WHICH POST WILL BE MORE ENGAGING? DISCOVERING EVALUATION CRITERIA VIA LLM-BASED REFLECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the relative popularity of social media posts plays an important role in understanding the relationship between content and user interactions, while informing content creation, candidate selection, and evaluative feedback for generative models. LLM-based evaluation and feedback-driven prompt optimization support text assessment and iterative instruction refinement, yet historical popularity outcomes provide only indirect guidance for revising evaluation criteria. The combined influence of content, creators, and platform distribution makes individual prediction errors an ambiguous basis for revision, increasing the risk of learning example-specific associations rather than reusable criteria. We introduce Grounded Reflection for Automated Criteria Exploration (GRACE), a framework that leverages LLM-based reflection to discover evaluation criteria from historical engagement feedback. GRACE examines discrepancies between model judgments and observed relative popularity across examples to refine and select topic-specific criteria for predicting the relative popularity of new posts. Across six topic-specific datasets, GRACE achieves 60.43% selective directional accuracy at 84.49% directional coverage, improving on the initial task prompt by 8.46 percentage points and outperforming the general prompt optimization methods evaluated in this study. These results support using historical engagement feedback to improve text-based relative popularity prediction, providing a methodological foundation for evaluation tools designed to aid content creation and selection.

1 INTRODUCTION

Social media platforms support knowledge sharing, opinion expression, and content creation. Understanding how posts attract attention and elicit interaction can inform content creation and behavioral analysis (Zhang et al., 2023), while supporting content selection and feedback for generative models (Yu et al., 2024). However, absolute popularity is difficult to predict because observed responses reflect the combined influence of content, creator visibility, and posting conditions (Hessel et al., 2017). We therefore adopt a comparative formulation that predicts which post will be more popular, shifting the prediction target from absolute magnitudes to relative ordering (Tan et al., 2014; Hessel et al., 2017). Recorded user interactions, including likes, comments, and saves, constitute the *engagement records* used to operationalize *relative popularity*.

The ability of large language models (LLMs) to interpret and compare text makes them a natural starting point for such judgments. LLM-as-a-judge research has established the use of natural-language instructions to guide text assessment (Zheng et al., 2023; Liu et al., 2023), with explicit evaluation criteria making these judgments more structured and adjustable (Kim et al., 2024b;a). Complementary work on automatic prompt optimization treats instruction design as a feedback-driven search problem (Pryzant et al., 2023; Opsahl-Ong et al., 2024; Agrawal et al., 2026). Together, these approaches provide mechanisms for eliciting and improving judgments. For popularity prediction, however, a task-specific issue remains: how can the mixed signals in historical outcomes inform reusable criteria for evaluating posts? Discrepancies between model judgments and observed popularity provide evidence for examining the current criteria, but using this evidence presents two challenges. First, a prediction error does not directly reveal which textual criterion should be revised because the observed popularity also reflects the creator and distribution conditions. Second, a revi-

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.

sion motivated by particular training errors may fit those examples without improving predictions on new posts, a risk also observed in prompt optimization (Pryzant et al., 2023). The goal is therefore to derive criteria from recurring error patterns and assess their predictive value on examples beyond those used to formulate them.

To address these challenges, we introduce Grounded Reflection for Automated Criteria Exploration (GRACE), a framework that discovers evaluation criteria through LLM-based reflection on historical engagement feedback. GRACE treats discrepancies between model judgments and observed relative popularity as evidence for examining and revising the current rubric. Across multiple training cases, it examines which textual cues are overemphasized, overlooked, or applied under inappropriate conditions. These cross-case patterns guide local rubric revisions, which are evaluated on held-out validation examples to select criteria that improve predictive performance. The selected criteria guide comparisons of new posts, while agreement across repeated judgments determines whether the system returns a directional prediction. Figure 1 illustrates how historical engagement feedback informs the evaluation criteria.

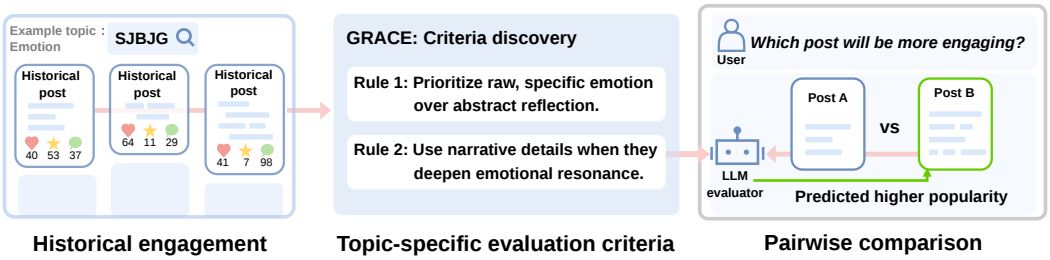


Figure 1: Discovering topic-specific evaluation criteria from historical engagement records for relative popularity prediction. The SJBJG rules are abbreviated for readability.

Across six topic-specific datasets, GRACE achieves 60.43% selective directional accuracy at 84.49% directional coverage, improving over the initial task prompt by 8.46 percentage points and outperforming general prompt optimization methods under the evaluated configurations. Selective accuracy improves across all six topics. Beyond predictions, GRACE produces explicit, editable evaluation criteria and records the revisions, allowing researchers to inspect how historical feedback informs the rubric. These findings support using historical engagement to improve relative popularity prediction, providing an evaluation basis for content selection and creation assistance, alongside inspectable records for examining the relationship between textual expression and observed popularity.

2 Related Work

Learning from social media engagement. Research on social media engagement investigates how content relates to user responses under different contextual conditions. Studies associate communication strategies with engagement (Zhang et al., 2023), while comparative designs and predictive models examine wording and title effects under controls for topic, authorship, or posting time (Tan et al., 2014; Weissburg et al., 2022). Relative popularity prediction similarly compares posts submitted to the same community in close succession (Hessel et al., 2017). A complementary line of work uses popularity feedback (Yu et al., 2024) and social-network context (Zhao et al., 2025) to guide content generation. GRACE uses historical engagement feedback to discover explicit, topic-specific evaluation criteria for relative popularity prediction.

Optimizing LLM evaluation criteria. LLM-based evaluation can be structured through criterion-conditioned judgments (Kim et al., 2024a) and interactive refinement of criteria and prompts (Kim et al., 2024b). Task-specific training provides another route to improving evaluator performance, as demonstrated in pairwise creativity assessment (Cao et al., 2026). Feedback-driven refinement operates at different levels: Self-Refine iteratively improves generated outputs (Madaan et al., 2023), whereas automatic prompt optimization searches over reusable instructions and demonstrations (Opsahl-Ong et al., 2024), using feedback such as critiques of training errors (Pryzant et al., 2023) and execution traces (Agrawal et al., 2026). GRACE builds on criterion-based evaluation and

feedback-driven prompt search, using discrepancies between model judgments and observed relative popularity to guide topic-specific rubric revisions.

Judgment reliability and selective prediction. Reliable evaluation requires understanding both judgment variability and the consequences of retaining only a subset of predictions. Position-bias studies examine changes induced by reordering candidates (Wang et al., 2024; Shi et al., 2025), while repeated-sampling analyses investigate reliability across model calls (Schroeder & Wood-Doughty, 2024). Selective prediction studies the trade-off between predictive error and coverage (Geifman & El-Yaniv, 2019). For LLM judges, Jung et al. (2025) develop a selective evaluation framework with guarantees of agreement with human judgments. GRACE uses agreement across repeated, order-balanced judgments to determine which relative-popularity predictions to retain and evaluates the resulting trade-off between accuracy and coverage against engagement-derived targets.

3 PRELIMINARIES

Relative popularity. We study the prediction of relative popularity between social media posts within the same topic. For topic k, each post i is represented by its observable text x_i, comprising its title, body text, and available tags, and its historical engagement records e_i. A predefined mapping g_k converts these records into a normalized score:

$$s_i = g_k(e_i) \in [0, 1]. \tag{1}$$

The score s_i represents the post's observed popularity. In this study, g_k aggregates engagement measures and converts the resulting scores into percentile ranks among the retained posts in topic k. The normalized score s_i expresses relative standing; a gap between two scores therefore represents rank separation rather than an absolute difference in interaction counts. The aggregation weights, ranking procedure, and tie threshold are specified in Appendix A.2.

Pairwise prediction. For two posts i and j from the same topic, the comparison target is defined by their score difference:

$$y_{ij} = \begin{cases} +1, & s_i - s_j > \tau, \\ -1, & s_j - s_i > \tau, \\ 0, & |s_i - s_j| \le \tau, \end{cases} \tag{2}$$

where $\tau \ge 0$ is a predefined tie threshold. The labels $+1$ and -1 indicate higher observed popularity for the first and second post, respectively, while 0 denotes a difference within the tie threshold. Reversing the presentation order reverses the directional target and preserves a tie, giving $y_{ji} = -y_{ij}$. This symmetry defines the relationship between the two presentations of the same comparison. Given a prompt p, the evaluator predicts $\widehat{y}_{ij}(p)$ from the text pair (x_i, x_j). The prompt specifies the task requirements and evaluation criteria, while historical engagement records supply the supervision targets.

Rubric search. Each topic has separate training, validation, and test partitions, with pairs constructed within each partition. With the language model parameters held fixed, search starts from an initial task prompt p_0 and uses training feedback to generate candidate evaluation criteria. Validation performance determines which prompt is selected. A candidate rubric specifies which textual cues should inform the comparison and under what conditions they are relevant. The same prompt is applied across pairs within a topic, so its validation score measures performance across examples rather than agreement with an individual training case.

Let $\mathcal{C}_k$ denote the candidate prompts evaluated during search for topic k, and let $\mathcal{V}_k^{\mathrm{dir}}$ contain the validation pairs with nonzero target labels. The directional validation accuracy of prompt p is

$$V_k(p) = \frac{1}{|\mathcal{V}_k^{\mathrm{dir}}|} \sum_{(i,j) \in \mathcal{V}_k^{\mathrm{dir}}} \mathbf{1}[\widehat{y}_{ij}(p) = y_{ij}], \tag{3}$$

where predictions are obtained under a fixed validation protocol. The selected topic-specific prompt is

$$p_k^* \in \arg\max_{p \in \mathcal{C}_k} V_k(p). \tag{4}$$

The selected prompt and inference settings are then frozen for test evaluation.

4 METHOD

GRACE consists of two stages: rubric search and inference (Figure 2). During search, the evaluator compares training pairs under the current prompt, and disagreements with the observed popularity targets provide feedback for revising the criteria. The reflector examines these errors across cases and proposes changes to the rubric. The prompt editor implements the proposed changes while preserving the task definition and input-output interface. Validation then determines whether the candidate improves predictive accuracy and should replace the current prompt for subsequent search.

During inference, the selected prompt and model settings remain fixed. The evaluator compares new pairs repeatedly under both presentation orders, with swapped-order predictions mapped back to the original orientation before aggregation. The system returns the aggregated prediction only when it favors one post over the other and its supporting vote fraction reaches the acceptance threshold; otherwise, it abstains.

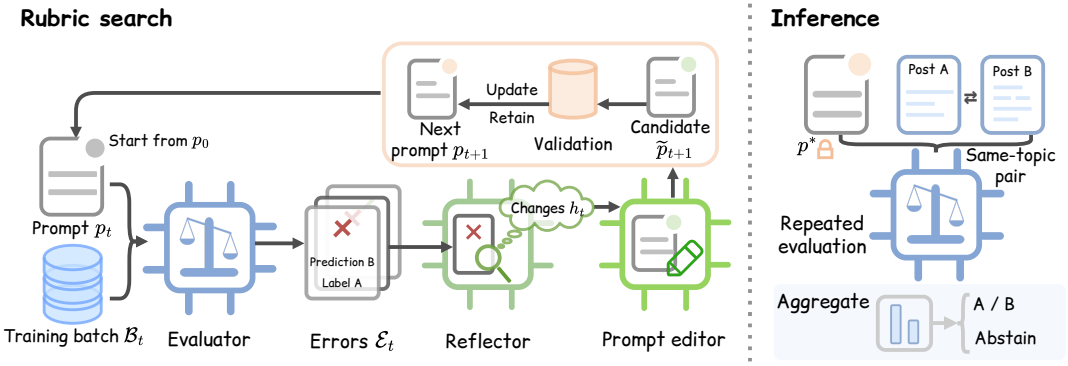


Figure 2: Overview of GRACE. During rubric search, the reflector proposes changes h_t from training errors, and the editor produces a candidate prompt. Validation then determines whether the candidate improves directional accuracy and should replace the current prompt for subsequent search. At inference, the fixed prompt p^* guides repeated judgments under both presentation orders. The system returns the aggregated prediction when it favors one post and its supporting vote fraction reaches γ; otherwise, it abstains.

4.1 FROM OUTCOME DISCREPANCIES TO RUBRIC REVISIONS

Collecting comparison errors. Incorrect comparisons provide concrete feedback for revising the current rubric: they identify pairs for which the model's judgment does not match observed relative popularity. Examining the paired texts alongside the prediction and target gives the reflector a basis for reconsidering the criteria used in those comparisons. GRACE therefore uses prediction errors as the starting point for rubric revision.

At search iteration t, the evaluator applies the current prompt p_t to a fresh training batch $\mathcal{B}_t$. A bounded subset of errors on pairs with directional targets is retained for reflection:

$$\mathcal{E}_t \subseteq \{(i,j) \in \mathcal{B}_t : y_{ij} \neq 0,\ \widehat{y}_{ij}(p_t) \neq y_{ij}\}\,. \tag{5}$$

For each selected pair, the reflector receives the text, prediction, and target. The outcome-score difference and difficulty category provide additional context about how far apart the observed outcomes are.

Identifying rubric changes. The goal of reflection is to turn errors on individual pairs into revisions that can guide subsequent comparisons. A revision based on a single example may give excessive weight to details specific to that pair. The reflector therefore examines $\mathcal{E}_t$ together with p_t to identify recurring problems in how the rubric uses textual cues. The analysis guides proposed changes that may introduce an overlooked cue, clarify when an existing criterion applies, or adjust the relative importance of competing cues. The proposed changes and their supporting observations are recorded in h_t for the prompt editor.

Implementing rubric revisions. To turn the proposed changes into instructions for subsequent comparisons, the prompt editor revises p_t according to h_t. The edits are confined to the decision

rubric, error warnings, and tie rule, while the task definition, input template, and output format remain unchanged. Keeping this interface fixed focuses the search on the evaluation criteria and allows candidate prompts to be compared under the same task specification.

The editor produces a complete candidate prompt $\widetilde{p}_{t+1}$ and a brief summary of the edits, linking the proposed changes to their implementation. This use of natural-language feedback builds on prior work on iterative refinement (Pryzant et al., 2023; Yuksekgonul et al., 2025; Agrawal et al., 2026). Let $\mathcal{R}$ and $\mathcal{M}$ denote the reflector and prompt editor, respectively. Candidate generation is expressed as

$$h_t = \mathcal{R}(p_t, \mathcal{E}_t), \qquad \widetilde{p}_{t+1} = \mathcal{M}(p_t, h_t). \tag{6}$$

Validation then determines whether the revised criteria improve directional accuracy.

4.2 Validation-Guided Rubric Selection

The revised criteria are intended to guide comparisons beyond the training cases that motivated them. GRACE therefore evaluates each candidate against the current prompt on a separate validation split, using predictive performance to determine whether the revision should be retained. Let $V(p)$ denote the directional validation accuracy defined in Section 3, evaluated under the same protocol for all candidates. The prompt used in the next iteration is

$$p_{t+1} = \begin{cases} \widetilde{p}_{t+1}, & V(\widetilde{p}_{t+1}) > V(p_t), \\ p_t, & \text{otherwise.} \end{cases} \tag{7}$$

Validation thus determines both whether a revision is accepted and which prompt serves as the basis for further search. Starting from p_0, GRACE retains the prompt with the highest validation accuracy observed so far. Search ends when the iteration budget is exhausted or a prespecified number of consecutive iterations yields no improvement. The retained prompt becomes p^* for inference. Algorithm 1 and Appendix B provide the complete procedure and configuration.

4.3 Inference and Selective Output

After search, GRACE applies the selected prompt p^* to new post pairs, with the evaluator and inference settings held fixed. Judgments of the same pair can nevertheless vary across model calls and presentation orders. GRACE therefore combines repeated judgments from both orders and uses their agreement to decide whether to return a prediction.

For each pair, the evaluator makes an equal number of judgments in each presentation order. Swapped-order predictions are mapped back to the original orientation before pooling, so that all judgments refer to the same first and second post. The label with the largest vote count becomes the aggregated prediction $\bar{y}$; a tie for the largest count is treated as a tie prediction. For a directional result, agreement a is the fraction of judgments supporting that label. The system returns $\bar{y}$ only if $\bar{y} \in \{+1, -1\}$ and $a \geq \gamma$; otherwise, it abstains. The two directional labels correspond to A and B in Figure 2.

A higher threshold restricts output to predictions with stronger agreement. Section 5 specifies the operating threshold and evaluation metrics, while Appendix B.2 provides the complete sampling and aggregation protocol.

5 Experiments

Datasets and protocol. The main experiments use six topic-specific datasets with engagement counts recorded approximately 30–90 days after publication. Normalized exact-text groups represented in the final pairs are disjoint across data splits. Relative-popularity targets follow a fixed canonical construction. Dataset statistics, cleaning procedures, target construction, and pair sampling are detailed in Appendix A.

Models and baselines. GRACE uses DeepSeek-V4-Flash as the evaluator and DeepSeek-V4-Pro for both the reflector and prompt editor. Baselines comprise the initial task prompt; three general prompt optimizers, ProTeGi (Pryzant et al., 2023), GEPA (Agrawal et al., 2026), and

MIPROv2 (Opsahl-Ong et al., 2024); and three supervised text models, TF-IDF logistic regression (Salton & Buckley, 1988), a frozen MacBERT linear probe (Cui et al., 2020), and MacBERT-RankNet (Burges et al., 2005). Model configurations, search budgets, and supervised training details are provided in Appendices B and C.

Evaluation. LLM-based methods in the main comparison use ten judgments per pair, five per presentation order, with an acceptance threshold of $\gamma = 0.70$. On test pairs with directional targets, selective directional accuracy measures correctness among accepted predictions, while directional coverage measures the fraction receiving an accepted prediction. The supervised baselines predict a direction for every pair, yielding 100% directional coverage. Results are micro-averaged across datasets, with paired-bootstrap 95% confidence intervals for between-method differences. Full metric definitions and statistical procedures appear in Appendix B.2.

5.1 Overall Results

GRACE improves selective accuracy while retaining high coverage. GRACE achieves 60.43% selective directional accuracy at 84.49% directional coverage, compared with 51.97% and 82.23%, respectively, for the initial task prompt (Table 1). The accuracy gain is 8.46 percentage points (95% CI: $[4.70, 12.32]$). The higher accuracy is accompanied by a modest increase in coverage, indicating that the improvement is achieved without reducing the overall proportion of accepted predictions.

Under the same inference protocol, GRACE also achieves higher selective directional accuracy than all three evaluated general prompt optimizers. The strongest of these, ProTeGi, reaches 55.91% accuracy at 88.29% coverage. GRACE improves accuracy by 4.52 percentage points while accepting a smaller proportion of test pairs. Among supervised text models, TF-IDF logistic regression provides a strong reference, reaching 62.91% accuracy at full directional coverage. Its performance indicates that lexical features carry predictive information about relative popularity. GRACE achieves higher accuracy on accepted predictions than either MacBERT baseline, although both baselines predict every pair with a directional target. GRACE learns its comparison criteria through prompt-level search with fixed evaluator parameters.

Table 1: Micro-averaged test results across the six topic-specific datasets.

Method	Acc. (%) ↑	Cov. (%)
No optimization		
Initial task prompt	51.97	82.23
General prompt optimization		
ProTeGi	55.91	88.29
GEPA	52.41	84.77
MIPROv2	51.18	83.78
Outcome-grounded optimization		
GRACE (ours)	**60.43**	84.49
Supervised text models		
TF-IDF logistic regression	**62.91**	100.00
Frozen MacBERT linear probe	57.69	100.00
MacBERT-RankNet	59.38	100.00

Acc.: selective directional accuracy; Cov.: directional coverage. Bold indicates the highest accuracy among LLM-based methods and among supervised models, separately. LLM-based methods use $\gamma = 0.70$.

Selective accuracy improves across all six topics. Figure 3 shows that GRACE improves selective directional accuracy in every topic, with directional coverage also increasing in four. The rights-protection advice dataset (WQGL) exhibits the largest joint improvement: accuracy rises by 25.15 percentage points and coverage by 13.46 points. In contrast, the civil-service examination (GKSA) and resume-refinement (JLJX) datasets achieve higher accuracy with lower coverage. The gains therefore span topics with both expanded and reduced acceptance.

With the acceptance threshold fixed at $\gamma = 0.70$, these contrasting shifts reflect differences in how often repeated judgments produce an accepted directional prediction. The size of the evaluation set also matters when interpreting the magnitude of topic-level changes. JLJX contains only 23 test pairs with directional targets, and its accepted count decreases from 21 to 17, making its accuracy estimate particularly sensitive to individual predictions. Exact per-topic results are provided in Appendix D.1.

5.2 Analysis of Rubric Search and Inference

Training feedback leads to inspectable rubric revisions. To examine how training feedback changes evaluation criteria, we trace a validation-accepted revision on the social-boundaries topic

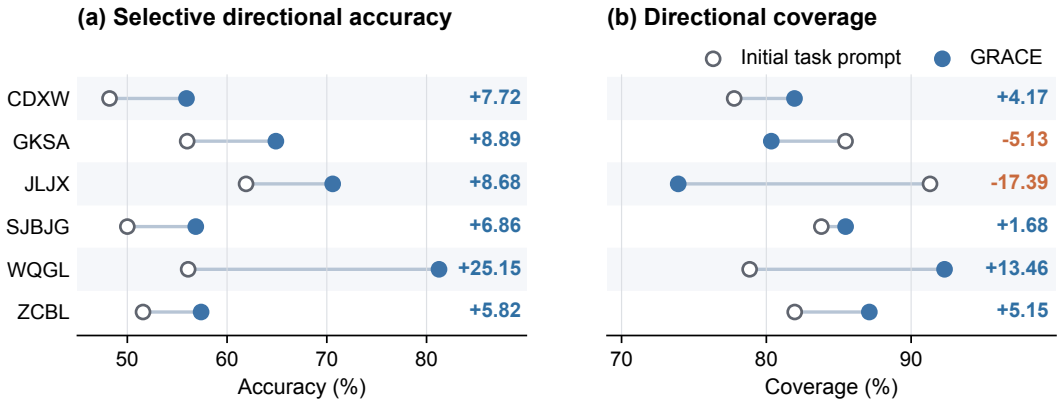


Figure 3: Topic-wise performance at $\gamma = 0.70$: (a) selective directional accuracy and (b) directional coverage. Annotations show GRACE-minus-initial differences in percentage points.

SJBJG | Selected excerpts from one recorded search iteration

Diagnosis. Distinguish advice grounded in personal struggle from a sales pitch; separate intimate emotional expression from polished literary presentation.

BEFORE REVISION

When does advice contribute?

"Structured advice is only influential if the author's personal, emotional struggle is the ***primary substance*** of the post, not just a brief framing device for a list of tips."

How is authenticity judged?

"Distinguish between the "loud" authenticity of a motivational coach (high-energy, many directives, emphatic punctuation) and the "quiet" authenticity of personal reflection (hesitant, questioning, emotionally raw)."

REVISED RUBRIC

Add a contextual distinction

"Advice or tips are engaging ***only if*** the author's personal, emotional struggle is the primary substance and the advice emerges organically from it."

Further distinction: empathetic extension vs. a hook for a solution, product, or persona.

Refine the expressive registers

Quiet reflection: strongest signal.

Motivational coaching: can engage when rooted in personal pain.

Literary reflection: polished abstraction may distance readers from immediate emotion.

Recorded validation accuracy: **61.54% → 63.08%**

Quotation marks identify verbatim excerpts; other text condenses the diagnosis or rule. Validation evaluates the complete candidate prompt.

Figure 4: A validation-accepted rubric revision for the social-boundaries topic (SJBJG). The diagram links the diagnosis from training discrepancies to excerpts and summaries of the rubric before and after revision. The reported directional validation accuracies evaluate the complete prompts.

(SJBJG), one of the six datasets (Figure 4). The feedback motivates finer distinctions between advice grounded in personal experience and advice used as a promotional hook, and between immediate emotional expression and polished literary presentation. The revised rubric clarifies these distinctions and specifies how quiet reflection, motivational language, and literary abstraction should influence a comparison. The complete revised prompt achieves 63.08% directional validation accuracy, compared with 61.54% before revision, and is retained.

The example illustrates how training discrepancies can sharpen the conditions under which a textual cue contributes to the evaluation. The trace links the proposed distinctions to explicit rubric

Table 2: Test performance of prompts selected under different validation rules from the same candidates generated during the default search. All rows use the same test-inference protocol, with acceptance threshold $\gamma = 0.70$.

Selection rule	Sel. acc. (%)	Cov. (%)	Dir. acc. (%)
Single judgment (S0)	55.66	88.43	51.34
Same-order repetition (S1)	58.81	85.61	54.02
Order-balanced repetition (S2)	**60.43**	84.49	**55.71**
Acceptance-aware selection (S3)	59.77	84.49	54.72

Sel. acc.: selective directional accuracy; Cov.: directional coverage; Dir. acc.: accuracy on all test pairs with directional targets before acceptance filtering. S3 selects by selective validation accuracy subject to a coverage constraint. Bold indicates the highest accuracy in each accuracy column.

changes and the candidate's validation result, making the revision process inspectable and providing a concrete basis for further refinement.

Repeated validation judgments improve prompt selection. To assess how validation affects the quality of the selected rubric, we repeat final prompt selection using different validation rules on the same candidates generated during the default GRACE search. The candidate prompts and test-inference protocol remain fixed, so the comparison examines which prompt each validation rule selects (Table 2).

Both same-order and order-balanced repetition select prompts with higher test accuracy than single-judgment selection. The default order-balanced rule improves directional accuracy on the full directional test set by approximately 4.4 percentage points and selective directional accuracy by approximately 4.8 points, with coverage approximately 4.0 points lower. The improvement before acceptance filtering indicates that repeated validation helps select more effective rubrics, beyond changes in the proportion of accepted predictions.

Selecting by selective validation accuracy subject to a coverage constraint provides no additional gain over the default rule. Among the tested rules, the results support repeated judgments with directional validation accuracy for prompt selection, followed by agreement-based acceptance at inference. Appendix D.2 provides the complete selection settings and comparisons.

Agreement helps identify more accurate predictions. With the prompt selected, we examine whether judgment agreement helps determine which predictions to return. We compare agreement-based retention with random retention of the same number of directional predictions, using the same frozen GRACE judgments (Figure 5a). At $\gamma = 0.70$, agreement-based retention achieves approximately 1.8 percentage points higher accuracy than the mean of matched-random retention. Matching the retention size shows that the advantage comes from which predictions are accepted, rather than simply how many are returned. The finding supports agreement as an output-selection signal in this task, consistent with prior work on estimating model confidence from sample consistency (Lyu et al., 2025).

Varying the acceptance threshold further characterizes the trade-off between accuracy and coverage. Relative to $\gamma = 0.70$, requiring unanimous judgments increases selective directional accuracy by approximately 2 percentage points while reducing directional coverage by approximately 30 points. Coverage decreases as the threshold rises, whereas accuracy does not increase monotonically across intermediate thresholds. Agreement therefore provides useful information for selective output, while stronger acceptance requirements can yield modest accuracy gains at a substantial cost in the number of comparisons answered.

5.3 Sensitivity and Additional Evaluations

Gains persist across alternative target definitions. In the main datasets, the mapping g_k in Equation 1 combines recorded counts of likes, comments, and saves with weights $(1, 3, 2)$, respectively, before within-topic percentile ranking (Appendix A.2). To assess sensitivity to this weighting, we assign equal weights to the three interaction types or use each type alone. Ranks and pair labels are recomputed using the same ranking procedure and tie threshold, while prompts, judgments, and acceptance decisions at $\gamma = 0.70$ remain fixed.

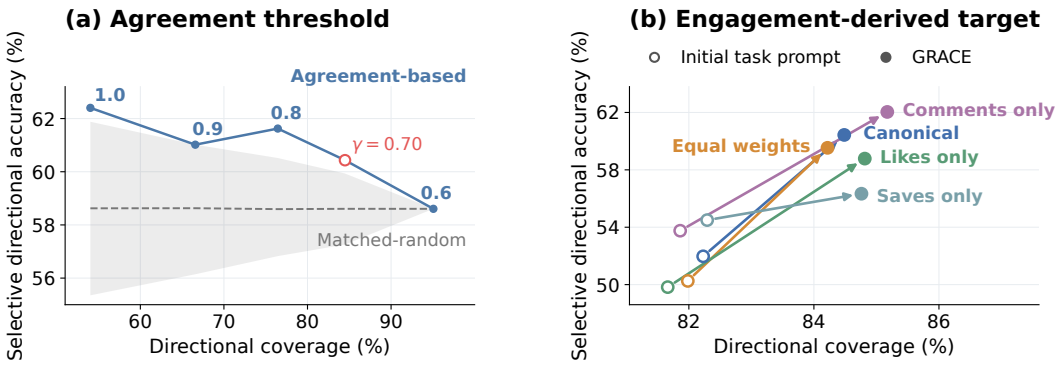


Figure 5: Frozen-prediction analyses. (a) Accuracy–coverage trade-off across agreement thresholds; the dashed line and shaded band show the mean and central 95% interval for random retention at matched output counts. (b) Target sensitivity at $\gamma = 0.70$; arrows connect the initial task prompt (open circles) to GRACE (filled circles).

GRACE retains selective directional accuracy gains of 1.84–9.28 percentage points over the initial task prompt, with higher directional coverage under every alternative (Figure 5b). The equal-weight gain of 9.28 points remains close to the canonical gain of 8.46 points, while the smallest gain occurs for saves-only. The comments-only target changes 36.35% of test labels relative to the canonical target, with directional reversals affecting 7.20% of all test pairs. The gains therefore extend beyond minor adjustments to the canonical labels, although their magnitude varies across targets. Coverage can vary because alternative labels change which pairs enter the directional evaluation set. Appendix D.3 reports target specifications, label changes, and rank comparisons.

Selective accuracy improves across three evaluator models. To examine whether rubric search remains useful with different evaluators, we conduct additional experiments on the workplace-bullying topic (ZCBL) using DeepSeek-V4-Flash, Qwen3.7-Plus, and GLM-5.2. The comparison holds the data partitions, initial task prompt, search budget, and test protocol fixed, with DeepSeek-V4-Pro serving as the reflector and editor throughout. Each evaluator independently searches for its own rubric and is compared with its initial-prompt baseline on the same held-out test set.

GRACE improves selective directional accuracy for all three evaluators, with gains of 1.31–5.52 percentage points. GLM-5.2 improves from 53.57% accuracy at 86.60% coverage to 59.09% at 90.72%, combining more accurate predictions with broader acceptance. Qwen3.7-Plus and DeepSeek-V4-Flash improve selective accuracy with lower coverage; for DeepSeek-V4-Flash, full-set directional accuracy remains unchanged. These results support rubric search across the tested evaluators on this topic, while showing that the accuracy–coverage trade-off depends on the model. Complete results, model identifiers, and search settings are provided in Appendix D.4.

Exploratory cases on Weibo. Two cases from Weibo-COV 2.0 (Hu et al., 2020; Weibo-COV Contributors, 2021) illustrate validation-guided rubric revision. Validation selects a revised rubric for case reporting and retains the initial prompt for infection prevention (Appendix D.5).

6 Conclusion

This work explores how historical engagement feedback can inform evaluation criteria for relative popularity prediction of social media posts. We introduce GRACE, which uses discrepancies between model judgments and observed popularity to guide topic-specific rubric revisions, with validation selecting the prompt used for subsequent comparisons. Across six topic-specific datasets, GRACE achieves 60.43% selective directional accuracy at 84.49% directional coverage, improving on the initial task prompt by 8.46 percentage points. Further analyses show that selective accuracy gains persist under the tested alternative target definitions and across three evaluators on one topic. The revision traces connect historical feedback to explicit changes in evaluation criteria and their validation outcomes, making the rubric search process inspectable. Together, these findings support learning predictive evaluation criteria from noisy behavioral feedback. Future work will examine which criteria can be reused across topics and evaluators, and when further revision is needed.

ETHICS STATEMENT

This study uses Xiaohongshu records supplied by a third-party provider and Weibo-COV 2.0 data to examine the relationship between visible text and behavioral engagement. The evaluator is blinded to author-identity fields. The Xiaohongshu provider did not grant redistribution rights, so raw post text and pairs containing that text will not be released. Use and release of Weibo materials are subject to the applicable source conditions. Textual examples considered for public release are subject to privacy and permission review, and released code and artifacts must exclude unauthorized source content. Engagement-oriented evaluation may amplify platform biases, encourage content homogenization, or facilitate attention manipulation. Creator-facing uses should retain human oversight and interpret outputs in light of content value and the intended context of use.

AI USE STATEMENT

We used generative AI tools to assist with various aspects of the research process, including: literature search and synthesis; discussion of the research framework and experimental design, implementation and debugging of methods and data-processing code, result interpretation and visualization, and manuscript drafting, translation, and revision. The roles and configurations of language models used as experimental components are documented in the method and appendix. The authors are responsible for reviewing AI-assisted text, code, analyses, and citations, and take full responsibility for the research design, findings, and manuscript.

REPRODUCIBILITY STATEMENT

Appendix A describes data preparation, target construction, and pair sampling. Appendices B and C document the search and evaluation protocols and baseline implementations, while Appendix D provides additional configurations and detailed results. The supplementary materials include GRACE source code, configuration files, prompt templates, and selected prompts. Raw text from the main datasets cannot be redistributed; rerunning the experiments requires authorized data access and access to the specified model services.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. Evaluating text creativity across diverse domains: A dataset and large language model evaluator. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3848fef259495bfd04d60cdc5c1b4db7-Abstract-Conference.html.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020. doi: 10.18653/v1/2020.findings-emnlp.58. URL https://aclanthology.org/2020.findings-emnlp.58/.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97

of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Jack Hessel, Lillian Lee, and David Mimno. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 927–936, 2017. doi: 10.1145/3038912.3052684.

Yong Hu, Heyan Huang, Anfan Chen, and Xian-Ling Mao. Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlpcovid19-2.34. URL https://aclanthology.org/2020.nlpcovid19-2.34/.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/08dabd5345b37fffcbe335bd578b15a0-Abstract-Conference.html.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. EvalLM: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024b. doi: 10.1145/3613904.3642216. URL https://doi.org/10.1145/3613904.3642216.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268, 2025. doi: 10.1609/aaai.v39i18.34120.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366, 2024. doi: 10.18653/v1/2024.emnlp-main.525.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, 2023. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0.

Kayla Schroeder and Zach Wood-Doughty. Can you trust LLM judgments? reliability of LLM-as-a-judge. *arXiv preprint arXiv:2412.12509*, 2024. URL https://arxiv.org/abs/2412.12509.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010. doi: 10.1145/1787234.1787254.

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 175–185, 2014. doi: 10.3115/v1/P14-1017. URL https://aclanthology.org/P14-1017/.

Luca Vassio, Michele Garetto, Emilio Leonardi, and Carla Fabiana Chiasserini. Mining and modelling temporal dynamics of followers' engagement on online social networks. *Social Network Analysis and Mining*, 12(1):96, 2022. doi: 10.1007/s13278-022-00928-2.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511.

Weibo-COV Contributors. Weibo-COV: Dataset repository and version 2.0 release, 2021. URL https://github.com/nghuyong/weibo-cov.

Evan Weissburg, Arya Kumar, and Paramveer S. Dhillon. Judging a book by its cover: Predicting the marginal impact of title on Reddit post popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 1098–1108, 2022. doi: 10.1609/icwsm.v16i1.19361.

Erxin Yu, Jing Li, and Chunpu Xu. PopALM: Popularity-aligned language models for social media trendy response prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12867–12878, 2024. URL https://aclanthology.org/2024.lrec-main.1127/.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639:609–616, 2025. doi: 10.1038/s41586-025-08661-4. URL https://www.nature.com/articles/s41586-025-08661-4.

Yu Zhang, Changyang He, Huanchen Wang, and Zhicong Lu. Understanding communication strategies and viewer engagement with science knowledge videos on Bilibili. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581476. URL https://doi.org/10.1145/3544548.3581476.

Yuying Zhao, Yu Wang, Xueqi Cheng, Anne Marie Tumlin, Yunchao Liu, Damin Xia, Meng Jiang, and Tyler Derr. Amplifying your social media presence: Personalized influential content generation with LLMs, 2025. URL https://arxiv.org/abs/2505.01698.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

APPENDIX CONTENTS

A Data Construction and Target Definition **14**

A.1 Topic Inventory and Dataset Statistics 14

A.2 Canonical Target Construction 14

A.3 Text Cleaning and Leakage Control 15

A.4 Pair Sampling 16

B GRACE Implementation and Evaluation Protocol **16**

B.1 Rubric Search Specification 16

B.2 Inference and Evaluation Protocol 18

C Baseline Implementations and Comparison Settings **19**

C.1 Prompt-Optimization Baselines 19

C.2 Supervised Text Baselines 19

D Additional Experimental Details and Results **20**

D.1 Detailed Topic-Level Results 20

D.2 Validation Selection and Selective Output 21

D.3 Sensitivity to Engagement-Target Construction 22

D.4 Multi-Evaluator Experiment 23

D.5 Exploratory Weibo Cases 23

A Data Construction and Target Definition

A.1 Topic Inventory and Dataset Statistics

The six main datasets are constructed around Chinese social media collection queries, which identify distinct discussion contexts. The topics cover emotional expression, personal experiences, practical services, and advice-seeking. This organization supports comparisons within a shared discussion context and allows rubric search to be evaluated across different communicative purposes. Table 3 gives the collection queries in pinyin transliteration with English glosses.

Publication and observation window. Engagement accumulates over time, making post age relevant to the interpretation of observed popularity (Vassio et al., 2022; Szabo & Huberman, 2010). Eligible posts were published between February 1 and April 8, 2026. Engagement counts were supplied by a third-party data provider whose delivery records place the snapshots in May 2026. The resulting observation window is approximately 30–90 days after publication, with every retained post observed at least 30 days after publication. This interval allows engagement to accumulate beyond the immediate post-publication period. The recorded counts provide popularity targets at these snapshots, with observation ages varying across posts.

Dataset composition and preparation. Table 4 summarizes raw and retained post counts, retention rates, post-level and pair-level split sizes, and test-target composition. Figure 6 organizes dataset construction into post preparation and pair construction. Textual cleaning first identifies eligible posts, and normalized exact-text groups identify records with identical textual content. Engagement records are converted to within-topic percentile ranks for split stratification and target construction. Posts are partitioned into training, validation, and test sets before pairing. Within each topic and split, pair sampling excludes same-group comparisons and limits the reuse of each group. Rank differences determine directional or tie targets and define difficulty strata, allowing sampling to include both closely matched and more widely separated pairs. Exact-text groups represented in the final pairs are confined to a single split.

Table 3: Collection queries, their pinyin transliterations with tones omitted, and English glosses.

Topic	Pinyin	English gloss
CDXW	Chedi xingwu	Complete realization
SJBJG	Shejiao bianjiegan	Sense of social boundaries
JLJX	Jianli jingxiu	Resume refinement
ZCBL	Zhichang baling	Workplace bullying
GKSA	Gongkao shangan	Success in the civil-service examination
WQGL	Weiquan gonglue	Guide to protecting one's rights

Table 4: Post-level cleaning, split sizes, and pair statistics for the six topic-specific datasets.

Topic	Context	Posts			Pairs	
		Raw	Retained (%)	T/V/Test	T/V/Test	Test A/B/tie
CDXW	Emotion	1260	1005 (79.76)	503/151/351	250/75/174	69/75/30
GKSA	Experience	2119	799 (37.71)	400/120/279	250/75/139	59/58/22
JLJX	Service	963	165 (17.13)	83/25/57	83/25/28	11/12/5
SJBJG	Emotion	3371	1215 (36.04)	606/183/426	250/74/210	98/81/31
WQGL	Advice	711	367 (51.62)	182/54/131	178/52/62	25/27/10
ZCBL	Experience	2316	1422 (61.40)	712/213/497	250/75/248	96/98/54
Total	–	10740	4973 (46.30)	2486/746/1741	1261/376/861	358/351/152

Retained entries report post counts, with percentages relative to raw posts in parentheses. T/V/Test denotes train/validation/test. A and B indicate higher observed popularity for the first and second post, respectively; directional targets comprise A and B.

A.2 Canonical Target Construction

For the main datasets, the engagement record $e_i = (l_i, c_i, f_i)$ contains the recorded counts of likes, comments, and saves, respectively. The aggregate engagement score is

$$r_i = w_l l_i + w_c c_i + w_f f_i, \qquad (w_l, w_c, w_f) = (1, 3, 2). \tag{8}$$

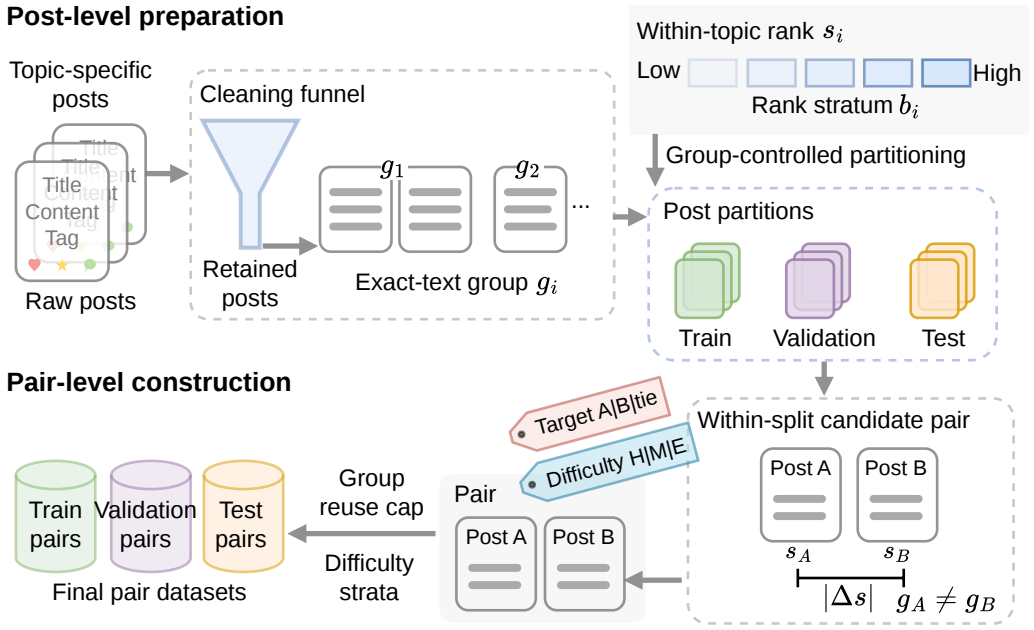


Figure 6: Post preparation and within-split pair construction. In the diagram, g_i, s_i, and b_i denote exact-text-group identifiers, normalized outcome ranks, and rank strata, respectively. H/M/E denotes hard/medium/easy pair difficulty.

These weights define the canonical target and are fixed across topics without tuning on validation or test outcomes.

For a topic with $n_k > 1$ retained posts, let rank_k assign ascending ranks from 1 to n_k, with equal scores receiving their average rank. The normalized popularity score is obtained as

$$R_{ik} = \mathrm{rank}_k(r_i), \qquad s_i = g_k(e_i) = \frac{R_{ik} - 1}{n_k - 1}. \tag{9}$$

The weighted aggregation and percentile transformation together specify the mapping g_k introduced in Equation 1. The implementation applies $\log(1 + r_i)$ before ranking; this strictly increasing transformation preserves the ranks and ties. The resulting score represents relative standing rather than the absolute number of interactions.

The tie threshold in Equation 2 is $\tau = 0.10$ for all topics. A pair therefore receives a tie target when the absolute difference between its normalized scores is at most 0.10. The implementation represents the mathematical labels $+1$, -1, and 0 as `A`, `B`, and `tie`, respectively.

Appendix D.3 examines alternative mappings and their effects on rankings, pair labels, and predictive performance.

A.3 Text Cleaning and Leakage Control

Deterministic screening. Screening aims to retain posts with sufficient visible text for comparison and to apply a consistent inclusion policy across topics. Textual cleaning uses fixed, outcome-blind rules shared across all topics. Visible core characters are counted after removing URLs, hashtags, and user mentions. Table 5 lists the recorded exclusion conditions. The rules screen for limited visible text, image-dependent content, and promotion or solicitation. Retention counts appear in Table 4.

Exact-text grouping and split control. Exact-text grouping does not capture small textual variations, motivating a separate audit of highly similar content across splits. Distinct publication records can contain identical text, so separating records alone does not ensure textual separation between training, validation, and test sets. We normalize each title and body and hash their concatenation

Table 5: Deterministic textual-cleaning rules. Core-character counts include Chinese characters, Latin letters, and digits after removing URLs, hashtags, and user mentions.

Rule	Recorded exclusion condition
Limited visible text	Fewer than 40 visible core characters.
Hashtag-dominated text	At most 20 core characters with at least two hashtags, or fewer than 40 core characters with a hashtag-character ratio of at least 0.65.
Image-dependent text	An image-reference cue with fewer than 120 visible core characters.
Promotion or solicitation	A match to fixed patterns for contact solicitation, gated resources, paid services, or explicit advertising.

to obtain an exact-text-group key. Distinct publication instances with the same key remain separate records. The key is used to enforce split-level separation and pair-construction constraints: a pair cannot contain two posts from the same group, and every group represented in the formal pairs occurs in only one split.

Group separation is enforced during group-aware post splitting or, for datasets with existing post-level assignments, during formal pair selection. Group-level assignments use seed 42, target a 0.50/0.15/0.35 train/validation/test ratio, and are stratified over five bins of the percentile score. Group integrity takes priority over exact split size; Table 4 reports the realized counts.

Near-duplicate audit. A label-blind audit uses character 5-gram Jaccard similarity and a length-ratio filter to identify near-duplicate posts. Candidate and high-similarity thresholds are fixed at 0.80 and 0.90, respectively. For each topic, the audit measures the fraction of test posts with a high-similarity neighbor in training or validation. The highest observed fraction across the six topics is below the prespecified 5% rebuild threshold. Exact-text groups remain the unit of formal split-level leakage control.

A.4 Pair Sampling

Pairs are sampled within each topic and split, excluding posts from the same exact-text group. Group-level degree caps are two in training, two in validation, and one in test. Thus, no test group contributes to more than one pair. When sufficient candidates are available, each topic contains at most 250 training pairs and 75 validation pairs.

To include comparisons with different degrees of outcome separation, we stratify candidate pairs by the absolute difference between their normalized popularity scores:

$$\text{difficulty}(i,j) = \begin{cases} \text{hard}, & 0 \le |s_i - s_j| < 0.15, \\ \text{medium}, & 0.15 \le |s_i - s_j| < 0.35, \\ \text{easy}, & 0.35 \le |s_i - s_j| \le 1. \end{cases} \tag{10}$$

Sampling targets an easy/medium/hard ratio of 0.45/0.30/0.25, subject to candidate availability and group-level reuse constraints. This stratification controls the mix of closely matched and more widely separated pairs.

Difficulty categories describe outcome separation, whereas the tie threshold $\tau = 0.10$ determines whether a pair has a directional target. Training uses directional-target pairs, while validation and test retain both directional and tie targets. Sampling configurations and random seeds are fixed before test evaluation.

B GRACE Implementation and Evaluation Protocol

B.1 Rubric Search Specification

Search procedure. GRACE begins with the initial task prompt p_0 and records its validation performance. At each iteration, the retained prompt evaluates a fresh training batch. Prediction errors on pairs with directional targets provide a bounded set of cases for reflection. The reflector examines these cases and proposes rubric changes, which the prompt editor implements in a candidate prompt.

Algorithm 1 Grounded Reflection for Automated Criteria Exploration (GRACE)

Require: Initial task prompt p_0; training split D_{tr}; validation split D_{val}; iteration budget T; patience K

$p_{\mathrm{best}} \leftarrow p_0$
$v_{\mathrm{best}} \leftarrow V(p_0; D_{\mathrm{val}})$; $c \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  Sample a fresh training batch $\mathcal{B}_t \subset D_{\mathrm{tr}}$
  Evaluate the text pairs in $\mathcal{B}_t$ using p_{best}
  Retain at most eight cases with $y_{ij} \neq 0$ and $\widehat{y}_{ij}(p_{\mathrm{best}}) \neq y_{ij}$ in $\mathcal{E}_t$
  if $\mathcal{E}_t = \emptyset$ **then**
    $c \leftarrow c + 1$
  else
    $h_t \leftarrow \textsc{Reflect}(p_{\mathrm{best}}, \mathcal{E}_t)$
    $\widetilde{p}_{t+1} \leftarrow \textsc{Edit}(p_{\mathrm{best}}, h_t)$
    $\widetilde{v} \leftarrow V(\widetilde{p}_{t+1}; D_{\mathrm{val}})$
    if $\widetilde{v} > v_{\mathrm{best}}$ **then**
      $p_{\mathrm{best}} \leftarrow \widetilde{p}_{t+1}$
      $v_{\mathrm{best}} \leftarrow \widetilde{v}$; $c \leftarrow 0$
    else
      $c \leftarrow c + 1$
    end if
  end if
  if $c \geq K$ **then**
    break
  end if
end for
return selected prompt $p^* \leftarrow p_{\mathrm{best}}$

The candidate replaces the retained prompt only after a strict increase in directional validation accuracy. Otherwise, search continues from the retained prompt. Search stops when the iteration budget or patience limit is reached. Algorithm 1 gives the complete procedure. Here, V denotes directional validation accuracy under the canonical order-balanced protocol.

Information flow and editing scope. GRACE turns training feedback into rubric revisions through reflection and editing. The reflector examines selected error cases in the context of the current prompt and proposes changes for the editor to implement. Table 6 maps the direct inputs supporting these operations and the evaluator's comparisons during search and inference.

To focus revisions on the comparison criteria, editing is restricted to the decision rubric, error warnings, and tie instructions. The task definition, input template, and output format remain unchanged, providing a consistent interface across candidates. Each revision is stored with a change summary and its validation result, linking the implemented edits to their observed performance.

Table 6: Direct inputs to GRACE components. A check mark indicates a supplied input; a dash indicates that it is not supplied. Changes refers to rubric revisions proposed by the reflector.

Component	Pair text	Prompt	Prediction and target	Score gap and difficulty	Changes
Evaluator (search)	✓	✓	–	–	–
Error identification	–	–	✓	–	–
Reflector	✓	✓	✓	✓	–
Prompt editor	–	✓	–	–	✓
Evaluator (inference)	✓	✓	–	–	–

Canonical configuration. The six main topic-specific searches share the model roles, search budget, validation criterion, and inference settings in Table 7.

Model-specific settings and terminal selection rules for the additional evaluator comparison are provided in Appendix D.4.

Table 7: Canonical GRACE configuration for the six main topic-specific datasets.

Component or decision	Setting
Evaluator	DeepSeek-V4-Flash
Reflector and prompt editor	DeepSeek-V4-Pro
Generation settings	Temperature 0.7; thinking disabled
Training batch and judgments	32 pairs; 3 original-order judgments per pair
Error cases	At most 8 cases with incorrect predictions on directional targets
Search budget	At most 8 iterations; patience 3
Editing scope	Decision rubric, error warnings, and tie instructions
Validation selection	Directional accuracy; strict improvement
Validation and test judgments	10 per pair, with 5 in each presentation order
Selective operating point	Directional prediction with agreement $a \geq 0.70$

B.2 Inference and Evaluation Protocol

Pooling and acceptance. Repeated judgments provide an aggregate comparison and an agreement signal for selective output. To give both presentation orders equal weight, the canonical validation and test protocol uses five judgments in each order. Swapped-order predictions are mapped back to the original orientation before pooling.

For pair (i, j), let $n_{ij}(z)$ denote the number of mapped judgments supporting label $z \in \{-1, 0, +1\}$. The pooled prediction $\bar{y}_{ij}$ is the unique label with the largest vote count; a tie for the largest count yields $\bar{y}_{ij} = 0$. Agreement is

$$a_{ij} = \frac{\max_{z\in\{-1,0,+1\}} n_{ij}(z)}{10}. \tag{11}$$

Let $\mathcal{D}_{\text{dir}}$ contain the test pairs with nonzero target labels. At test time, the accepted subset is

$$\mathcal{A}_\gamma = \{(i, j) \in \mathcal{D}_{\text{dir}} : \bar{y}_{ij} \in \{-1, +1\},\ a_{ij} \geq \gamma\}, \qquad \gamma = 0.70. \tag{12}$$

Evaluation metrics. Because GRACE can abstain, evaluation considers both the correctness of returned predictions and the proportion of comparisons they cover. Selective directional accuracy measures correctness on $\mathcal{A}_\gamma$, while directional coverage measures the fraction of $\mathcal{D}_{\text{dir}}$ that is accepted. Directional accuracy provides a complementary reference by evaluating pooled predictions on the full directional set without agreement filtering:

$$\text{SelDirAcc}(\gamma) = \frac{\sum_{(i,j)\in\mathcal{A}_\gamma} \mathbf{1}[\bar{y}_{ij} = y_{ij}]}{|\mathcal{A}_\gamma|}, \tag{13}$$

$$\text{DirCov}(\gamma) = \frac{|\mathcal{A}_\gamma|}{|\mathcal{D}_{\text{dir}}|}, \tag{14}$$

$$\text{DirAcc} = \frac{\sum_{(i,j)\in\mathcal{D}_{\text{dir}}} \mathbf{1}[\bar{y}_{ij} = y_{ij}]}{|\mathcal{D}_{\text{dir}}|}. \tag{15}$$

Under directional accuracy, pooled tie predictions count as incorrect, whereas directional predictions are evaluated regardless of their agreement scores. Canonical prompt selection uses the same directional accuracy definition on validation pairs, without agreement filtering. Aggregate results are micro-averaged by pooling the relevant counts across topics. Pairs with tie targets are excluded from these directional metrics; their counts are reported in Table 4.

Uncertainty estimation. Differences between methods are evaluated with 10,000 paired bootstrap resamples of shared test pairs. Each resample uses the same pair indices for both methods, and the metrics and their differences are recomputed within that resample. The 2.5th and 97.5th percentiles form the reported 95% interval. Accepted subsets may differ between methods. The intervals quantify test-sample uncertainty conditional on the recorded predictions, rather than variation across independent searches or API runs.

C Baseline Implementations and Comparison Settings

Baselines use the same topic splits and held-out test pairs as GRACE. Training supports prompt search or parameter fitting, and validation selects prompts and model configurations. Test evaluation follows after these choices are frozen. The initial task prompt uses the same evaluator and test-inference protocol as GRACE. Request-retry and parsing rules are fixed before testing.

C.1 Prompt-Optimization Baselines

ProTeGi, GEPA, and MIPROv2 start from the shared initial task prompt p_0 and use the same topic splits and task-model family. Each retains its method-specific candidate-generation and selection procedure. ProTeGi uses the budget-aligned implementation below, whereas GEPA and MIPROv2 use their specified `auto=light` configurations. Candidate counts and model-call allocations therefore vary across optimizers.

ProTeGi. The implementation preserves textual-gradient generation, gradient-guided instruction editing, beam search, and bandit-based candidate selection (Pryzant et al., 2023). It uses eight search rounds, training batches of 32 pairs, three evaluator judgments per training pair, a beam of four prompts, and at most two optimizer-model calls per round. Candidates remain zero-shot and instruction-only. The selected prompt is evaluated under the same frozen test protocol as GRACE.

GEPA. GEPA evolves prompt candidates through reflection on execution and evaluation feedback (Agrawal et al., 2026). The experiments use DSPy 3.2.1 with GEPA 0.0.27 and `auto=light`. Each topic uses seed 42, a reflection minibatch of three examples, Pareto candidate selection, and at most five merge invocations. Feedback includes expected and predicted labels, output-format validity, and an instruction to re-examine the pair using its visible text. The selected prompt is frozen for the common test-inference protocol.

MIPROv2. MIPRO optimizes instructions and demonstrations (Opsahl-Ong et al., 2024). The experiments use MIPROv2 in DSPy 3.2.1 with Optuna 4.9.0. The shared configuration sets `auto=light`, seed 42, and zero bootstrapped or labeled demonstrations, yielding an instruction-only final program. The minibatch size is 35 examples, and full evaluation is scheduled every five trials.

C.2 Supervised Text Baselines

TF-IDF logistic regression. Title, body text, and tags are concatenated for each post. A shared character TF-IDF vectorizer uses n-grams for $n \in \{2, 3, 4\}$. Pairs are represented by the antisymmetric difference $\phi(x_i) - \phi(x_j)$, and logistic regression predicts the directional label. Vectorizer and regularization settings are selected using validation data, with the model frozen before test evaluation.

Frozen MacBERT linear probe. Each post is encoded using the last-layer `[CLS]` representation from `hfl/chinese-macbert-base`. The encoder is frozen, embeddings are L2-normalized, and a logistic head without an intercept is trained on the difference between the two post embeddings. The regularization value is selected from $\{0.01, 0.1, 1, 10\}$ using directional validation pairs. All topics use a maximum sequence length of 512 and an embedding batch size of 16.

MacBERT-RankNet. The model uses `hfl/chinese-macbert-base` with a shared post scorer. Pair probabilities are defined from differences between post scores. Training uses at most six epochs, validation-based early stopping, and seed 42 for the reported run.

Appendix D.1 reports the complete topic-level results for the prompt-based and supervised methods.

D Additional Experimental Details and Results

D.1 Detailed Topic-Level Results

Aggregate results can obscure differences in gains, acceptance rates, and baseline rankings across topics. We therefore examine whether the improvements extend across discussion contexts and whether they persist when all directional test pairs are evaluated. Topic abbreviations and their English glosses are provided in Table 3.

Selective gains accompany different coverage changes. At the fixed threshold $\gamma = 0.70$, GRACE improves selective directional accuracy across all six topics, with coverage increasing in four (Table 8). For civil-service examination experiences (GKSA) and resume refinement (JLJX), higher selective accuracy instead accompanies lower coverage. The aggregate improvement therefore combines topic-level outcomes with different acceptance patterns. To distinguish improvements in the predictions from changes in which predictions are accepted, we also examine directional accuracy before agreement-based filtering.

Table 8: Topic-wise selective results at $\gamma = 0.70$. Sel. acc. denotes selective directional accuracy; Cov. denotes directional coverage. Metric values are percentages. Differences are GRACE minus the initial task prompt, calculated from unrounded values and reported in percentage points.

	Initial task prompt		GRACE		Difference	
Topic	Sel. acc.	Cov.	Sel. acc.	Cov.	Sel. acc.	Cov.
CDXW	48.21	77.78	55.93	81.94	+7.72	+4.17
GKSA	56.00	85.47	64.89	80.34	+8.89	-5.13
JLJX	61.90	91.30	70.59	73.91	+8.68	-17.39
SJBJG	50.00	83.80	56.86	85.47	+6.86	+1.68
WQGL	56.10	78.85	81.25	92.31	+25.15	+13.46
ZCBL	51.57	81.96	57.40	87.11	+5.82	+5.15

Accuracy gains persist before agreement filtering. Table 9 evaluates the aggregated predictions on all test pairs with directional targets. GRACE improves directional accuracy over the initial task prompt in every topic and achieves the highest accuracy among the evaluated LLM-based methods in five of the six topics. On workplace bullying (ZCBL), ProTeGi performs better, reaching 54.12% compared with GRACE's 53.09%. Thus, the improvements are not confined to the accepted subset, although the strongest LLM-based method still depends on the topic.

The magnitude of a topic-level gain should also be interpreted alongside its evaluation-set size. For resume refinement (JLJX), the increase from 60.87% to 65.22% corresponds to one additional correct prediction among 23 directional test pairs. Dataset sizes in Table 4 therefore provide important context for comparing the observed changes across topics.

Table 9: Directional test accuracy (%) of LLM-based methods by topic and micro-averaged across topics. All test pairs with directional targets are included, without agreement-based filtering; aggregated tie predictions count as errors.

Method	CDXW	GKSA	JLJX	SJBJG	WQGL	ZCBL	Micro
Initial task prompt	44.44	52.99	60.87	44.69	50.00	47.94	47.81
ProTeGi	43.75	53.85	56.52	50.84	67.31	54.12	52.19
GEPA	41.67	53.85	52.17	46.37	48.08	53.09	48.80
MIPROv2	49.31	50.43	52.17	48.60	48.08	44.33	47.95
GRACE	50.00	58.97	65.22	53.63	76.92	53.09	55.71

Supervised comparisons reveal topic-dependent strengths. To compare predictive performance on the same evaluation set, Table 10 reports supervised accuracy on all directional test pairs. TF-IDF logistic regression achieves the highest accuracy among the supervised models in five topics, including a tie with MacBERT-RankNet on workplace bullying (ZCBL). MacBERT-RankNet performs best on resume refinement (JLJX), so the encoder-based models do not consistently outperform the lexical baseline.

Across all directional test pairs, TF-IDF logistic regression reaches 62.91% micro-averaged accuracy, exceeding GRACE's 55.71%. The ordering differs for rights-protection advice (WQGL), where GRACE reaches 76.92% compared with 73.08% for TF-IDF logistic regression. These results establish a strong supervised reference and show that the relative advantage of rubric search varies across discussion contexts.

Table 10: Directional test accuracy (%) of supervised text models. All methods predict a direction for every test pair with a directional target. Bold indicates the highest accuracy within each row, including ties.

Topic	TF-IDF logistic regression	Frozen MacBERT linear probe	MacBERT-RankNet
CDXW	**68.75**	59.72	61.81
GKSA	**69.23**	59.83	58.97
JLJX	69.57	69.57	**73.91**
SJBJG	**59.78**	55.31	58.10
WQGL	**73.08**	67.31	71.15
ZCBL	**54.12**	53.09	**54.12**
Micro	**62.91**	57.69	59.38

D.2 Validation Selection and Selective Output

Validation determines which prompt is retained, while agreement determines which predictions are returned during inference. We examine these decisions separately: the validation analysis compares final prompt choices from the same recorded candidates, and the selective-output analysis compares agreement-based and random retention at matched coverage.

Comparing validation selection rules. To examine whether repeated validation judgments improve final prompt selection, we apply four selection rules to the same recorded candidate prompts for each topic, including the initial task prompt. The rules differ in how validation judgments are collected or used to rank candidates:

- **Single judgment (S0).** Select by directional validation accuracy using one original-order judgment per pair.
- **Same-order repetition (S1).** Pool five original-order judgments per pair and select by directional validation accuracy.
- **Order-balanced repetition (S2).** Pool five judgments in each presentation order, mapping reversed-order predictions to the original orientation before aggregation. Select by directional validation accuracy.
- **Acceptance-aware selection (S3).** Use the same judgments as S2, but select by selective directional validation accuracy at $\gamma = 0.70$, requiring validation coverage at least as high as that of the initial task prompt.

The candidates were generated during GRACE search using the default order-balanced rule. The comparison therefore evaluates final selection within this recorded set, rather than independently rerunning search under each rule. Every selected prompt is evaluated using the same frozen test-inference protocol, so differences in test performance reflect the prompt choices rather than different test-time sampling settings.

Repeated judgments improve final prompt selection. Both same-order and order-balanced repetition select prompts with higher selective and full-set directional test accuracy than single-judgment selection (Table 2). The paired-bootstrap intervals for these accuracy differences lie above zero, while coverage decreases (Table 11). The results support using repeated validation judgments to identify more accurate prompts within the recorded candidate set, with a corresponding reduction in accepted predictions.

Order-balanced repetition yields the highest accuracy point estimates, but its accuracy-difference intervals relative to same-order repetition include zero. The two rules also differ in judgment count,

so their comparison does not isolate the contribution of presentation-order balancing. Acceptance-aware selection provides no further observed improvement over the default rule. Within this comparison, selecting by full-set directional validation accuracy remains effective without incorporating the acceptance threshold into the selection objective.

Table 11: Paired-bootstrap 95% intervals for test-metric differences between validation selection rules, using 10,000 resamples. Differences subtract the second rule from the first and are reported in percentage points. Sel. acc., Cov., and Dir. acc. denote selective directional accuracy, directional coverage, and full-set directional accuracy, respectively.

Contrast	Sel. acc.	Cov.	Dir. acc.
S1–S0	$[1.15, 5.17]$	$[-5.13, -0.56]$	$[0.56, 4.85]$
S2–S0	$[2.54, 7.07]$	$[-6.62, -1.30]$	$[2.08, 6.78]$
S3–S0	$[1.93, 6.28]$	$[-6.50, -1.43]$	$[1.09, 5.71]$
S2–S1	$[-0.36, 3.63]$	$[-3.55, 1.40]$	$[-0.43, 3.88]$
S3–S2	$[-1.83, 0.49]$	$[-1.68, 1.65]$	$[-2.26, 0.28]$

Agreement-based and matched-random retention. The analysis in Figure 5a examines whether agreement identifies more accurate predictions than random selection at the same retention size. GRACE's selected prompts and pooled test predictions remain fixed throughout this analysis.

For each agreement threshold, we record the number of accepted predictions on test pairs with directional targets. The random reference samples the same number uniformly without replacement from GRACE's pooled directional predictions on those pairs, before agreement filtering. Matching the retained count also matches directional coverage, allowing the comparison to focus on which predictions are retained.

We draw 10,000 random subsets at each threshold using random seed 42. The dashed line reports their mean accuracy, and the shaded band spans the 2.5th and 97.5th percentiles. The band describes variation across random subsets, rather than uncertainty from repeated model calls. At $\gamma = 0.70$, agreement-based retention achieves approximately 1.8 percentage points higher accuracy than the matched-random mean.

D.3 Sensitivity to Engagement-Target Construction

The canonical target combines several engagement measures, but different weightings may change which posts are considered more popular. We examine how alternative targets change post rankings and pair labels, and whether GRACE's improvement over the initial task prompt persists under those changes. The alternatives use equal weights or a single engagement measure, as specified in Table 12. Retained posts, pair identities, data partitions, the percentile-ranking procedure, and the tie threshold remain unchanged.

Rank and label changes. To establish how substantially the alternatives change the evaluation task, we examine both post-level rankings and the labels of the existing test pairs. For each topic, Spearman correlation measures agreement between the alternative and canonical rankings. Table 12 reports the correlation range across the six topics and the tie rate among the 861 test pairs.

Table 12: Target construction and its effects on ranks and test labels. Weights correspond to likes, comments, and saves. Rank ρ is the range of within-topic Spearman correlations with the canonical ranking. The final three columns are percentages of all 861 test pairs. Changes and reversals are measured relative to the canonical labels.

Target	(w_l, w_c, w_f)	Rank ρ range	Tie rate	Label changes	Direct reversals
Canonical	$(1, 3, 2)$	1.000–1.000	17.65	0.00	0.00
Equal weights	$(1, 1, 1)$	0.987–0.995	16.84	6.85	0.00
Likes only	$(1, 0, 0)$	0.856–0.947	18.93	19.40	2.09
Comments only	$(0, 1, 0)$	0.550–0.844	26.36	36.35	7.20
Saves only	$(0, 0, 1)$	0.753–0.883	19.98	29.50	5.81

Rank agreement alone does not fully describe changes to the pairwise task, because score differences near the tie threshold can change labels even when the overall rankings remain similar. Table 12

therefore reports label changes and direct directional reversals. A label change includes transitions between a directional label and a tie, whereas a direct reversal exchanges the two directional labels. Both rates use all 861 test pairs as the denominator. Equal weights closely preserve the canonical rankings, with correlations of 0.987–0.995. Although 6.85% of test labels change, none undergo a direct directional reversal. The single-channel targets produce larger changes. Comments-only has the lowest rank agreement, changes 36.35% of test labels, and directly reverses 7.20% of all test pairs. It also produces the highest tie rate. The alternatives therefore range from modest changes around the tie boundary to substantial changes in the ordering and composition of directional comparisons.

Evaluation with fixed predictions. To assess the effect of target construction without introducing changes from another rubric search, we rescore the recorded predictions of the initial task prompt and GRACE's prompts selected under the canonical target. Prompts, judgments, and acceptance decisions at $\gamma = 0.70$ remain fixed.

For each target, metrics are recomputed over the test pairs whose new labels are directional. Changing labels can alter both prediction correctness and membership in the directional evaluation set. Consequently, that set and its accepted subset can change even though the acceptance decision for each individual pair remains unchanged. The resulting coverage changes reflect this relabeling, rather than a change in the acceptance rule.

GRACE maintains higher selective directional accuracy and directional coverage than the initial task prompt under all four alternatives (Figure 5b). The largest changes to the target do not coincide with the smallest accuracy gain: comments-only retains an 8.28-percentage-point gain, whereas saves-only yields a 1.84-point gain despite changing fewer labels. Together with the rank and label analyses, these results show that the predictive gains extend beyond alternatives that closely preserve the canonical target. Their magnitude nevertheless varies with the engagement behavior used to define relative popularity.

D.4 Multi-Evaluator Experiment

We examine whether evaluator-specific rubric search remains useful when the task evaluator changes. The experiment uses the workplace-bullying topic (ZCBL), which has the largest directional test set among the six main datasets, with 194 directional targets among 248 pairs. The three evaluators share the data partitions, initial task prompt, search budget, and test protocol, while the reflector and editor remain fixed. Each evaluator independently searches for its own prompt, which is compared with its initial-prompt baseline on the same held-out test set. The DeepSeek-V4-Flash results are obtained from separate baseline and rubric-search runs for this experiment, rather than reused from the main ZCBL comparison.

Shared settings and final selection. Table 13 specifies the models and shared experimental settings. Search follows the prompt-update procedure described in Appendix B. Final selection additionally applies the validation eligibility conditions and tie-breakers listed in the table. These terminal selection conditions are specific to this additional experiment and are identical across evaluators. All three selected prompts satisfy the eligibility conditions and are revisions of the initial task prompt.

Selective gains extend across the tested evaluators. GRACE improves selective directional accuracy for all three evaluators, with gains of 1.31–5.52 percentage points (Table 14). GLM-5.2 achieves the largest gain while increasing directional coverage by 4.12 points. The other two evaluators improve selective accuracy with lower coverage; DeepSeek-V4-Flash's full-set directional accuracy remains unchanged. These results support the usefulness of evaluator-specific rubric search on this topic, with model-dependent accuracy–coverage trade-offs.

D.5 Exploratory Weibo Cases

We complement the main experiments with two exploratory cases from Weibo-COV 2.0 (Hu et al., 2020; Weibo-COV Contributors, 2021), covering infection prevention and case reporting. These cases illustrate how the workflow operates with different post formats and engagement measures, including both adoption of a revised rubric and retention of the initial task prompt.

Table 13: Model identifiers and shared settings for the three-evaluator experiment on ZCBL. Final selection uses validation data only.

Setting	Specification
Models and serving conditions	
DeepSeek-V4-Flash evaluator	`deepseek-v4-flash-0731`
Qwen3.7-Plus evaluator	`qwen3.7-plus-2026-05-26`
GLM-5.2 evaluator	`glm-5.2`
Reflector and editor	`deepseek-v4-pro-0813`
Provider	Model Studio (DashScope), shared endpoint
Generation settings	Temperature 0.7; thinking disabled for all model roles
Search and evaluation	
Pairs: train/validation/test	250/75/248
Search budget	At most eight iterations; batch size 32; early-stopping patience three; random seed 42
Training judgments	Three per pair in the original order
Validation and test judgments	Ten per pair, five per presentation order
Test acceptance threshold	$\gamma = 0.70$
Final prompt selection	
Eligibility conditions	Swap consistency ≥ 0.60 and ungated directional-output fraction ≥ 0.55, both measured over all validation pairs
Primary criterion	Directional validation accuracy
Tie-breakers, in order	(1) Accuracy among validation pairs receiving an ungated directional prediction; (2) fraction of validation pairs with pooled vote agreement ≥ 0.70; (3) lower predicted tie rate

Swap consistency compares the order-specific aggregate predictions after mapping them to a common orientation. The first tie-breaker includes pairs with tie targets and does not apply agreement filtering. The second includes agreement on any label, including a tie.

Table 14: ZCBL test results for three task evaluators. Sel. acc. and Cov. denote selective directional accuracy and directional coverage at $\gamma = 0.70$. Dir. acc. evaluates all 194 test pairs with directional targets before agreement filtering. All values are percentages.

	Initial task prompt			GRACE		
Evaluator	Sel. acc.	Cov.	Dir. acc.	Sel. acc.	Cov.	Dir. acc.
DeepSeek-V4-Flash	50.92	84.02	48.45	52.23	80.93	48.45
Qwen3.7-Plus	52.81	91.75	49.48	58.18	85.05	52.06
GLM-5.2	53.57	86.60	50.00	59.09	90.72	55.67

Data and target construction. The cases use the January, March, and May 2020 partitions of Weibo-COV 2.0. We retain original posts with at least 40 core-text characters and an observation age of at least 30 days. Topic subsets are identified through keyword retrieval followed by text-based topic classification. Unlike the main datasets, these posts contain body text and hashtags but no titles. The target combines likes, comments, and reposts with weights $(1, 3, 2)$, respectively, replacing saves with reposts as the third engagement measure. Percentile ranks are computed separately for each topic and calendar month, and pairs are formed within the same month to keep comparisons within a common calendar context.

The split-before-pairing procedure and exact-text-group separation follow Appendix A. For these cases, splitting is stratified by topic, month, and outcome bin, and pair construction excludes same-group endpoints. Each topic contains 300 training, 90 validation, and 180 test pairs. Training pairs have directional targets, while validation and test pairs retain ties under $\tau = 0.10$. Endpoint reuse is capped at two pairs in training and validation and one pair in test. The infection-prevention and case-reporting test sets contain 118 and 100 directional-target pairs, respectively.

Validation selection and test results. The requested model aliases are `deepseek-v4-flash` for the evaluator and `deepseek-v4-pro` for the reflector and editor. Prompts are selected by directional validation accuracy and frozen before testing. Test evaluation follows the order-balanced

aggregation and selective-output protocol in Appendix B.2, with $\gamma = 0.70$. Within each topic, the initial and selected conditions are evaluated on the same test pairs and targets.

Table 15 shows two distinct selection outcomes. For case reporting, the selected revision raises directional validation accuracy from 66.67% to 78.95%. Its selective test accuracy improves by 11.47 percentage points, with directional coverage decreasing by 7.00 points. For infection prevention, no candidate improves on the initial prompt's 55.22% validation accuracy, so the initial prompt is retained. The two infection-prevention rows therefore represent separate inference runs of the same prompt. Their differences reflect inference variability, not an improvement attributable to rubric search.

Table 15: Exploratory Weibo results at $\gamma = 0.70$. Acc. denotes selective directional accuracy; Cov. denotes directional coverage. The retained initial prompt is evaluated in a separate inference run without any prompt changes.

Topic	Prompt condition	Acc. (%)	Cov. (%)
Infection prevention	Initial task prompt	59.05	88.98
	Retained initial prompt	63.37	85.59
Case reporting	Initial task prompt	62.92	89.00
	Selected revision	74.39	82.00

Together, these cases illustrate the role of validation in deciding whether rubric revision is warranted: a revised prompt is adopted when validation accuracy improves, while the initial prompt remains available when the proposed changes provide no validation gain.