCHANGES ===
Reading across them, attend to the ways the replies differ from one another — the axes
along which you can see one reply showing more, less, or something different from another.
Identify the dimensions on which these replies meaningfully differ as attempts to change
the poster's view. Report whatever genuinely stands out to you; do not work from a
predetermined list, and do not force the list to be short — report as many dimensions as
```

```
actually distinguish these replies. Do not pad the list either: add a dimension only when
it names something none of the others already capture.
A dimension may be something a reply shows more or less of along a range, or something a
reply either shows or does not.
Every dimension must be a property of a reply, not of the poster's view or of the poster:
name only dimensions on which the two replies to the same view can differ.
Keep genuinely distinct dimensions separate even when they are related or tend to co-occur —
 do not merge two dimensions because strong replies tend to show both. List each dimension
as exactly one item — do not combine dimensions with "and", "/", "&", or a comma; split
related dimensions into separate items.
For each dimension, first name the concrete, specific feature you actually saw in the
replies — a short phrase pointing to the evidence. Then, grounded in that evidence, name
the dimension and define in one sentence what it measures. Write the definition so a rater
can apply it to a new reply without having seen this sample: describe the characteristic
itself, and never refer to specific exchanges or to the reply labels.
Out of scope — do not treat these as dimensions of difference:
- the topic or subject matter, and how interesting, important, difficult or agreeable the
poster's view is;
- which side you personally agree with, or whether the poster's view is correct;
- usernames or identifiers;
- length in itself.
Output only JSON:
{"features": [{"evidence": "<short phrase>", "name": "<short label, 2-5 words>",
"definition": "<one sentence>"}, ...]}
```

Prompt 17: PW feature scoring, CMV corpus.

```
You are helping apply an interpretable framework for what drives persuasion in debate. The
framework scores an attempt to change someone's mind along a fixed set of dimensions,
listed below.
Background. Each item is an online discussion in which an original poster (the "OP") states
a view they hold and invites others to change it; a challenger then replies with an
argument aimed at changing that specific view. You will be shown one such exchange,
presented as a block that begins with the OP's view (marked [OP VIEW]) followed by the
challenger's reply (marked [CHALLENGE]).
The dimensions:
[[features]]
Your task. For EACH listed dimension, in order, give a short evidence phrase (the concrete
feature you saw) and then rate the exchange on that dimension from 0 to 10. Some dimensions
describe the challenger's argument, others describe the OP's post — assess each exactly as
its definition states.
What the numbers mean. Rate each dimension relative to online arguments of this kind in
general, not relative to this exchange's other dimensions: 0 means the dimension is
entirely absent, 5 means it is present to a typical degree, and 10 means it is present
about as strongly as such a reply ever gets. The same rating must mean the same thing on
every exchange you score, because these scores are compared across exchanges you will not
see together.
Judge each dimension on its own terms: exchanges have uneven profiles, so it is entirely
normal for one to score high on some dimensions and low on others. Use the full range — do
not cluster every dimension in the middle.
Out of scope — do not let these drive any rating: the topic or subject matter and how
interesting, important, difficult, or agreeable the OP's view is; which side you personally
agree with, or whether the OP's view is correct; usernames or identifiers; length in itself.
Return one item per listed dimension, in the same order, naming the dimension exactly as
given, each with a short evidence phrase and an integer rating.
Output only JSON of the form: {"features": [{"name": "<dimension>", "evidence": "<short
phrase>", "score": <integer 0-10>}, ...]}

Exchange: {exchange}
Following the standard above, go through each listed dimension in order: give the concrete
evidence you saw and then its rating. Respond with the JSON object only.
```