FOMCBENCH: A HUMAN-ANNOTATED BENCHMARK FOR FEDERAL RESERVE Q&A RESPONSIVENESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cautious language can answer a question, while a fluent, on-topic response can leave it unanswered. We introduce FOMCBENCH, a human-annotated benchmark of answer responsiveness: 2,154 English question–answer pairs from 89 Federal Reserve press conferences (2011–2025). Three finance professionals independently annotated every pair, blind to model predictions. The release provides 6,462 judgments, fixed splits, and human-majority references for binary responsiveness and three-class evasion degree. Three-class inter-annotator agreement is 0.791 (Fleiss' κ). Human training labels disagree with a three-model majority on 26.3% of pairs for three-class degree and 21.5% for binary responsiveness. We compare eight open-weight and three API models under pair-only and full-meeting inputs, alongside Eva-4B's native pair-input baseline. All use common references, scoring and meeting-bootstrap uncertainty. On the primary test, meeting-minus-pair binary Macro-F1 differences range from -8.0 to $+24.9$ percentage points; three-class comparisons can reverse direction. The contrast does not isolate context or model openness. Versioned votes, cached predictions and executable scoring support reproducible comparisons, with explicit limits from shared meetings, subjective judgments and source-text fidelity.

1 INTRODUCTION

A central-bank answer can be cautious yet informative, or discuss the right topic without resolving the question. Distinguishing these cases requires comparing the information requested with the answer supplied. FOMCBENCH provides an English-language text benchmark for this problem: fixed question–answer pairs, documented reference labels, and reproducible evaluation of answer responsiveness.

For example, a reporter asks which policy path would produce the Federal Reserve's projected economic conditions. Bernanke replies, "I don't think it would be constructive for me to give you a tentative projection…" The answer does not provide the requested path; its recorded human-majority label is fully evasive (abridged Test-Challenge example; Appendix H). That label describes responsiveness, not whether withholding a projection is justified or what the speaker intends. Conversely, uncertainty or qualification alone does not make an answer evasive.

Federal Reserve communication conveys market-relevant information through language and delivery, including vocal tone and facial expressions (Gorodnichenko et al., 2023; Curti & Kazinnik, 2023). Central-bank messages can also be interpreted differently by different listeners (Blinder et al., 2008). We evaluate responsiveness in the text of the exchange.

The closest evasion benchmarks evaluate earnings calls against a largely single-annotator gold set (Ma et al., 2026) or political interviews (Thomas et al., 2026). FOMCBENCH provides three professional degree judgments for every pair and human-majority references for Train, Dev and both tests. This makes model-generated references directly auditable: human and model-consensus Train labels agree on only 73.7% of three-class judgments. We use this audit and an eleven-model evaluation to show how the reference source, scoring target and input protocol change benchmark conclusions.

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.

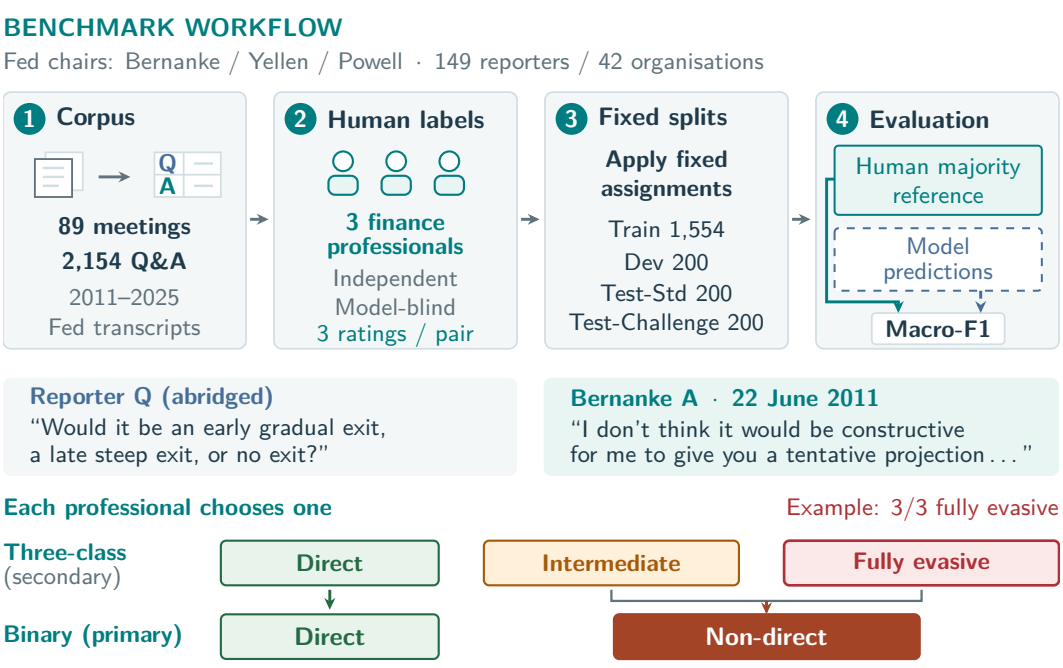


Figure 1: **FOMCBENCH: from press-conference exchanges to evaluation.** Corpus construction, professional degree annotation and fixed split assignments support model evaluation. Each professional chooses direct, intermediate or fully evasive; binary scoring groups the latter two as non-direct. Human judgments and model predictions enter scoring separately. The abridged exchange illustrates a unanimous fully-evasive judgment.

Contributions.

1. **A dataset and benchmark.** We release 2,154 Q&A pairs with three professional degree judgments each, fixed splits, model-derived auxiliary annotations, and an executable scorer (Section 3).
2. **An auditable evaluation.** We evaluate eight open-weight and three API models with pair-only and full-meeting inputs, plus Eva-4B with its native pair input, using common human references, scoring and meeting-level uncertainty. We report both targets, invalid outputs and provenance; supplementary adapted baselines retain their training conditions (Section 4).
3. **Evidence about reference and protocol choices.** Human and model-majority Train references differ on 21.5% of pairs under binary scoring and identify 120 versus 49 fully evasive pairs. Meeting input raises every open-weight binary point estimate, API directions are mixed, and three-class comparisons can reverse (Section 5).

2 Related Work

Evasion and responsiveness benchmarks. Earlier work measures managerial non-answers to analysts' questions in corporate conference calls using linguistic analysis (Gow et al., 2021). Ma et al. (2026) introduced EVASIONBENCH (84K training samples, 1K gold evaluation) and fine-tuned Eva-4B. Its Gold-1K uses one annotator with a 100-item second-annotator check; validation beyond earnings calls is future work. FOMCBENCH retains three professional judgments throughout Train, Dev and tests. CLARITY (Thomas et al., 2026) extends the political-interview response-clarity dataset of Thomas et al. (2024), retaining its three-way clarity and nine-class fine-grained taxonomy. Financial Q&A studies established the three-axis framework (Al Nuaimi et al., 2025b) and reported 1,728 pairs from 83 FOMC conferences (Al Nuaimi et al., 2025a). FOMCBENCH adopts this domain and framework, providing 2,154 pairs from 89 conferences with professional degree references, fixed splits, executable scoring and protocol/reference analyses. Exact record overlap is unverified; the difference in corpus totals is not a count of new examples (Appendix G). Unlike the

multipart-question decomposition in Thomas et al. (2024), our unit is the recorded Q&A exchange. CLARITY's any-annotator-match rule (Thomas et al., 2026) also differs from our fixed-majority evaluation. Ferracane et al. (2021) study subjective conversation acts and perceived intentions in U.S. congressional hearings. FOMCBENCH measures professional judgments of answer responsiveness without inferring speaker intent.

Protocol sensitivity. Laban et al. (2026) report an average 39% performance drop across six generation tasks when instructions are distributed over simulated turns; complementary benchmarks examine multi-turn conversational performance (Kwan et al., 2024; Bai et al., 2024; Sirdeshmukh et al., 2025). These findings motivate testing protocol sensitivity, but do not establish that full-meeting context improves FOMC classification.

Evasion taxonomies. Bull & Mayer (1993) developed a political non-reply typology; Bull (2003) examined equivocation in interviews. Rasiah (2010) distinguished answers, intermediate responses and evasions in parliamentary discourse. Bavelas et al. (1986) developed a situational account of equivocal political communication. Rogers & Norton (2011) showed that listeners can miss a dodge when a speaker answers a similar question. Our degree, mechanism and tactic categories operationally adapt these traditions.

Central bank communication. Text analyses document systematic patterns in Federal Reserve communication (Hansen et al., 2018; Shah et al., 2023; Ahrens & McMahon, 2021); Blinder et al. (2008) survey the broader communication literature. The World Central Banks corpus (Shah et al., 2025) benchmarks stance, temporal orientation, and uncertainty across 25 central banks, while Fed-specific systems support exploratory analysis (Lee et al., 2021). Vocal tone (Gorodnichenko et al., 2023) and facial expressions (Curti & Kazinnik, 2023; Alexopoulos et al., 2024) provide information beyond transcripts. These studies examine policy language, stance, and delivery; FOMCBENCH focuses on question–answer responsiveness in press conferences.

Discourse and hedging detection. Hedging—linguistic devices that attenuate commitment—has been studied in academic discourse (Hyland, 1998); related tasks include stance detection (Mohammad et al., 2016) and the rhetorical analysis of parliamentary Q&A (Ilie, 2015).

LLM annotation. LLMs can outperform crowd workers on selected text-labeling tasks (Gilardi et al., 2023; Törnberg, 2023); Ziems et al. (2024) distinguish classification agreement from gains in free-form explanations. Annotation performance depends on model settings and fine-tuning (Alizadeh et al., 2025), and nuanced constructs require task-specific validation (Pangakis et al., 2023; Ashwin et al., 2026). Meta-evaluations show judge bias (Chen et al., 2024), while human-agreement-aware evaluation provides a framework for deciding when to trust model judgments (Jung et al., 2025). Our dataset documentation is informed by Gebru et al. (2021) and Bender & Friedman (2018) (Appendix G). Benchmark-quality work recommends stating purpose, documenting design and limitations, and planning maintenance (Longjohn et al., 2024); BetterBench also specifies contamination safeguards (Reuel et al., 2024).

3 THE FOMCBENCH DATASET

3.1 CORPUS CONSTRUCTION AND RELEASE

FOMCBENCH contains 2,154 English-language Q&A pairs from all 89 Federal Reserve press conferences between April 2011 and December 2025, covering Bernanke (12), Yellen (16), and Powell (61), with 149 reporters from 42 organisations. Official PDF transcripts provide speaker-attributed text. Construction also uses public video recordings and Whisper-based transcripts for timestamp alignment, but the released benchmark is text-based.

The pipeline parses speaker turns, normalises names and organisations, aligns timestamps, removes non-substantive turns, and constructs Q&A pairs before annotation (Figure 1; details in Appendix A). Each released row contains the question, answer, meeting and speaker metadata, and annotations. The package supplies fixed splits, cached predictions, a scorer, dataset documentation, licenses, and

Croissant metadata. Per-speech and per-section representations and aligned media are construction artifacts outside this release. Table 6 in Appendix B gives detailed corpus statistics.

Access and reuse. The accompanying anonymous data/code archive contains the complete text benchmark and human degree references, licensed under CC BY 4.0, with scoring code under MIT. Checksums bind the records and results. Cached predictions allow rescoring without access to the evaluated models.

The frozen text retains two source-confirmed Dev omissions and 20 other editorial-text differences from a separately processed version. Appendix A.1 identifies the records and affected meeting contexts. Existing references and predictions have not been validated for corrected transcripts.

3.2 Responsiveness and Diagnostic Labels

We retain the three-axis operational schema of prior financial Q&A work (Al Nuaimi et al., 2025b;a):

- **Degree** adapts the three-way distinction of Rasiah (2010). In our rubric, *direct* answers address the requested information; *intermediate* answers engage with the question but leave information unresolved; *fully evasive* answers substantially fail to engage with its core request. Hedging alone does not imply evasion, while omissions in multipart questions can warrant an intermediate label.
- **Mechanism** describes how a non-direct answer falls short: vagueness, non-sequitur, omission, or restatement. These are operational categories informed by equivocation theory (Bavelas et al., 1986).
- **Tactic** uses 12 operational strategies adapted from the prior financial-Q&A scheme (Al Nuaimi et al., 2025b), including agenda shifting, partial answer, and refusal to answer. Table 35 specifies the current definitions.

We refer to these annotation axes as Rasiah degree, Bavelas mechanism, and Bull tactic. Released mechanism, tactic, and score fields come from Opus 4.6 under Condition D and lack independent human validation. The prompt assigns N/A to mechanism and tactic for direct outputs; these auxiliary fields can differ from the released degree reference. The 1–5 score, stored as `evasion_score`, increases with *responsiveness*. It is distinct from model-reported confidence and can disagree with the degree label (Appendix G).

3.3 Annotation and Reference Sources

Professional annotation across the corpus. Three finance professionals, including one author, independently annotated all 2,154 pairs. Their expertise included quantitative research and signal development. They worked through separate annotation access, without model predictions or one another's ratings. The task concerns the information requested and the response supplied, rather than the speaker's intent. The documented clarification rule permits an immediately following, journalist-confirmed continuation when the chair explicitly requests a reminder (Appendix B.2).

Every pair has three retained judgments and resolved binary and three-class references, produced through professional review and majority vote. Agreement statistics describe the retained ratings before that review: full-corpus three-class Fleiss' κ (Fleiss, 1971) is 0.791 (95% interval $[0.772, 0.810]$), and binary κ is 0.826 $[0.805, 0.846]$. Intervals resample all 89 meetings as whole units (10,000 draws, seed 20260926), without estimating new-rater or protocol variation. Appendix B.2 describes the aggregation and retained records.

On the 400 test pairs, three-class κ is 0.676, and all 108 disagreements are between adjacent degrees; binary κ is 0.736 on Test-Std and 0.594 on Test-Challenge, which oversamples model disagreement.

Model-derived comparison labels. Opus 4.6, GPT-5.2 and Gemini 3 Pro supplied a separate candidate reference under meeting-level prompting with reasoning (Condition D, Appendix D). Their majority vote uses Opus 4.6 to break five three-way ties. These model outputs informed split selection: Dev contains 200 model-unanimous pairs, while the tests use model-derived strata and disagreements (Section 3.4). Section 5.1 compares this candidate reference with human judgments.

Supplementary diagnostics distinguish the human references from the *Opus diagnostic reference* and model-consensus Dev labels (Tables 25 and 31). These diagnostics measure model-reference

Table 1: Fixed, pair-disjoint splits with current human-majority degree references. Split selection preceded full human annotation and remains unchanged. All splits share meetings. Class counts describe the selected splits, not population prevalence.

Split	N	Direct	Interm.	Evasive	Original selection
Train	1,554	936	498	120	Remaining pairs
Dev	200	113	82	5	Model-unanimous
Test-Std	200	88	102	10	Chair/model-class strata
Test-Challenge	200	53	124	23	Model disagreement
Total	2,154	1,190	806	158	

Every pair has resolved binary and three-class references.

agreement. Shared model families and inputs limit their independence; LLM judges also raise consistency concerns (Wang et al., 2026).

3.4 Evaluation Splits

The four pair-disjoint splits are Train (1,554), Dev (200), Test-Std (200), and Test-Challenge (200); a convenience file combines the two test splits. Test-Std is the primary split and uses chair and model-derived class strata. Test-Challenge enriches disagreement among the three selection models and between Eva-4B and Opus 4.6 (Appendix B.2); its composition is therefore conditional on those systems. Neither test split estimates population prevalence.

All 89 meetings occur in Train, including all 80 Test-Std and 81 Test-Challenge meetings. Evaluation therefore concerns held-out exchanges from shared meetings. The human-majority references contain 158 fully evasive pairs: 120 Train, five Dev, ten Test-Std, and 23 Test-Challenge. Sparse test support motivates the primary binary task and explicit uncertainty for three-class results.

We report the test splits separately and pool them only for stated diagnostics. A lower challenge score reflects difficulty under disagreement-based selection.

4 Evaluation Setup

4.1 Tasks and Scoring

Binary responsiveness (primary; T1b) predicts direct versus non-direct, grouping intermediate and fully evasive. **Three-class degree (secondary; T1)** retains all three labels. Both use the fixed human-majority test references. The primary metric is Macro-F1, which weights the fixed target classes equally. The scorer also reports accuracy, per-class F1, a confusion matrix, and Cohen's κ (Cohen, 1960). Missing and invalid predictions count as errors on the complete split; label aliases are normalised. Appendix J gives executable commands. Throughout, "F1" means Macro-F1 unless qualified.

Test-Std contains only ten fully evasive examples. Binary scoring reduces dependence on this rare class, while secondary three-class and per-class scores retain the distinctions it obscures. We therefore report both targets, invalid-output counts and uncertainty. Appendix D.5 gives the cached baselines' class-level intervals and boundary sensitivity.

Auxiliary analyses of mechanisms, cross-corpus labels and chair partitions use model-derived references. Their scope and limitations appear in the supplement (Figure 3); they are not human-validated evaluation targets. No cross-domain task or unseen-chair retraining is implemented in this release.

4.2 Models and Input Protocols

Eleven models have **pair-only (G)** and **full-meeting (B)** predictions. G targets one frozen Q&A record; B supplies its released meeting records in chronological order. The eight open-weight runs use a shared degree/mechanism/tactic framework, domain-neutral financial-Q&A wording and fixed

Table 2: Models and evaluated inputs. Parameter sizes are publisher declarations in billions (B), with component conventions retained.

Model	Size (B)	Developer	Country/ region	Input
Qwen3.8-Flash-Next	125[a]	Qwen / Alibaba	China	G/B
Gemma-4-31B	30.7	Google DeepMind	UK	G/B
GPT-OSS-120B	117	OpenAI	USA	G/B
Mistral-Small-4	119	Mistral AI	France	G/B
Nemotron-3-Super	120	NVIDIA	USA	G/B
Apertus-70B	70	Swiss AI Initiative	Switzerland	G/B
GLM-4.7-Flash	30	Z.ai / ZhipuAI	China	G/B
DeepSeek-V4.1-Flash	552[b]	DeepSeek	China	G/B
Opus 4.6	N/V	Anthropic	USA	G/B
GPT-5.2	N/V	OpenAI	USA	G/B
Gemini 3 Pro	N/V	Google DeepMind	UK	G/B
Eva-4B	4	EvasionBench team	China (HK/Macao)	Native pair

G: pair-only; B: full-meeting. Eva-4B uses its native pair input; meeting input is not evaluated. Locations identify documented developer or academic offices, not data origin or exclusive development location. N/V: size not verified in inspected sources. [a]Language-model component; 51B n-gram embeddings and 4B MTP are separate. [b]Backbone; 196B conditional memory is separate. Appendix C gives sources and counting qualifications.

Table 3: Original model-derived Train labels versus final human-majority references on the same 1,554 pairs. Rows are the original labels; columns are the human references. This is a label-source audit, not a held-out model comparison.

Original / human	**Direct**	**Intermediate**	**Fully evasive**	**Total**
Direct	734	129	3	866
Intermediate	202	366	71	639
Fully evasive	0	3	46	49
Total	936	498	120	1,554

demonstrations. The three API systems—Opus 4.6, GPT-5.2 and Gemini 3 Pro—use cached G/B degree exports from historical runs of the project's three-axis framework prompt; their exact wire inputs and dated snapshots are not recorded (Appendix D). Eva-4B supplies a separate native pair-input baseline, not a G-framework run; no meeting-input result is evaluated for it. Scoring is common across groups, without established prompt or native inference equivalence.

Meeting inputs include later exchanges and text assigned to other splits, without reference labels. This is offline classification of exchanges from shared meetings. The protocols also differ in batching and wrapper structure, so their contrast does not isolate context alone. The annotation guideline's narrow continuation rule can give human raters evidence absent from a pair-only input.

Open-weight degree extraction follows a declared policy chosen without human-reference scores; API evaluation uses the frozen exported degrees. Neither view filters on auxiliary-field completeness. Invalid or missing degrees count as errors on all 200 targets per split. We report native settings and known exposure where recoverable; missing provenance is not a verified match.

Additional API conditions and adapted baselines appear in Appendix D, with their model-derived training labels where applicable. The main comparison holds scoring and references fixed; differences in prompts, training and inference settings preclude attributing group differences to model openness.

5 Results and Analysis

5.1 Human and Model Reference Comparison

Model-consensus and human-majority references disagree on 408 of 1,554 Train pairs for three-class degree (26.3%) and on 334 for binary responsiveness (21.5%). Agreement is 73.7% $[71.5, 75.9]$ and 78.5% $[76.5, 80.5]$, respectively, using 95% meeting-bootstrap intervals. These compare candidate references, not held-out prediction errors. Differences between the model prompt and the human clarification rule are a possible source of disagreement, not an identified cause (Appendix B.2).

The discrepancy extends to unanimous human judgments: 299 of 1,317 such Train pairs differ from the model-consensus reference. Humans identify 120 fully evasive Train pairs, compared with 49 in the model reference; only 46 receive that label from both sources. Reference choice therefore changes both class support and which examples represent each degree.

Dev's model-unanimous reference agrees with human majorities on 189 of 200 pairs in both tasks (94.5% $[91.4, 97.3]$). Its selection, class composition and tuning history differ from Train and preclude treating it as an untouched test. The contrast supports documenting label provenance, without establishing that model unanimity predicts annotation quality. Table 3 gives the full Train cross-tabulation.

5.2 Model Evaluation under Pair and Meeting Inputs

Among the eight open-weight models, all binary point estimates increase under meeting inputs. Test-Std increases range from 3.3 percentage points (pp) for GLM to 24.9 pp for Nemotron; Test-Challenge increases range from 0.4 pp for Apertus to 17.3 pp for DeepSeek. Qwen has this group's highest observed Test-Std score, 86.1% $[80.5, 91.1]$.

The API results show a different pattern. Gemini 3 Pro's pair-only Test-Std score is 88.0% $[82.7, 92.6]$, the highest point estimate in the G/B panel. Its meeting-input score is 79.9%, a change of -8.0 pp $[-14.2, -1.9]$; its Challenge change is -8.8 pp $[-16.6, -0.9]$. GPT-5.2's binary point estimates rise on both splits, while Opus 4.6 rises on Test-Std and falls on Test-Challenge. These API systems also contributed to split construction, so their scores are conditional on model-informed selection (Section 3.4). The table does not establish a model-openness effect or a statistically resolved overall ranking.

The paired intervals qualify these point differences. Gemma's Test-Std increase is 5.5 pp with a 95% meeting-bootstrap interval of $[-0.7, 11.6]$ pp; its Challenge increase is 2.5 pp $[-6.1, 11.4]$. All main model intervals use the same 10,000 whole-meeting resamples (seed 20260924). They are exploratory and pointwise, without simultaneous coverage or confirmatory significance claims (Appendix C).

The complete secondary results (Table 5) show that the scoring target can reverse a protocol comparison. For Apertus on Test-Std, binary Macro-F1 rises from 35.9% to 42.8%, while three-class Macro-F1 falls from 38.1% to 29.7%. On Test-Challenge its binary change is $+0.4$ pp $[-1.9, 3.8]$, whereas its three-class change is -11.5 pp $[-19.4, -3.3]$. These are pointwise descriptive contrasts, not a general benefit or cost of meeting context.

Per-class scores help interpret these aggregate differences (Table 10). On Test-Std, Apertus's direct-class F1 rises from 0.0% under G to 14.0% under B, while fully evasive F1 falls from 46.2% to 10.0%. Binary scoring merges the latter class with intermediate answers and consequently hides that distinction. Reporting the secondary task exposes this tradeoff, although only ten fully evasive test examples support its class-level estimate.

Apertus has one invalid Test-Std/B degree, counted as an error; a fallback view that resolves it gives 42.7% rather than 42.8% binary Macro-F1. Recorded inference settings match across G and B on all 400 test pairs for Apertus and DeepSeek; they are unknown for Qwen, for nine pairs in each other open-weight model, and for all API systems. No targets are filtered on this basis. Excluding nine open-weight test pairs with known prior exposure is a separate sensitivity (Appendix C); no model is certified unseen.

Native pair-input baseline. Eva-4B scores 81.3% binary Macro-F1 on Test-Std and 73.5% on Test-Challenge under its native pair input, with one invalid prediction per split counted as an error. Its

Table 4: Binary Macro-F1 (%) on the fixed human-reference tests. Pair/Meeting use G/B for framework systems; Eva-4B uses its native pair prompt.

	Test-Std ($n = 200$)		Test-Challenge ($n = 200$)	
Model	Pair	Meeting	Pair	Meeting
Open-weight models				
Qwen3.8-Flash-Next	67.3 [60.7, 73.3]	86.1 [80.5, 91.1]	61.9 [52.8, 71.0]	74.6 [68.8, 80.3]
Gemma-4-31B	77.2 [71.1, 82.7]	82.7 [76.8, 87.9]	71.9 [62.7, 80.3]	74.3 [67.5, 80.9]
GPT-OSS-120B	51.0 [43.8, 57.9]	71.5 [63.9, 78.4]	55.7 [48.6, 62.9]	63.3 [56.1, 70.5]
Mistral-Small-4	37.1 [34.0, 40.3]	45.3 [39.5, 51.4]	42.4 [40.4, 44.1]	52.5 [45.7, 59.5]
Nemotron-3-Super	43.0 [37.9, 48.2]	67.8 [61.1, 74.3]	46.2 [41.7, 52.0]	56.4 [49.6, 63.5]
Apertus-70B	35.9 [33.4, 38.2]	42.8 [37.4, 48.4]†	42.4 [40.4, 44.1]	42.7 [39.8, 46.6]
GLM-4.7-Flash	41.6 [36.4, 47.3]	44.9 [39.0, 50.9]	48.7 [42.6, 55.8]	53.2 [46.1, 59.8]
DeepSeek-V4.1-Flash	67.0 [60.7, 72.7]	82.4 [77.1, 87.2]	61.9 [53.4, 70.0]	79.2 [72.0, 85.5]
Closed-source/API models				
Opus 4.6‡	80.8 [74.6, 86.5]	83.0 [77.3, 88.0]	75.5 [67.4, 82.8]	71.5 [64.6, 78.6]
GPT-5.2‡	82.8 [77.1, 87.8]	85.3 [80.2, 89.9]	75.6 [67.4, 82.9]	77.6 [71.5, 83.0]
Gemini 3 Pro‡	88.0 [82.7, 92.6]	79.9 [74.6, 84.8]	77.6 [71.0, 83.2]	68.8 [62.2, 75.1]
Native pair-input baseline				
Eva-4B§	81.3 [75.6, 86.5]	—	73.5 [66.9, 80.1]	—

Every evaluated cell retains 200 targets; invalid degrees count as errors. Apertus Std/B has one invalid; Eva-4B has one per split; other cells have none. Dashes mean not evaluated; Eva-4B is not Condition G. Brackets: exploratory, unadjusted 95% whole-meeting intervals. Open rows use qualified original-output degrees; API rows use frozen degree exports. Common scoring uses the same references, scorer and interval method; identical prompts and native settings are not established (Appendix C). ‡ These systems' Condition-D outputs informed Test-Std strata and Test-Challenge selection. § Disagreements between Eva-4B and Opus 4.6 B also informed Test-Challenge selection.

disagreements with Opus 4.6 B also informed Test-Challenge selection. No meeting-input result is evaluated, so it does not enter the G/B contrasts.

6 Discussion and Conclusion

FOMCBench provides professional references and reproducible evaluation for Federal Reserve Q&A. The reference source matters: human and model-consensus Train labels differ on 21.5% of binary labels and disagree on which pairs are fully evasive (120 versus 49). The input protocol also matters, but not uniformly: meeting input raises every open-weight binary point estimate, API directions are mixed, and three-class comparisons can reverse. Scores should therefore be reported with their reference, target and input access.

Using the benchmark. Hold the target, split, input and reference fixed; report Test-Std binary, three-class and per-class scores; use Test-Challenge for diagnostics; and check scoring against the cached predictions.

Table 5: Three-class Macro-F1 (%) on the same fixed human-reference tests. Classes: direct, intermediate and fully evasive; Pair/Meeting use G/B for framework systems; Eva-4B uses its native pair prompt.

	Test-Std ($n = 200$)		Test-Challenge ($n = 200$)	
Model	Pair	Meeting	Pair	Meeting
Open-weight models				
Qwen3.8-Flash-Next	64.5 [53.8, 72.2]	82.3 [72.8, 89.1]	70.3 [62.7, 77.5]	78.1 [72.5, 83.3]
Gemma-4-31B	63.4 [54.1, 71.9]	80.0 [70.4, 87.0]	75.4 [67.9, 82.3]	78.4 [72.0, 84.2]
GPT-OSS-120B	47.3 [36.7, 56.0]	61.6 [48.8, 71.7]	53.7 [45.8, 61.1]	59.4 [50.7, 67.5]
Mistral-Small-4	42.6 [32.9, 49.9]	46.5 [37.7, 54.1]	52.4 [46.3, 57.4]	56.2 [48.1, 63.1]
Nemotron-3-Super	40.0 [28.1, 51.0]	54.4 [40.9, 66.6]	52.9 [45.3, 59.3]	46.2 [36.5, 56.3]
Apertus-70B	38.1 [23.0, 48.3]	29.7 [23.5, 37.9]	48.6 [42.2, 54.3]	37.1 [29.2, 44.8]
GLM-4.7-Flash	39.0 [28.1, 48.5]	40.8 [28.1, 51.0]	52.3 [44.6, 59.5]	49.6 [40.6, 57.5]
DeepSeek-V4.1-Flash	62.0 [50.3, 71.0]	75.6 [64.0, 84.3]	68.7 [61.2, 75.2]	76.1 [67.4, 83.4]
Closed-source/API models				
Opus 4.6‡	75.2 [61.7, 84.4]	72.9 [58.8, 83.1]	78.5 [71.1, 85.3]	67.9 [59.2, 75.4]
GPT-5.2‡	84.2 [76.0, 90.1]	83.5 [74.3, 90.0]	76.4 [68.2, 84.0]	80.9 [75.3, 85.8]
Gemini 3 Pro‡	78.0 [68.0, 85.8]	77.1 [68.0, 84.0]	80.4 [74.4, 85.4]	72.6 [66.3, 78.0]
Native pair-input baseline				
Eva-4B§	77.6 [67.6, 84.9]	—	74.8 [67.2, 81.6]	—

Every evaluated cell retains 200 targets; Apertus Std/B and each Eva-4B split have one invalid degree; the remaining cells have none. Shared scoring, interval, output-provenance and ‡/§ split-selection conventions follow Table 4.

Limitations and interpretation. The evidence concerns English-language U.S. Federal Reserve exchanges and a subjective professional rubric (Section 3.3). Auxiliary labels lack independent human validation. Only ten fully evasive Test-Std examples support three-class estimates; binary scoring reduces rare-class sensitivity but retains direct/intermediate boundary uncertainty.

Prior test inspection makes the metric prioritisation and analyses retrospective. Main intervals resample whole meetings; supplementary analyses identify any pair resampling. Intervals condition on retained votes and outputs, excluding new-rater or training variation and dependence across meetings. Shared meetings limit generalisation; contamination cannot be ruled out for this public corpus. Appendix G details the assumptions constraining small-gap progress tracking.

Future work. A next step is to test transfer to other central banks, earnings-call Q&A and additional languages, using locally validated labels and appropriate redistribution rights. Prospective splits based on unseen meetings would better assess generalisation. Controlled studies should separate the effects of context and instructions, and compare training on human versus model-derived labels. Extending professional annotation to mechanisms and tactics, while retaining individual judgments and disagreements, would support richer evaluation of how answers avoid the question.

AI USE STATEMENT

Generative AI tools assisted organisation extraction, transcript quality control, prompt-based model evaluation, research software, manuscript drafting and revision, figure preparation, and analysis and scoring code. The model-reference panel and auxiliary annotations are described in their respective methods. Codex assisted reference and provenance checks, cached-score reproduction, scientific critique, editing, and artifact verification; Claude Code and Claude Cowork provided additional critique. Feedback on analysis and interpretation was checked against retained data, code, and cited sources. These checks cannot eliminate all errors. The authors remain responsible for all claims, references, code, and data, including AI-assisted material.

ETHICS STATEMENT

FOMCBENCH can support research and public-accountability analyses of how officials answer policy questions. It measures answer responsiveness against a rubric; it does not measure deception, honesty, beliefs, motives, speaker credibility, or institutional trustworthiness. We discourage individual-speaker surveillance or scoring and employment, credentialing, or credibility decisions about identifiable people. Automated judgments of institutional trustworthiness are outside the benchmark's validated use. This guidance describes intended use and does not add restrictions to the CC-BY-4.0 license. Three finance professionals independently annotated all 2,154 public-transcript pairs, blind to model outputs and one another's ratings. They contributed voluntarily to a joint institutional research project and agreed to the research release of their anonymized ratings. Whether this annotation activity required institutional ethics review remains unresolved; no approval or exemption is claimed. Section 3.3 describes the rubric, aggregation and annotation limits. Auxiliary mechanism, tactic and score labels are model-derived. No label establishes an utterance's sincerity.

REPRODUCIBILITY STATEMENT

The anonymous supplement contains all 2,154 Q&A records, fixed splits, 6,462 human degree judgments, and resolved binary and three-class majority references. Underlying ratings and reference derivation are preserved; reported agreement describes retained ratings before final reference review. Dev and test ratings are released as anonymous triplets. Source hashes bind the data, cached predictions, scoring code and numerical outputs. Agreement intervals use 10,000 whole-meeting bootstrap samples (seed 20260926); the main G/B results for all eleven models use 10,000 whole-meeting samples with seed 20260924. Missing or invalid predictions count as errors on the full reference population. Common scoring does not establish matching prompts or native inference settings; the package identifies those limits and adapted baselines' training-label provenance. Appendix J gives executable commands and identifies the unavailable full-corpus Opus 4.6 Condition-D reference, inference/training inputs, per-system auxiliary predictions, individual Dev/test exports, the earnings-call subset and multimodal media. The source-text erratum documents known differences between the frozen benchmark records and official transcripts.

REFERENCES

Maximilian Ahrens and Michael McMahon. Extracting economic signals from central bank speeches. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pp. 93–114, 2021. doi: 10.18653/v1/2021.econlp-1.12. URL https://aclanthology.org/2021.econlp-1.12/.

Khaled Al Nuaimi, Gautier Marti, Alexis Marchal, Mohamad Alansari, and Andreas Henschel. Evasive answers in financial Q&A: Earnings calls vs. FOMC press conferences. In *NeurIPS 2025 Workshop on Generative AI in Finance*, 2025a. URL https://openreview.net/forum?id=A1FDpZ0Kdg.

Khaled Al Nuaimi, Gautier Marti, Alexis Marchal, and Andreas Henschel. Detecting evasive answers in financial Q&A: A psychological discourse taxonomy and lightweight baselines. In *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing*, pp. 191–196, 2025b. doi: 10.18653/v1/2025.finnlp-2.12. URL https://aclanthology.org/2025.finnlp-2.12/.

Michelle Alexopoulos, Xinfen Han, Oleksiy Kryvtsov, and Xu Zhang. More than words: Fed chairs' communication during congressional testimonies. *Journal of Monetary Economics*, 142:103515, 2024. doi: 10.1016/j.jmoneco.2023.09.002.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D. Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17, 2025. doi: 10.1007/s42001-024-00345-9.

Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. Using large language models for qualitative analysis can introduce serious bias. *Sociological Methods & Research*, 55(3):795–839, 2026. doi: 10.1177/00491241251338246.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, 2024. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Janet Beavin Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. A situational explanation of discourse: The case of political disqualification. *Working Papers of the Linguistics Circle*, 5(1):33–44, 1986. URL https://journals.uvic.ca/index.php/WPLC/article/view/5046.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041.

Alan S. Blinder, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen. Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–945, 2008. doi: 10.1257/jel.46.4.910.

Peter Bull. *The Microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge, London, 2003. doi: 10.4324/9780203417843.

Peter Bull and Kate Mayer. How not to answer questions in political interviews. *Political Psychology*, 14(4):651–666, 1993. doi: 10.2307/3791379.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Filippo Curti and Sophia Kazinnik. Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences. *Journal of Monetary Economics*, 139:110–126, 2023. doi: 10.1016/j.jmoneco.2023.06.007.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1626–1644, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.129. URL https://aclanthology.org/2021.naacl-main.129/.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971. doi: 10.1037/h0031619.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. The voice of monetary policy. *American Economic Review*, 113(2):548–584, 2023. doi: 10.1257/aer.20220129.

Ian D. Gow, David F. Larcker, and Anastasia A. Zakolyukina. Non-answers during conference calls. *Journal of Accounting Research*, 59(4):1349–1384, 2021. doi: 10.1111/1475-679X.12371. URL https://doi.org/10.1111/1475-679X.12371.

Stephen Hansen, Michael McMahon, and Andrea Prat. Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, 133(2):801–870, 2018. doi: 10.1093/qje/qjx045.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Ken Hyland. *Hedging in Scientific Research Articles*. John Benjamins Publishing, Amsterdam, 1998. doi: 10.1075/pbns.54.

Cornelia Ilie. Parliamentary discourse. In Karen Tracy, Cornelia Ilie, and Todd Sandel (eds.), *The International Encyclopedia of Language and Social Interaction*. Wiley, 2015. doi: 10.1002/9781118611463.wbielsi201.

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.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 2004. doi: 10.1111/j.1468-2958.2004.tb00738.x.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, 2024. doi: 10.18653/v1/2024.emnlp-main.1124. URL https://aclanthology.org/2024.emnlp-main.1124/.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/59f6421e64707225fdf5b28840679a07-Abstract-Conference.html.

Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. FedNLP: An interpretable NLP system to decode federal reserve communications. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2560–2564, 2021. doi: 10.1145/3404835.3462785. URL https://doi.org/10.1145/3404835.3462785.

Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. Benchmark data repositories for better benchmarking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 86435–86457. Curran Associates, Inc., 2024. doi: 10.52202/079017-2744. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9d41fddf3169f4df3025fc60bfac94fa-Paper-Datasets_and_Benchmarks_Track.pdf.

Shijian Ma, Yan Lin, and Yi Yang. EvasionBench: A large-scale benchmark for detecting managerial evasion in earnings call Q&A. *arXiv preprint arXiv:2601.09142*, 2026. URL https://arxiv.org/abs/2601.09142v2.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41, 2016. doi: 10.18653/v1/S16-1003.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*, 2023. URL https://arxiv.org/abs/2306.00176.

Parameswary Rasiah. A framework for the systematic analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42(3):664–680, 2010. doi: 10.1016/j.pragma.2009.07.010.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21763–21813. Curran Associates, Inc., 2024. doi: 10.52202/079017-0685. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf.

Todd Rogers and Michael I. Norton. The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology: Applied*, 17(2):139–147, 2011. doi: 10.1037/a0023439.

Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6664–6679, 2023. doi: 10.18653/v1/2023.acl-long.368. URL https://aclanthology.org/2023.acl-long.368/.

Agam Shah, Siddhant Sukhani, Huzaifa Pardawala, Saketh Budideti, Riya Bhadani, Rudra Gopal, Siddhartha Somani, Rutwik Routu, Michael Galarnyk, Soungmin Lee, Arnav Hiray, Akshar Ravichandran, Eric Kim, Pranav Aluru, Joshua Zhang, Sebastian Jaskowski, Veer Guda, Meghaj Tarte, Liqin Ye, Spencer Gosden, Rachel Yuh, Sloka Chava, Sahasra Chava, Dylan Patrick Kelly, Aiden Chiang, Harsit Mittal, and Sudheer Chava. Words that unite the world: A unified framework for deciphering central bank communications. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2764.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5204–5233, 2024. doi: 10.18653/v1/2024.findings-emnlp.300. URL https://aclanthology.org/2024.findings-emnlp.300/.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. SemEval-2026 task 6: CLARITY – unmasking political question evasions. In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pp. 3704–3715, 2026. doi: 10.18653/v1/2026.semeval-1.449. URL https://aclanthology.org/2026.semeval-1.449/.

Petter Törnberg. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023. URL https://arxiv.org/abs/2304.06588.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xin-Yu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. TrustJudge: Inconsistencies of LLM-as-a-judge and how

to alleviate them. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ecb87120e1e9c68ad186385e146094f4-Abstract-Conference.html.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/.

RESEARCH TRANSPARENCY

The statements before the references describe AI use, ethics, and reproducibility. FOMCBENCH provides 6,462 professional human degree judgments and resolved majority references for all 2,154 pairs. These appendices document annotation reliability, source-text fidelity, evaluation inputs, cached baseline results, and reproducible scoring. Human-reference evaluations and model-reference diagnostics retain distinct interpretations. Uncertainty estimates condition on their stated sampling designs and do not capture new-rater, training-run, or provider variation. The labels assess answer responsiveness, not intent, honesty, or institutional trustworthiness.

Appendix roadmap. Appendix A details corpus construction, and Appendix B documents human annotation and reference provenance. Appendix C gives G/B evaluation details, open-weight model sources, supplemental scores, and reproduction procedures. Appendix D reports API and supervised baselines together with their human-reference, model-reference, and cross-domain analyses. Appendices E and F report model-label temporal patterns and configuration ablations; Appendices G–J document the dataset, illustrative cases, model prompt, and remaining reproduction procedures.

A CONSTRUCTION PIPELINE DETAILS

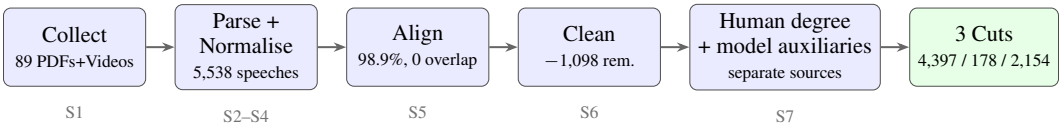


Figure 2: FOMCBENCH construction pipeline. The benchmark comprises 2,154 processed Q&A records with professional human degree judgments and separately sourced model-derived auxiliary fields. The 4,397-speech and 178-section representations are construction artifacts. Source-text limitations appear in Appendix A.1; annotation and aggregation are described in Appendix B.2.

Stage 1: Collection. We download 89 PDF transcripts and videos from the Federal Reserve website using automated downloads with resume capability. Meeting times are sourced from the Federal Reserve press calendar; no common start time is assumed.

Stage 2: PDF parsing. A custom parser extracts speaker-attributed turns from PDF transcripts, handling layout-specific challenges: continuation vs. paragraph indentation thresholds (with 0.5-point tolerance for floating-point edge cases), speaker name detection from capitalised headers, and special cases such as the first meeting (2011-04-27) where the PDF used generic "QUESTION" labels requiring manual speaker mappings. The parser also handles PDF typos (e.g., "CHAIMRAN POWELL"). This produces 5,538 raw speeches across 89 meetings.

Stage 3: Organisation extraction. Each reporter's media organisation is identified using Claude Opus 4.6, with the full parsed CSV sent per meeting so the LLM can extract self-introductions ("Steve Liesman, CNBC"). Cross-meeting gap filling propagates identified organisations to meetings where a reporter did not self-introduce, with provenance tracking. Canonical normalisation maps 73 raw organisation names to 42 unique outlets (e.g., "Bloomberg News" / "Bloomberg Television" → "Bloomberg"). Organisations are identified for 1,556 of 1,562 reporter–meeting records (99.6%); 18 reporters changed outlets over the 14-year span. These records count a reporter separately at each meeting; the final Q&A dataset contains 149 unique normalised reporter names.

Stage 4: Name normalisation. 27 rules correct nicknames (Mike → Michael), spelling errors (Applebaum → Appelbaum), PDF typos (Michelle Smth → Michelle Smith), and casing issues (Mcgrane → McGrane), yielding 1,284 corrections across all outputs.

Stage 5: Timestamp alignment. Official transcripts lack timestamps; ASR transcripts have timestamps but unreliable speaker diarisation. We align the two via greedy fuzzy text matching (RapidFuzz token_sort_ratio) between official transcript segments and ASR segments, with uncapped matching for opening remarks and final speeches to prevent truncation of long turns. Boundary refinement detects reporter self-introductions and skips moderator transitions. Result: 4,348 of 4,397 speeches aligned (98.9%), zero timestamp overlaps, monotonic ordering verified.

Stage 6: Quality control. LLM-assisted cleaning (Claude Sonnet 4.5 with reasoning) flags non-substantive speeches for removal: brief acknowledgements ("Thank you"), inaudible questions, and moderator turns. All 89 meetings maintain strict question–answer count parity after cleaning. 1,098 speeches removed, 43 merged (interrupted answers concatenated). We prefer removal over merging to maintain audio/video timestamp contiguity.

Stage 7: Human degree annotation and model-derived fields. Three finance professionals independently annotated every Q&A record, blind to model predictions. Professional review and majority aggregation produce complete binary and three-class degree references (Appendix B.2). Mechanism, tactic, and auxiliary score fields are produced by Opus 4.6 under Condition D and remain distinct from the human judgments.

Model-consensus labels also support split construction and a label-source audit. The three-family Condition D panel comprises Opus 4.6, GPT-5.2, and Gemini 3 Pro with meeting-level prompting and reasoning. Its majority rule uses Opus to resolve five model three-way disagreements; Dev selects model-unanimous cases. These choices shape the evaluation populations and do not make the panel's shared models/prompts independent annotators. The six-configuration prompt panel and 15-configuration consensus ablation are separate diagnostic designs. Training-label variants for the supervised baselines are documented in Table 52.

A.1 Known text-version defects

The benchmark preserves the frozen processed records to which its annotations and cached predictions are linked. A source comparison identifies two substantive omissions in Dev: `M20241107_P011` retains a 113-character answer but omits the continuation after brief reporter acknowledgements (official 7 November 2024 transcript, physical pp. 12–13); `M20250319_P025` omits 93 characters from its question, including the SEP statement about 17 of 19 participants (19 March 2025 transcript, p. 23). Independent reconstruction from those official pages matches the separately processed counterpart fields of 2,627 and 569 characters, respectively. This check does not assign labels to the reconstructed text.

The packaged `ERRATA.md` and `text_difference_crosswalk.json` provide official source links, page locators, field hashes and input associations. The broader comparison contains 22 differing fields across 13 meetings: 11 Train, five Dev, one Test-Std and five Test-Challenge. The other 20 changes delete editorial corrections or footnote markers; their semantic appropriateness is not established. The two confirmed omissions are not directly scored test targets, but their meetings contain six Std and three Challenge targets whose B inputs include those Dev passages. Across all 13 meetings, the corresponding counts are 26 and 36. These are input-dependency counts, not counts of wrong predictions or labels. The known-nine prior-exposure sensitivity does not address these text differences.

For the cached API G/B runs, predictions are linked to records by pair ID only; their historical wire-input text version is unverified. It is therefore unknown whether those calls contained the two confirmed omissions or the other editorial variants, including the affected meeting passages available to B.

Released votes and predictions are bound to the frozen processed records. Retained assignment files and G/B messages identify this field version, but per-item presentation is not reconstructable for every annotation. Context availability or an opening event does not establish what a professional read. The experiments use the frozen text; invariance of labels or scores under correction is unestablished. Correcting the records requires reassessing their labels and dependent meeting inputs.

B Human Annotation and Evaluation Scope

B.1 Evaluation scope

Table 6 summarises the corpus. The primary binary task (T1b) and secondary three-class task (T1) use human degree references. Figure 3 distinguishes their evidence from the narrower model-reference diagnostics.

Table 6: FOMCBENCH dataset summary. Degree counts use current human majorities; percentages use all 2,154 pairs. Auxiliary labels remain model-derived.

Statistic	Value	Coverage
Meetings / Q&A pairs	89 / 2,154	Apr 2011–Dec 2025
Pairs per meeting	24.2 ± 5.7	range: 10–36
Reporters / organisations	149 / 42	
Bernanke / Yellen / Powell	266 / 296 / 1,592	12 / 16 / 61 meetings
Questions: tokens	68.6 ± 36.6	median 66
Answers: tokens	208.9 ± 124.5	median 189.5
Professional degree judgments	6,462	three per pair
Direct / non-direct (binary)	1,190 / 964	all references resolved
Direct / intermediate / fully evasive	1,190 / 806 / 158	55.2% / 37.4% / 7.3%
Resolved three-class references	2,154	all four splits

Text length uses whitespace tokens; dispersions are sample standard deviations.

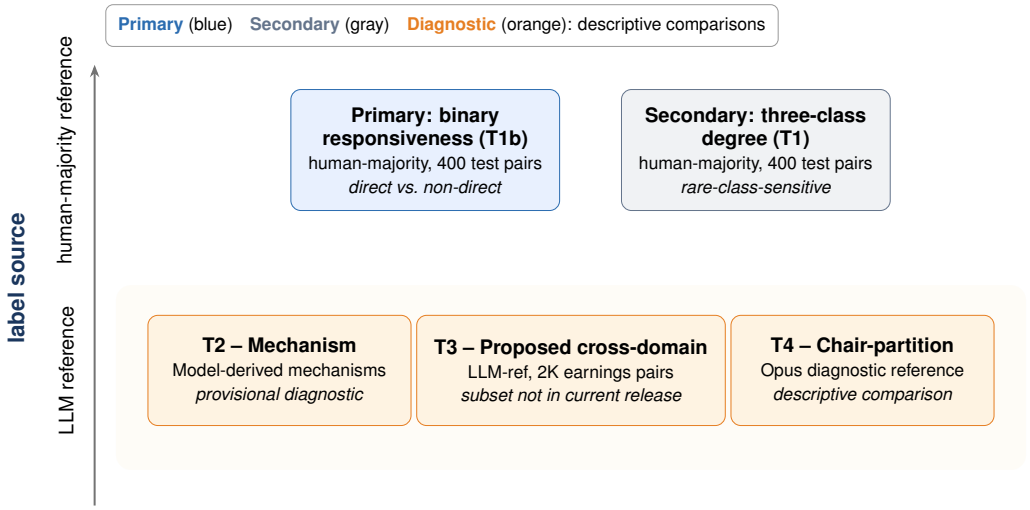


Figure 3: **Task claim-scopes: what each task can and cannot support.** T1b is primary and T1 secondary, with human degree references on all 2,154 pairs. Test-Std and Test-Challenge provide the fixed human-reference test populations. Train/Dev selection and tuning exposure limit broader evaluation to a corpus audit. T2/T3/T4 use model-derived or provisional references and support narrower diagnostic analyses.

Supplemental task inventory. Table 7 retains the core degree tasks alongside auxiliary diagnostics and the proposed T3 task. The model-derived T2/T4 references do not provide human validation of mechanisms or unseen-chair generalisation. T3 is unimplemented and its proposed earnings-call subset is absent; it is not part of the core benchmark evaluation. API B/D input definitions appear in Appendix D.1.

B.2 PROFESSIONAL HUMAN ANNOTATION

Collection and rubric. Three finance professionals from the same institution, including one author, independently annotated all 2,154 pairs. Their expertise covered financial analysis, quantitative research, and signal development. They used separate annotation access without model predictions or one another's responses; the collecting author confirms this independent, model-blind human procedure. Every record has three retained judgments, yielding 6,462 degree labels. The rubric assesses whether the answer supplies the requested information, including multipart questions, numerical requests, substantive qualifications, and partial substitutes; it does not assess honesty or intent.

Table 7: FOMCBENCH tasks and diagnostics, including the proposed T3 task. Inputs, outputs, metrics, and reference sources are specified below.

Task	Description	Input	Output	Metric	Labels
T1	3-way (secondary)	Q+A pair	{direct, interm., evasive}	Macro-F1	Human
T1b	Binary (primary)	Q+A pair	{direct, non-direct}	Macro-F1	Human
T2	Mechanism identification	Non-direct Q+A	{vague., non-seq., omis., restat.}	W-F1	LLM
T3	Proposed cross-domain task[†]	Earnings call Q+A	3-way label	Macro-F1	LLM
T4	Chair-stratified diagnostic	Chair-partition Q+A	3-way label	Macro-F1	Cond. D

[†]T3 is not implemented. Its 1K Chinese + 1K S&P subset is absent pending redistribution permission.
Labels: Human = 3-annotator majority vote; LLM = model-derived; Cond. D = Opus 4.6 reference.
The prediction target is one exchange; B/D also receive the full unlabeled meeting context (Appendix D.1).
T4 partitions cached predictions by chair against Condition D; it does not retrain models with a chair excluded.

The documented instructions provide the target question/answer and ordered meeting Q&A context. The FC1 rule permits an immediately following response to complete the target only when the chair requests a reminder, the journalist explicitly confirms or restates the question, and the following reply belongs to that continuation. Unrelated later answers are excluded, and forgetting alone does not determine the label. The retained model prompt differs from FC1 in its treatment of repetition requests (Appendix C).

Reference aggregation and agreement. Final references were produced through professional review and majority vote. For binary responsiveness, each vote is collapsed to direct versus non-direct before aggregation; three-class references require at least two identical votes. The complete reference set contains 1,190 direct, 806 intermediate, and 158 fully evasive labels, or 1,190 direct and 964 non-direct labels for binary evaluation. Every pair has a resolved reference for both tasks.

Agreement statistics describe the versioned retained ratings before that professional review. Three-class Fleiss' κ is 0.7911 (95% interval $[0.7717, 0.8101]$), and binary κ is 0.8264 ($[0.8054, 0.8461]$). The rating snapshot used for agreement and the reviewed references used for scoring are separately identified in the machine-readable release metadata.

Intervals resample the 89 meetings as whole units 10,000 times with replacement (seed 20260926), carrying all eligible items/votes and retaining repeated-meeting multiplicity. Point estimates weight items equally, and sampled pair counts may vary. Intervals are the linearly interpolated 2.5th and 97.5th percentiles; they are descriptive, pointwise, and unadjusted. This design captures empirical meeting variation for these three professionals, not new-rater uncertainty or variation in presentation protocols.

Record format and label-source audit. Dev and test ratings are released as anonymous triplets in degree order; Train records retain separate annotation slots. This format preserves majority references and Fleiss' κ; individual-rater performance is not reported.

Model-consensus Train labels differ from human majorities on 334/1,554 binary references and 408/1,554 three-class references; Dev differs on 11/200 under either task. These are comparisons of label sources on the same items, not held-out predictive accuracy. Split selection and supervised-baseline training provenance remain relevant to interpretation.

Test-set agreement. The retained ratings on the 400 test pairs are unanimous for 292 pairs. All 108 disagreements concern adjacent degrees: 89 direct/intermediate and 19 intermediate/fully_evasive.

Recorded three-class Fleiss' κ is 0.689 on Test-Std and 0.654 on Test-Challenge, with 200 pairs each. The three-family Condition-D model panel has Fleiss' $\kappa = 0.516$ on 2,154 pairs; the separate six-configuration B/D panel has $\kappa = 0.524$ on 2,153 complete pairs. These coefficients describe different panels and weighting choices.

On the binary target, observed pairwise agreement in each split is 87.0% for Test-Std and 83.3% for Test-Challenge, with Fleiss' $\kappa = 0.736$ and 0.594.

Test-Std construction. The retained construction script specifies sampling after excluding Test-Challenge, stratified by Condition-D model-consensus class and chair, with seed 42. It allocates

classes proportionally with a minimum of five, then chairs within each class proportionally with a minimum of one, caps draws at available stratum sizes, and randomly trims any overshoot. For the released Test-Std IDs, the retained model-label counts (direct/intermediate/fully evasive) are 16/8/1 for Bernanke, 15/12/1 for Yellen, and 81/61/5 for Powell: 112/81/7 overall. The human-reference composition, 88/102/10, is an annotation outcome rather than the sampling target. `data/splits/CONSTRUCTION.md` records the rule, source bindings, and released-membership tabulation. An isolated replay of the Test-Std sampling function with the bound inputs, seed 42, and its implemented class iteration order reproduces all 200 released IDs. Scoring uses the supplied fixed-ID membership.

Test-Challenge selection. The retained selection manifest reproduces all 200 challenge IDs: 100 disagreement cases, 50 model-unanimous cases, 30 Eva–Opus disagreement cases, and 20 randomly selected remaining cases (seed 42). The selection panel uses Opus 4.6, GPT-5.2, and Gemini 3 Pro under D; the additional contrast uses unadapted Eva-4B and Opus 4.6 B. Opus 4.7, GPT-5.4, and Gemini 3.1 Pro did not participate. This selection makes difficulty conditional on the selecting systems; it does not force them to be wrong against the later human-majority labels. Comparisons with newer systems inherit this selection asymmetry.

C G/B EVALUATION DETAILS

The main G/B comparison contains eight open-weight systems and three cached API systems: Opus 4.6, GPT-5.2, and Gemini 3 Pro. Each uses the same final human-majority references on 200 Test-Std and 200 Test-Challenge pairs, fixed target classes, and invalid-as-error scoring. Binary direct/non-direct Macro-F1 is primary; three-class degree is secondary. The common 10,000 whole-meeting resampling procedure is described below. Source recovery and native metadata remain specific to each group; common scoring does not establish identical input messages or runtime settings. This section details the open-weight group, with API protocol and cache limitations in Appendix D.

C.1 MODEL INVENTORY AND OPEN-WEIGHT INPUTS

For the open-weight group, G denotes the domain-neutral financial-Q&A benchmark prompt on the target record; B denotes its meeting-level counterpart over ordered released records, including later exchanges, without reference labels. These are processed inputs subject to the source-text limits in Appendix A.1. Both use the degree/mechanism/tactic framework, the same domain-neutral financial-Q&A wording, and fixed demonstrations. Degree scoring does not validate the requested auxiliary fields. G/B also change batching and message wrappers, so their contrast is a protocol comparison rather than an isolated causal effect of context. Full-meeting input can include exchanges assigned to other splits; evaluation concerns held-out exchanges from shared meetings.

The model prompt judges only the answered parts after a request for repetition, whereas the human FC1 rule credits a confirmed immediate continuation against the material question. This rubric difference is a possible source of disagreement, not an identified cause, and must be considered alongside the source-text limitations.

The source-bound fresh-run profiles record temperature 0.1, top-$p = 1$, sampling seed 222026, an 8,192-token G output budget, and a B budget of $\max(8192, 512n + 1024)$ for n target exchanges. Recorded context caps are 32,768 tokens for Qwen and DeepSeek and 65,536 for the other six systems. These are execution caps, not model-capacity claims. GPT-OSS permits native low-effort reasoning; the other recorded profiles disable thinking or select no reasoning effort. Nemotron uses FP8 weights/quantized linear activations with BF16 other tensors; DeepSeek combines MXFP4 routed experts, FP8 dense/Engram components, and BF16 embeddings. Retained or reused outputs retain their own native provenance rather than inheriting these settings by model name.

C.2 ORIGINAL-OUTPUT RECOVERY AND ADAPTIVE SENSITIVITY

For the eight open-weight models, the headline view, `required_field_degree_v1`, was selected before inspecting the numerical results. It recovers the explicitly qualified degree from the original stored output. It uses a degree only when the source audit marks that field qualified and its value belongs to the three-class taxonomy. Missing auxiliary mechanism, tactic, score, rationale,

Table 8: Exact checkpoints in the eight-model open-weight panel. Repository names link to immutable model cards; revision prefixes abbreviate the full revisions in those links.

Pinned model card / exact repository	Revision	Published license
`Qwen/Qwen3.8-Flash-Next-FP8`	`236dfdf285`	Qwen Community License 1.0
`google/gemma-4-31B-it`	`842da3794e`	Apache 2.0
`openai/gpt-oss-120b`	`b5c939de8f`	Apache 2.0
`mistralai/Mistral-Small-4-119B-2603`	`a11f36bebf`	Apache 2.0
`nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-FP8`	`744b1880a3`	NVIDIA Nemotron Open Model License
`swiss-ai/Apertus-70B-Instruct-2509`	`0f2767662a`	Apache 2.0
`zai-org/GLM-4.7-Flash`	`7dd20894a6`	MIT
`deepseek-ai/DeepSeek-V4.1-Flash`	`dba1be0a40`	MIT

We use *open-weight* to describe published weights. Qwen Community License 1.0 and the NVIDIA Nemotron Open Model License have their own terms; the Apertus card also includes its LLM Acceptable Use Policy (1.0, 1 September 2025). License names reproduce publisher declarations; this table makes no assessment of legal compliance. Public card and revision metadata were checked on 26 September 2026; public availability does not establish which tensor bytes were loaded in a run.

Table 9: Developers and advertised parameter sizes for the eleven G/B models. Eva-4B, the separate native pair-input baseline, is described as a 4B model in EVASIONBENCH (Ma et al., 2026). B denotes one billion parameters. Linked publisher sources define the counting conventions; component counts are not summed into inferred whole-model totals. N/V means no parameter count was verified in the inspected official sources.

Model	Developer	Advertised components (B)	Active (B)
Qwen3.8 Flash Next	Qwen Team / Alibaba	125 language model; 51 n-gram embeddings; 4 MTP	6
Gemma 4 31B	Google DeepMind	30.7 (dense)	—
gpt-oss-120b	OpenAI	117	5.1
Mistral Small 4	Mistral AI	119	6.5
Nemotron 3 Super	NVIDIA Corporation	120	12
Apertus 70B	Swiss AI Initiative (ETH Zurich, EPFL, CSCS)	70 (dense)	—
GLM-4.7-Flash	Z.ai (ZhipuAI)	30	3
DeepSeek-V4.1-Flash	DeepSeek	552 backbone; 196 Engram memory	8 prefill; 16 decode
Claude Opus 4.6	Anthropic	N/V	N/V
GPT-5.2	OpenAI	N/V	N/V
Gemini 3 Pro	Google DeepMind	N/V	N/V

Qwen's 6B active count refers to its language-model component; its listed embedding/MTP components and DeepSeek's Engram memory remain separate. Their overview descriptions do not quantify all vision components. Gemma's card separately lists approximately 0.55B vision-encoder parameters without establishing whether they are included in 30.7B. Mistral's pinned card reports 6.5B active; its launch article reports 6B, or 8B including embedding/output layers, an unresolved publisher-convention difference. Dense models have no separately reported active count. These declarations are not normalized compute measures; N/V is bounded source non-verification, not a claim of universal nondisclosure.

or evidence fields do not invalidate an independently qualified degree. Missing, conflicting, or unqualified degree values remain invalid. An original three-class conflict is not resolved merely

because both candidates would collapse to binary non-direct. No human-reference label or model score is used to select a record or infer a degree.

The separately declared adaptive view first takes the original qualified degree. Only when that degree is unqualified does it use the degree from the already admitted, chronologically selected first wholly qualifying record. A later differing degree does not replace a qualified original, and fields from different records are not combined. If neither source qualifies, the prediction remains invalid. This adaptive projection is a sensitivity analysis, not the original-output headline. The source exports also preserve their original degree, strict-object, and earlier recovery views where available; an unavailable GLM earlier-recovery view is not manufactured.

Every intended test pair remains in the scoring denominator, with missing or invalid predictions counted as errors under fixed target classes. Within the open-weight group, Apertus Test-Std B has 199 valid original-output degrees and one invalid prediction out of 200; every other open-weight model/arm/split cell has 200 valid predictions. The adaptive fallback changes that Apertus cell and remains reported separately in Table 13; greater output coverage need not raise Macro-F1.

C.3 Uncertainty, exposure, and native-comparability limits

Per-arm intervals and paired B-minus-G intervals use the same 10,000 whole-meeting PCG64 draws (seed 20260924). Every sampled meeting contributes all eligible pairs and retains its multiplicity when sampled repeatedly; G/B and the source views share draw indices. Fixed classes and invalid-as-error scoring are retained in every replicate. The reported 95% intervals are the linearly interpolated 2.5th and 97.5th percentiles. Paired intervals come from paired score differences, not subtraction of marginal interval endpoints. They are exploratory, pointwise, and unadjusted: no simultaneous coverage, new p-value family, borrowed API-panel Holm correction, or confirmatory winner claim follows. The G/B comparison seed is 20260924 with whole meetings as resampling units; annotation-κ intervals use seed 20260926 with whole meetings, while the additional API pair-level calculations identify seed 42 in Appendix J.

For the open-weight sensitivity analysis, the known-nine mask identifies the four Test-Std and five Test-Challenge items that overlap a 71-pair development census. All nine come from the 17 September 2025 meeting. The extraction procedure was developed with that census's output failures visible; this is extraction-development exposure, not evidence of model-weight training. Their exclusion gives denominators of 196 and 195, respectively, with separately recomputed pointwise intervals (Tables 11 and 12). The known-exposed-only views contain four or five pairs and report descriptive point estimates without intervals. Excluding this known set does not establish that the remaining examples were unseen. Test-Challenge remains disagreement-enriched, and both test sets share meetings with Train.

Native metadata matching for the open-weight group is reported without filtering the principal test population. Qwen has 400 unknown G/B comparisons. Gemma, GPT-OSS, Mistral, Nemotron, and GLM each have 391 matched and nine unknown comparisons; Apertus and DeepSeek each have 400 matched. Unknown does not mean equivalent. DeepSeek's matching is confined to its admitted single collection with a common 32,768-token context cap, exact model/profile settings, and no reused outputs; it does not establish cross-job date equivalence. Some per-request inference and publisher-prefix dates are unrecorded, and export timestamps are not substituted for them. Metadata matching does not make the G/B contrast causal. The intervals condition on fixed predictions and references and omit annotation uncertainty, repeated-run variation, and temporal dependence between meetings.

C.4 Supplemental scores and protocol differences

Table 5 in the main paper gives secondary three-class scores for both model groups; Table 10 reports available class-level breakdowns. Table 14 contains the full-split G/B contrasts for all eleven systems. The known-nine-excluded contrasts in Table 15, the exposure sensitivities, and the adaptive-output table apply only to the eight open-weight systems. Their mask and recovery policy are not transferred to the cached API group.

Table 10: Per-class F1 (%) for the two source panels under their stated degree-output policies. D: direct; I: intermediate; E: fully evasive.

		Test-Std ($n = 200$)			Test-Challenge ($n = 200$)		
Model	Arm	D	I	E	D	I	E
Open-weight models							
Qwen3.8-Flash-Next	G	54.8	75.5	63.2	37.1	83.3	90.5
	B	83.8	84.8	78.3	63.0	82.6	88.9
Gemma-4-31B	G	70.1	75.0	45.2	54.5	85.3	86.3
	B	79.0	82.7	78.3	62.3	83.3	89.8
GPT-OSS-120B	G	27.2	68.6	46.2	25.4	77.5	58.3
	B	67.8	69.9	47.1	52.5	65.8	60.0
Mistral-Small-4	G	2.2	68.3	57.1	0.0	79.9	77.3
	B	18.0	66.4	55.2	20.9	78.2	69.4
Nemotron-3-Super	G	12.8	67.1	40.0	7.3	79.7	71.8
	B	65.6	64.4	33.3	43.2	67.0	28.6
Apertus-70B	G	0.0	68.2	46.2	0.0	79.1	66.7
	B	14.0	65.2	10.0	3.2	72.2	35.9
GLM-4.7-Flash	G	10.6	66.4	40.0	13.1	78.6	65.3
	B	16.5	68.3	37.5	22.9	77.4	48.5
DeepSeek-V4.1-Flash	G	54.1	74.7	57.1	37.1	82.5	86.4
	B	79.3	80.9	66.7	68.7	84.6	75.0
Closed-source/API models							
Opus 4.6[‡]	G	75.2	83.6	66.7	61.0	86.6	87.8
	B	82.8	78.7	57.1	62.2	73.0	68.6
GPT-5.2[‡]	G	78.4	85.5	88.9	60.8	86.5	82.1
	B	83.8	84.3	82.4	68.9	82.9	90.9
Gemini 3 Pro[‡]	G	87.2	83.2	63.6	70.1	80.2	90.9
	B	78.9	76.2	76.2	59.6	71.4	87.0

The fixed class set is retained even when a model predicts no examples of a class. Invalid outputs contribute false negatives rather than being assigned to a class. [‡] These systems' Condition-D outputs informed Test-Std strata and Test-Challenge selection.

C.5 Prediction provenance and reproduction

The reviewer package stores predictions under `predictions/open_weight_panel/`, partitioned by model, source view, arm, and test split. Each JSONL row has its exact pair identifier and a degree or explicit invalid value. `analysis/open_weight_panel/expected_scores.json` contains 1,128 score records and 376 paired-difference records. `cohort_membership.jsonl` and `protocol.json` in the same directory bind fixed populations, exposure flags, and the scoring plan; `manifest.json` binds the prediction files.

The same directory contains `models.json`, with the eight exact model IDs, full revisions, recorded precision, and pinned model-card links; `benchmark_requests.jsonl`, preserving canonical messages before model-specific native wrapping; and `native_conditions.json`, preserving the recorded native profiles. Per-model/arm files under `analysis/open_weight_panel/provenance/` retain source-message hashes and per-pair native-eligibility flags. Exact per-output operational bindings remain internal; the fresh-collection profile must not be assigned to every reused or recovery output by implication. The package reproduces scores from fixed degree projections and documents native metadata limits; it does not reconstruct every original inference call or validate auxiliary labels.

From the reviewer-package root, reproduce the stored point estimates with:

Table 11: Open-weight binary Macro-F1 (%) after excluding the nine known previously exposed test pairs (four Std; five Challenge).

	Test-Std ($n = 196$)		Test-Challenge ($n = 195$)	
Model	G	B	G	B
Qwen3.8-Flash-Next	67.4 [60.7, 73.6]	86.9 [81.3, 91.7]	62.1 [52.8, 71.5]	74.6 [68.4, 80.3]
Gemma-4-31B	77.5 [71.4, 83.1]	82.9 [77.0, 88.1]	72.2 [62.8, 80.7]	74.6 [67.7, 81.1]
GPT-OSS-120B	51.0 [43.9, 58.1]	72.0 [64.4, 78.7]	55.9 [48.6, 62.9]	63.4 [56.0, 70.5]
Mistral-Small-4	37.0 [33.8, 40.3]	45.3 [39.5, 51.5]	42.3 [40.3, 44.1]	52.6 [45.8, 59.5]
Nemotron-3-Super	42.9 [37.8, 48.3]	68.7 [61.9, 75.0]	46.3 [41.6, 52.1]	56.0 [48.9, 63.1]
Apertus-70B	35.7 [33.2, 38.1]	42.7 [37.3, 48.5]	42.3 [40.3, 44.1]	42.7 [39.6, 46.8]
GLM-4.7-Flash	41.5 [36.2, 47.6]	44.8 [38.9, 50.9]	48.8 [42.7, 55.8]	53.3 [46.2, 60.0]
DeepSeek-V4.1-Flash	67.1 [60.8, 73.0]	82.1 [76.7, 87.0]	62.1 [53.5, 70.6]	79.6 [72.6, 85.8]

This is a declared sensitivity; the remaining pairs are not certified unseen. Every retained target contributes to the score, including invalid degrees. The scoring delivery preserves the exposed-only point scores and separate paired intervals for this population.

Table 12: Open-weight three-class Macro-F1 (%) after excluding the nine known previously exposed test pairs (four Std; five Challenge).

	Test-Std ($n = 196$)		Test-Challenge ($n = 195$)	
Model	G	B	G	B
Qwen3.8-Flash-Next	64.5 [53.4, 72.3]	82.8 [72.9, 89.8]	70.3 [62.3, 77.8]	78.0 [72.1, 83.3]
Gemma-4-31B	63.5 [54.1, 72.0]	80.1 [70.3, 87.4]	75.4 [67.5, 82.3]	78.5 [71.9, 84.5]
GPT-OSS-120B	47.3 [36.7, 56.4]	63.0 [49.6, 73.3]	54.4 [45.6, 61.9]	58.8 [49.6, 67.2]
Mistral-Small-4	42.4 [32.3, 49.8]	46.4 [37.3, 54.1]	52.0 [45.5, 57.2]	56.4 [48.0, 63.8]
Nemotron-3-Super	39.9 [27.8, 50.7]	55.0 [41.7, 66.7]	52.4 [44.5, 59.1]	44.1 [34.8, 54.2]
Apertus-70B	38.0 [23.0, 48.3]	29.6 [23.4, 37.6]	47.9 [41.1, 53.7]	35.9 [28.2, 43.7]
GLM-4.7-Flash	38.9 [27.9, 48.5]	40.7 [27.5, 50.5]	51.9 [43.8, 59.0]	48.5 [39.1, 56.8]
DeepSeek-V4.1-Flash	62.0 [50.5, 71.2]	75.4 [63.7, 83.9]	68.6 [60.9, 75.6]	76.0 [66.8, 83.5]

This is a declared sensitivity; the remaining pairs are not certified unseen. Every retained target contributes to the score, including invalid degrees. The scoring delivery preserves the exposed-only point scores and separate paired intervals for this population.

```
python3 -B code/reproduce_open_weight_panel.py \
  --output-dir ../open-panel-points
```

Adding `--bootstrap` uses NumPy to regenerate the shared 10,000 whole-meeting draws, all 752 per-arm intervals, and all 376 paired intervals. The output directory must be new and outside the reviewer package. The routine retains all available declared views, with original-output required-field

Table 13: Declared adaptive-output sensitivity for Apertus Std/B. The adaptive rule uses a qualified whole-record fallback only when the original degree is unresolved.

Output view	Binary F1 (%)	Three-class F1 (%)	Invalid
Original required field	42.8 [37.4, 48.4]	29.7 [23.5, 37.9]	1
Adaptive fallback	42.7 [37.3, 48.3]	29.7 [23.4, 37.8]	0

Both rows retain all 200 intended targets. Brackets give pointwise 95% whole-meeting intervals. In this frozen snapshot, adaptive and original required-field predictions coincide in every other fixed-test cell. The adaptive view is separate from the original-output headline evaluation.

Table 14: Observed B minus G protocol differences in Macro-F1 percentage points, with pointwise 95% whole-meeting intervals.

	Binary (primary)		Three-class (secondary)	
Model	Std	Challenge	Std	Challenge
Open-weight models				
Qwen3.8-Flash-Next	+18.8 [+11.7, +26.1]	+12.7 [+3.5, +21.7]	+17.8 [+7.5, +27.8]	+7.8 [+0.5, +14.9]
Gemma-4-31B	+5.5 [-0.7, +11.6]	+2.5 [-6.1, +11.4]	+16.6 [+6.9, +25.8]	+3.1 [-3.6, +10.3]
GPT-OSS-120B	+20.6 [+11.6, +29.6]	+7.6 [-1.0, +16.1]	+14.3 [+4.1, +24.3]	+5.7 [-4.0, +15.4]
Mistral-Small-4	+8.2 [+2.6, +14.4]	+10.1 [+3.1, +17.2]	+4.0 [-6.2, +15.3]	+3.8 [-5.2, +12.4]
Nemotron-3-Super	+24.9 [+17.4, +32.6]	+10.2 [+1.3, +19.1]	+14.5 [+1.6, +25.2]	-6.7 [-17.0, +4.4]
Apertus-70B	+6.9 [+1.5, +12.6]	+0.4 [-1.9, +3.8]	-8.4 [-18.9, +3.0]	-11.5 [-19.4, -3.3]
GLM-4.7-Flash	+3.3 [-3.1, +9.5]	+4.5 [-4.5, +12.9]	+1.7 [-13.7, +16.1]	-2.8 [-12.3, +6.6]
DeepSeek-V4.1-Flash	+15.5 [+8.1, +22.9]	+17.3 [+8.2, +26.3]	+13.6 [+3.5, +23.2]	+7.4 [-2.1, +16.4]
Closed-source/API models				
Opus 4.6‡	+2.2 [-5.2, +9.5]	-4.0 [-14.3, +7.2]	-2.3 [-12.6, +5.3]	-10.5 [-21.3, -0.6]
GPT-5.2‡	+2.5 [-3.6, +8.6]	+2.0 [-8.0, +12.0]	-0.8 [-8.3, +5.0]	+4.4 [-4.6, +13.6]
Gemini 3 Pro‡	-8.0 [-14.2, -1.9]	-8.8 [-16.6, -0.9]	-0.9 [-8.9, +7.9]	-7.8 [-15.1, -0.7]

Intervals use 10,000 paired whole-meeting draws and retain 200 targets per split, including invalid predictions. They are exploratory, pointwise and unadjusted; no confirmatory or familywise claim follows. G/B jointly changes context, batching and wrappers. Full-test native comparability is unknown for all 400 Qwen pairs; Gemma, GPT-OSS, Mistral, Nemotron and GLM each have 391 matched and nine unknown pairs; Apertus and DeepSeek each have 400 matched. All 400 G/B comparisons per API model have unknown native equivalence. Shared scoring does not establish identical prompts or runtime; no matched-subset scores are reported. ‡ These systems' Condition-D outputs informed Test-Std strata and Test-Challenge selection.

recovery as the headline and adaptive recovery separate. Reproduction reads fixed predictions and references; it does not run inference, re-extract raw responses, or train a model.

Table 15: Open-weight B minus G differences after known-nine exclusion in Macro-F1 percentage points, with pointwise 95% whole-meeting intervals.

	Binary (primary)		Three-class (secondary)	
Model	Std	Challenge	Std	Challenge
Qwen3.8-Flash-Next	+19.5 [+12.3, +26.7]	+12.4 [+3.0, +21.9]	+18.2 [+7.8, +28.7]	+7.6 [+0.1, +15.2]
Gemma-4-31B	+5.4 [-0.9, +11.7]	+2.4 [-6.1, +11.4]	+16.6 [+6.8, +26.3]	+3.1 [-3.5, +10.5]
GPT-OSS-120B	+21.0 [+11.8, +29.9]	+7.5 [-1.3, +16.0]	+15.8 [+5.2, +25.4]	+4.5 [-4.9, +14.3]
Mistral-Small-4	+8.3 [+2.7, +14.7]	+10.3 [+3.0, +17.2]	+4.0 [-6.6, +15.5]	+4.4 [-5.0, +13.5]
Nemotron-3-Super	+25.8 [+18.1, +33.4]	+9.7 [+0.6, +18.6]	+15.1 [+2.8, +25.7]	-8.3 [-18.6, +3.0]
Apertus-70B	+6.9 [+1.7, +12.7]	+0.4 [-1.9, +4.0]	-8.4 [-18.9, +2.9]	-12.0 [-20.2, -3.3]
GLM-4.7-Flash	+3.3 [-3.3, +9.6]	+4.6 [-4.5, +13.0]	+1.8 [-14.1, +15.9]	-3.3 [-13.6, +6.4]
DeepSeek-V4.1-Flash	+15.0 [+7.7, +22.6]	+17.4 [+8.1, +26.6]	+13.3 [+3.0, +22.9]	+7.3 [-2.4, +16.4]

Intervals use 10,000 paired whole-meeting draws and retain 196 Std and 195 Challenge targets, including invalid predictions. They are exploratory, pointwise and unadjusted; no confirmatory or familywise claim follows. G/B jointly changes context, batching and wrappers. Full-test native comparability is unknown for all 400 Qwen pairs; Gemma, GPT-OSS, Mistral, Nemotron and GLM each have 391 matched and nine unknown pairs; Apertus and DeepSeek each have 400 matched. No matched-subset scores are reported.

D API AND SUPERVISED BASELINES

The main G/B comparison includes cached Opus 4.6, GPT-5.2, and Gemini 3 Pro predictions under the common human-reference scoring rules. Additional comparisons here cover the B-only Opus 4.7, GPT-5.4, and Gemini 3.1 Pro caches, degree-only A and other prompt conditions, and supervised baselines. Their model versions, inputs, training-label sources, and uncertainty procedures remain explicit. The model-reference diagnostics use separately identified references and are not human-reference performance estimates.

D.1 API INPUT PROTOCOLS

Configuration letters identify cached runs. In compact labels, unversioned Opus, GPT, and Gemini denote Opus 4.6, GPT-5.2, and Gemini 3 Pro; other versions are explicit:

- **Pair-only, degree-only (A):** the target Q&A with the Rasiah degree task.
- **Pair-only, three-axis (G; API framework prompt):** the target Q&A with the project's full degree, mechanism, and tactic framework. This cached API G is distinct from degree-only A and from the open-weight group's domain-neutral G prompt.
- **Full-meeting, three-axis (B):** the ordered released meeting Q&A records with degree, mechanism, and tactic instructions.
- **Reasoning-enabled variants (C and D):** the corresponding pair-only and full-meeting configurations with reasoning enabled.

G was chosen as the API pair-only arm to retain the full three-axis framework in both arms, rather than by selecting the pair-only condition with the highest score. A remains a separate degree-only comparison. The exact historical API G prompt text and per-pair builder are not included in the release; the released authored prompt covers B/D meeting inputs.

For the three main API systems, degree predictions are taken from the frozen exported labels, without a new raw-output extraction or auxiliary-field eligibility filter. The main comparison standardizes the human references, complete test denominators, released scorer, and 10,000 whole-meeting PCG64 draws (seed 20260924) for per-arm and paired intervals. Exact cached API wire-input text and

native G/B equivalence are not established; nominal G/B labels do not certify prompt or runtime identity with the open-weight panel. All intervals remain exploratory, pointwise, and unadjusted. The open-weight known-nine mask is not applied to this cached API regime.

The source G configurations record the aliases `claude-opus-4.6`, `gpt-5.2`, and `gemini-3-pro`. These are configured aliases, not dated model snapshots. Query dates and dated snapshot identifiers are not recorded in those configurations, and prediction-file modification times are not treated as inference dates. The prediction CSVs associate outputs with benchmark records by pair ID, without per-pair input text or input-text hashes. The provider documentation linked in Table 9 does not authenticate a historical query snapshot.

The three API families also supplied the corpus-construction model labels, so all 400 test exchanges had been submitted to those providers before this comparison. Whether the providers retained those requests or used them for training is unknown. This prior provider submission is distinct from the open-weight extraction-development mask and does not establish identical historical input strings.

Full-meeting inputs include later exchanges and text assigned to other splits, with speaker/date metadata and pair IDs, but no reference labels. This is offline full-meeting classification. Comparing A with B changes both input access and instructions; C/D also change generation settings. These contrasts do not isolate the causal effect of context or reasoning. The recovered B/D prompt and configuration details are in Appendices I and J.

Baseline inputs also differ: classical and adapted systems use retained pair-level or answer-only procedures, whereas the majority baseline receives no text. Adapted runs use documented model-derived training-label variants on the shared-meeting pool; exact checkpoint-to-release-label linkage is incomplete (Table 52). Results compare documented system/protocol combinations, rather than matched-input model capabilities.

D.2 Performance on Human-Reference Tests

Strong scores, uncertain ordering. Among systems using full-meeting prompting on Test-Std, Opus 4.7 scores 86.7% binary Macro-F1 and GPT-5.2 scores 85.3% (Table 16). Their 1.4 pp difference is unresolved: the 95% interval is $[-4.4, +7.0]$ pp when resampling pairs and $[-4.4, +7.2]$ pp when resampling whole meetings (10,000 draws, seed 42; Appendix D.5). Neither establishes equivalence. The three-class point estimates reverse their order: GPT-5.2 scores 83.5%, versus 81.7% for Opus 4.7.

The reversal is concentrated in the ten fully evasive cases. GPT-5.2's per-class F1 differences relative to Opus 4.7 are -1.1 pp for direct, -1.0 pp for intermediate, and $+7.4$ pp for fully evasive. For the nine-system cached panel (six frontier systems, both Qwen systems, and Eva-4B+FOMC), binary/three-class rank agreement is Kendall $\tau_b = 0.78$ (95% bootstrap interval $[0.39, 0.89]$) and Spearman $\rho = 0.90$. These results support reporting both targets with their uncertainty.

Classical and adapted baselines. Cached Qwen3-4B and Qwen3-FOMC predictions score 40.5% and 57.3% on three-class evaluation, and 40.7% and 63.4% on binary evaluation. Table 51 separates these runs from summary-only results. DeBERTa-v3 (He et al., 2023) has a retained three-class result, but its missing checkpoint and predictions prevent binary rescoring. Earnings-trained Eva-4B (Ma et al., 2026), without FOMC adaptation, scores 77.6% three-class F1 when all 200 cases are retained and one invalid output counts as an error. The distinct FOMC-adapted Eva-4B+FOMC scores 78.4% three-class and 84.8% binary F1.

Pooled-test comparisons. For three-class F1 on the combined 400 tests, four paired comparisons contrast GPT-5.2 with Opus 4.7, GPT-5.4, Gemini 3.1 Pro, and Eva-4B+FOMC (10,000 resamples, seed 42; Table 23). Only the Gemini 3.1 Pro comparison has a Holm-adjusted value below 0.05 in these independent-pair calculations. Meeting-cluster and post-selection error control are not established. The comparisons also condition on a challenge split selected partly through older models and differing input/training regimes; they do not establish a general benefit of model recency or a resolved advantage over the adapted baseline.

Table 16: Human-reference Test-Std results (200 pairs): binary direct/non-direct Macro-F1 (T1b, primary) and three-way Macro-F1 (T1, secondary). A uses pair-only, degree-only prompting; B uses full-meeting, three-axis prompting.

System	Type	Binary Macro-F1	Three-class Macro-F1
Majority†	Trivial	$0.306_{[0.27,0.34]}$	$0.204_{[0.18,0.23]}$
Lexical†	Rule-based	$0.489_{[0.42,0.56]}$	$0.346_{[0.30,0.39]}$
TF-IDF+LR†	Classical ML	$0.535_{[0.46,0.60]}$	$0.585_{[0.44,0.69]}$
Qwen3-4B	Zero-shot 4B	$0.407_{[0.36,0.46]}$	$0.405_{[0.28,0.50]}$
DeBERTa-v3†	Fine-tuned encoder	—	$0.664_{[0.55,0.76]}$
Qwen3-FOMC	Fine-tuned 4B	$0.634_{[0.56,0.70]}$	$0.573_{[0.44,0.68]}$
Eva-4B	Earnings-trained 4B	$0.813_{[0.76,0.87]}$	$0.776_{[0.67,0.85]}$
Eva-4B+FOMC	FOMC-adapted 4B	$0.848_{[0.79,0.89]}$	$0.784_{[0.68,0.86]}$
Opus 4.6 A	Frontier LLM	$0.880_{[0.83,0.92]}$	$0.805_{[0.70,0.88]}$
GPT-5.2 A	Frontier LLM	$0.785_{[0.72,0.84]}$	$0.725_{[0.63,0.80]}$
Gemini 3 A	Frontier LLM	$0.870_{[0.82,0.91]}$	$0.678_{[0.60,0.75]}$
Opus 4.6 B	Frontier LLM	$0.830_{[0.78,0.88]}$	$0.729_{[0.58,0.83]}$
Gemini 3 B	Frontier LLM	$0.799_{[0.74,0.85]}$	$0.771_{[0.67,0.85]}$
GPT-5.2 B	Frontier LLM	$0.853_{[0.80,0.90]}$	$0.835_{[0.74,0.90]}$
Opus 4.7 B	Frontier LLM	$0.867_{[0.82,0.91]}$	$0.817_{[0.72,0.89]}$
GPT-5.4 B	Frontier LLM	$0.825_{[0.77,0.87]}$	$0.788_{[0.67,0.87]}$
Gemini 3.1 Pro B	Frontier LLM	$0.794_{[0.73,0.85]}$	$0.715_{[0.61,0.80]}$

Inputs differ: B receives the full unlabeled meeting; retained baseline procedures use target-pair text, answer text, or no text (Majority; Section 4). FOMC adaptation uses a shared-meeting training pool and historical label variants; checkpoint linkage to released labels is unverified (Table 52). These system/protocol comparisons do not isolate capability at matched input. Pointwise 95% pair-bootstrap intervals use 10,000 resamples; they are not simultaneous and do not establish adjacent ranks. †Archived evaluations; DeBERTa lacks predictions for binary rescoring. Both Qwen metrics use the same cache (Table 51). Eva-4B's one invalid output counts as an error among all 200 pairs.

D.3 Remaining Responsiveness Errors

Under full-meeting, three-axis prompting (B), 570 three-class model–pair predictions across six frontier systems and both test splits disagree with the reference. Of these, 77% assign a lower evasion degree, including 389 intermediate-to-direct predictions. Thus a recurring disagreement is whether an answer that engages with a question has actually supplied the requested information. The direction depends on protocol: pooling both tests, the less-evasive share for Opus 4.6, GPT-5.2, and Gemini 3 Pro is 22%, 6%, and 53% under A, versus 83%, 66%, and 72% under B (Appendix D.5). These are repeated model–pair outcomes, not distinct examples or population error rates.

The existing targeted rationale audit identifies 17 instances across 12 pairs and all six systems in which a rationale notes an answer shortfall but the final label is direct (Appendix D.6). This provides a lower-bound diagnostic of rationale–label inconsistency, without explaining every disagreement. Disagreements also occur on unanimously rated items: GPT-5.2 B disagrees with 31 of the 292 such test pairs.

Errors are partly complementary: mean pairwise system agreement is 0.72 and 60% of test pairs are contested. However, neither of two tested plurality-voting tie rules produces a statistically resolved gain; results depend on the 45 tied cases. A reference-assisted per-pair oracle reaches three-class F1 of 0.988, versus 0.833 for the best single system, but is not a deployable selector (Appendix D.5).

D.4 Protocol and Reference Sensitivity

Prompting changes point estimates. On human-reference Test-Std, switching from pair-only, degree-only (A) to full-meeting, three-axis (B) prompting changes primary binary F1 by -5.0, $+6.8$, and -7.0 pp for Opus 4.6, GPT-5.2, and Gemini 3 Pro. The corresponding three-class changes are

-7.6, $+11.0$, and $+9.3$ pp: Gemini's direction reverses across targets. Pair-only Opus 4.6 reaches 88.0% binary F1, above the 86.7% meeting-level Opus 4.7 point estimate, but their difference is unresolved (95% paired interval $[-4.5, +7.4]$ pp).

Appendix Table 17 reports both targets and splits. Among 12 retrospective protocol contrasts (three models, two targets, two splits), only Gemini's binary Test-Challenge contrast has a Holm-adjusted value below 0.05 ($p = 0.017$; 10,000 pair resamples, seed 42); none on Test-Std crosses that threshold. These are independent-pair calculations, without established meeting-cluster or post-selection error control. Input access and instructions change together, so the results do not establish a general advantage of meeting-level prompting.

The reference can reverse a protocol comparison. For fixed predictions on the same 200 Test-Std pairs, Opus 4.6's binary B-minus-A difference is $+15.6$ pp against its Condition-D labels and -5.0 pp against human-majority labels (Appendix Table 18). This reference change shifts the contrast by -20.7 pp (pointwise 95% paired interval $[-29.3, -12.8]$); three-class differences are $+28.4$ and -7.6 pp. The human-reference protocol contrast remains unresolved. The model reference shares Opus's family and B's meeting-level input, but this comparison cannot separate those influences. Full-corpus ablations and their separate corrections appear in Appendix F.

Secondary labels support narrower analyses. Mechanism identification has Weighted-F1 of 0.827 [0.80, 0.85] for Opus B on 960 pairs selected to be non-direct under both its prediction and the Opus diagnostic reference, with a valid predicted mechanism. This same-family, prediction-selected agreement lacks human validation; the majority baseline uses a different set of 1,092 pairs (Table 34). Model-reference agreement also varies by meeting size and chair, but those comparisons change content, period, model, or prompting together and do not isolate their effects (Tables 43 and 30).

Cross-benchmark differences remain descriptive. Unadapted Eva-4B's 77.6% FOMC three-class score is about 7.3 pp below its reported earnings-call score of 84.9%. Different data, labels, and protocols prevent interpreting this as an isolated transfer effect. The cross-corpus analyses describe model-assigned labels and do not isolate transfer, annotation-noise, or institutional effects (Appendix D.9). The earnings-call analysis populations are documented there; their record overlap with the earlier datasets (Al Nuaimi et al., 2025a;b) has not been established.

D.5 Human-reference performance and uncertainty

The starter leaderboard (Table 16) and binary evaluation (Table 19) use human-reference labels on Test-Std. Table 20 reports the separate difficulty-enriched test split. The class-level and scoring-boundary analyses below use the stated test populations and cached predictions. Table 17 covers both metrics and splits for all three available A/B contrasts, including results unfavourable to full-meeting prompting. Its joint retrospective Holm family contains 12 tests. The separate model-version/protocol contrast, Opus 4.6 A minus Opus 4.7 B on binary Test-Std, is $+0.013$ (95% paired interval $[-0.045, +0.074]$, exploratory raw $p = 0.694$); the point estimate does not establish a better model or protocol. Table 23 preserves earlier three-test and pooled-test families and separately labels full-corpus model-reference results.

Meeting-cluster uncertainty. Test-Std contains 200 pairs across 80 meetings (mean 2.5, range 1–7 pairs per meeting). For binary Macro-F1 of Opus 4.7 B minus GPT-5.2 B, the observed difference is 0.014. Sampling 80 whole meetings with replacement, preserving all their test pairs and draw multiplicities, gives a 95% percentile interval of $[-0.0438, 0.0722]$, compared with $[-0.0435, 0.0704]$ from paired item resampling (10,000 draws, seed 42). Both systems share each draw; scores weight pairs rather than meetings equally. The conclusion is unchanged for this comparison. The cluster diagnostic assumes exchangeable sampled meetings and does not repair shared Train/Test meetings, retrospective test use, uncertain reference labels, or missing run-to-run variation. It does not establish generalisation to unseen meetings.

Protocol-specific error direction. The six full-meeting B systems yield 438 less-evasive predictions among 570 three-class disagreements across both tests (76.8%). For the three models with A/B caches, the corresponding A versus B counts are Opus 4.6: 14/64 versus 85/102; GPT-5.2: 5/88 versus 48/73; and Gemini 3 Pro: 49/92 versus 77/107. Each denominator counts disagreements for

Table 17: Protocol sensitivity on human-majority references ($N = 200$ per split). A uses pair-only input and degree-only instructions; B uses full-meeting input and three-axis instructions. Scores and $\Delta = B - A$ are Macro-F1. Pointwise 95% paired-bootstrap intervals use 10,000 shared pair resamples and seed 42. Holm12 adjusts inclusive two-sided sign-tail p-values (with a $+1$ Monte Carlo correction) jointly across the retrospective 12-contrast family. Only Gemini 3 Pro on Test-Challenge binary has Holm12 $p < .05$. These fixed-run comparisons change both input access and instructions and do not identify a causal context effect.

Model	**Task**	**A**	**B**	**Δ [95% CI]**	**Holm12 p**
Test-Std					
Opus 4.6	Binary	0.880	0.830	-0.050 [$-0.119, +0.020$]	0.795
GPT-5.2	Binary	0.785	0.853	$+0.068$ [$-0.006, +0.144$]	0.614
Gemini 3 Pro	Binary	0.870	0.799	-0.070 [$-0.135, -0.009$]	0.274
Opus 4.6	Three-way	0.805	0.729	-0.076 [$-0.217, +0.055$]	0.795
GPT-5.2	Three-way	0.725	0.835	$+0.110$ [$-0.009, +0.231$]	0.614
Gemini 3 Pro	Three-way	0.678	0.771	$+0.093$ [$-0.011, +0.194$]	0.614
Test-Challenge (disagreement-enriched)					
Opus 4.6	Binary	0.778	0.715	-0.063 [$-0.156, +0.031$]	0.795
GPT-5.2	Binary	0.750	0.776	$+0.026$ [$-0.081, +0.131$]	0.822
Gemini 3 Pro	Binary	0.806	0.688	-0.118 [$-0.192, -0.047$]	**0.017**
Opus 4.6	Three-way	0.811	0.679	-0.132 [$-0.233, -0.037$]	0.077
GPT-5.2	Three-way	0.773	0.809	$+0.036$ [$-0.052, +0.126$]	0.822
Gemini 3 Pro	Three-way	0.792	0.726	-0.065 [$-0.141, +0.010$]	0.614

Table 18: Reference sensitivity on identical Test-Std pairs ($N = 200$). B-minus-A Macro-F1 differences in percentage points hold predictions, target, pair IDs, and output policy fixed; only the reference changes. H denotes human majority; D denotes Opus 4.6's Condition-D diagnostic labels. The change in the contrast is H minus D, with pointwise 95% paired-bootstrap intervals (10,000 resamples, seed 42). These retrospective intervals are not adjusted for multiple comparisons or meeting clustering.

	Binary			Three-class		
System	H	D	H−D [95% CI]	H	D	H−D [95% CI]
Opus 4.6	-5.0	+15.6	-20.7 [$-29.3, -12.8$]	-7.6	+28.4	-36.0 [$-56.7, -17.4$]
GPT-5.2	+6.8	+16.3	-9.5 [$-17.3, -1.9$]	+11.0	+24.4	-13.4 [$-28.6, +1.6$]
Gemini 3 Pro	-7.0	-3.0	-4.0 [$-10.1, +2.0$]	+9.3	+11.8	-2.6 [$-13.0, +8.8$]

that model/protocol, so the sets of affected pairs differ. This is a descriptive protocol contrast, not an isolated context effect. The analysis code reports each split separately as well as the explicitly pooled diagnostic.

Matched reference sensitivity. Table 18 re-scores the same A/B predictions on the same Test-Std IDs under human-majority and Opus-D references, separately for each target. Opus changes sign under both targets; GPT-5.2 remains positive, and Gemini remains negative for binary and positive for three-class F1. This controlled scoring comparison establishes reference sensitivity for these fixed outputs, not a causal explanation for reference disagreement. The broader full-corpus effects in Table 23 also change population and are separate diagnostics. We emphasise Test-Std because Test-Challenge was selected partly through the Opus-D panel; its reference contrasts additionally reflect that selection. The test-only Opus-D reference slice and reproduction code accompany the analysis.

The nine-system rank analysis in Section D.2 uses six frontier systems, both Qwen systems, and Eva-4B+FOMC, with 2,000 pair-bootstrap resamples and seed 20260727. The ten-system boundary analysis below additionally includes unadapted Eva-4B; its panel and handling of invalid outputs are distinct.

Table 19: Primary binary evaluation on the 200 human-reference Test-Std pairs. Intermediate and fully evasive are collapsed into non-direct; invalid outputs count as errors rather than being assigned to a class.

System	T1b Macro-F1 [95% CI]	κ
Majority†	$0.306_{[0.27,0.34]}$	0.000
Lexical†	$0.489_{[0.42,0.56]}$	0.004
TF-IDF+LR†	$0.535_{[0.46,0.60]}$	0.163
Qwen3-4B	$0.407_{[0.36,0.46]}$	0.051
DeBERTa-v3†	—	—
Qwen3-FOMC	$0.634_{[0.56,0.70]}$	0.304
Eva-4B	$0.813_{[0.76,0.87]}$	0.627
Eva-4B+FOMC	$0.848_{[0.79,0.89]}$	0.696
Opus 4.6 B	$0.830_{[0.78,0.88]}$	0.664
Gemini 3 B	$0.799_{[0.74,0.85]}$	0.601
GPT-5.2 B	$0.853_{[0.80,0.90]}$	0.707
Opus 4.7 B	$0.867_{[0.82,0.91]}$	0.735
GPT-5.4 B	$0.825_{[0.77,0.87]}$	0.654
Gemini 3.1 Pro B	$0.794_{[0.73,0.85]}$	0.603

Conventions and historical-result markers follow Table 16. No binary estimate is available for DeBERTa-v3.

Table 20: Diagnostic Test-Challenge results (200 human-reference pairs), which oversample model disagreement. Binary and three-way Macro-F1 use the same prediction cache and invalid-output error convention as Test-Std.

System	T1b Macro-F1 [95% CI]	T1 Macro-F1 [95% CI]
Opus B	$0.715_{[0.65,0.78]}$	$0.679_{[0.59,0.75]}$
GPT-5.2 B	$0.776_{[0.71,0.84]}$	$0.809_{[0.75,0.86]}$
Gemini 3 B	$0.688_{[0.62,0.75]}$	$0.726_{[0.66,0.78]}$
Opus 4.7 B	$0.772_{[0.70,0.83]}$	$0.758_{[0.69,0.82]}$
GPT-5.4 B	$0.720_{[0.65,0.78]}$	$0.740_{[0.67,0.80]}$
Gemini 3.1 Pro B	$0.708_{[0.64,0.77]}$	$0.720_{[0.65,0.78]}$
Eva-4B	$0.735_{[0.66,0.80]}$	$0.748_{[0.68,0.81]}$
Qwen3-4B	$0.478_{[0.42,0.54]}$	$0.528_{[0.44,0.60]}$
Qwen3-FOMC	$0.609_{[0.53,0.68]}$	$0.584_{[0.49,0.66]}$
Eva-4B+FOMC	$0.732_{[0.66,0.80]}$	$0.736_{[0.66,0.80]}$

Pointwise pair-bootstrap intervals use 10,000 resamples and seed 42. Differences from Test-Std reflect the altered sampling frame as well as model performance.

Sensitivity to the scoring boundary. We re-score fixed labels and predictions from the same ten-system cached-run panel under four alternatives: binary direct versus non-direct, fully-evasive versus the other two classes, quadratic-weighted Cohen's κ, and negative mean absolute ordinal error. The ordinal encodings are direct $= 0$, intermediate $= 1$, and fully_evasive $= 2$; higher scores are better after negating mean absolute error. The sensitivity calculation uses 199 parseable predictions per test split for Eva-4B (OOD) and all 200 for the other systems; it therefore differs from the complete-coverage primary table. These analyses change the scoring rule without collecting new labels. Table 22 reports rank correlations with three-way Macro-F1. Rankings are not invariant: agreement is weakest when the scoring rule isolates the rare fully-evasive class. Agreement under ordinal metrics supports reporting degree information alongside the primary binary task, but does not validate a unique semantic boundary.

Rare-class uncertainty. The sensitivity analysis uses 10,000 pair-level percentile-bootstrap samples with shared sample indices across systems (seed 20260725). Class-F1 is omitted from a resample only if that class has neither reference nor predicted instances. Across the ten cached systems in that sensitivity analysis, the median Test-Std interval widths are 0.118 for direct, 0.125 for intermediate, and 0.504 for fully_evasive. The latter rests on only ten reference examples; Test-Challenge contains 23. Table 21 separately recomputes the class-level intervals for the principal systems with 10,000

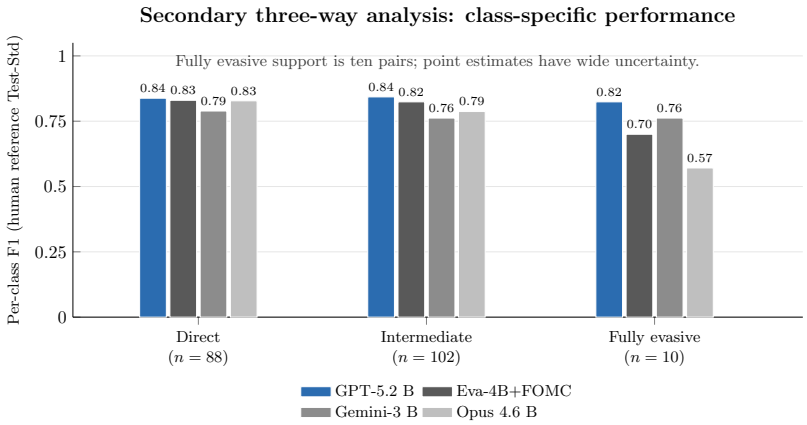


Figure 4: **Per-class F1 reveals rare-class sensitivity.** On 200 human-reference Test-Std pairs, GPT-5.2 B exceeds Opus 4.6 B by 10.6 pp three-class Macro-F1 (0.835 versus 0.729). Fully evasive F1 (0.824 versus 0.571) contributes about 8.4 pp of this difference. Class supports are 88 direct, 102 intermediate, and ten fully evasive. The small fully evasive class makes this comparison uncertain; it does not establish suitability for policy monitoring. Table 21 reports class scores and intervals.

Table 21: Per-class F1 and pointwise 95% pair-bootstrap intervals on human-reference Test-Std. Support is 88 direct, 102 intermediate and 10 fully evasive examples.

System	Direct	Intermediate	Fully evasive
GPT-5.2 B	$0.838_{[0.78,0.89]}$	$0.843_{[0.78,0.89]}$	$0.824_{[0.55,1.00]}$
Eva-4B+FOMC	$0.830_{[0.77,0.88]}$	$0.824_{[0.76,0.88]}$	$0.700_{[0.40,0.90]}$
Gemini 3 B	$0.789_{[0.72,0.85]}$	$0.762_{[0.69,0.82]}$	$0.762_{[0.50,0.94]}$
Opus B	$0.828_{[0.77,0.88]}$	$0.787_{[0.72,0.85]}$	$0.571_{[0.17,0.86]}$

10,000 resamples, seed 42. These wide rare-class intervals do not establish an ordering between frontier systems. Eva-4B+FOMC denotes the adapted model, not the earnings-only model.

draws and seed 42. These intervals condition on the sampled pairs and fixed predictions; they do not establish stable frontier rankings or remove annotation uncertainty.

Ensemble comparison. On the combined 400 test pairs, the six-system plurality ensemble has 45 ties. Resolving tied labels by ordinal median gives three-class Macro-F1 0.024 below the best single model (95% paired-bootstrap interval $[-0.065, +0.014]$, raw $p = 0.22$). Choosing the most evasive tied label instead gives 0.854 versus the best model's 0.833 (raw $p = 0.225$). Neither tested rule yields a statistically resolved gain. The reference-assisted per-pair oracle selects a correct prediction when any system supplies one and reaches 0.988; this uses the reference label and is not a deployable selection method. These are retrospective fixed-run analyses with pair-resampling limitations.

D.6 Human-reference error analysis

To complement the aggregate statistics, Table 24 presents 24 curated examples from the human-reference 400-pair test set, organised by failure pattern, covering four behaviourally distinct buckets. The pool sizes characterise the relative prevalence of each pattern: (a) *Direct reference: A disagrees, B agrees*: 33 of 400 pairs (8.3%) whose reference is direct, where the Condition A majority predicts intermediate or fully evasive and the Condition B majority matches the reference; this within-test-set contrast is distinct from the full-corpus model-reference comparison in Table 29. (b) *Direct–intermediate boundary*: 218 of 400 pairs (54.5%) whose human reference is direct or intermediate, for which the six model-configuration votes (3 models $\times$ 2 conditions) contain at least two labels and at least one vote assigns the opposite direct/intermediate label. (c) *Fully evasive reference: B agrees*: 25 of 400 pairs (6.3%) where the reference is fully evasive and Condition B agrees—a positive-control bucket showing that B does not assign direct to every answer. (d) *Reverse pattern A-corrects-B*: 70

Table 22: Sensitivity of system rankings to the scoring scheme. Kendall τ_b compares each scheme with three-way Macro-F1 over ten systems: six frontier systems, the two Qwen systems and the two Eva variants.

Alternative scoring scheme	**Test-Std τ_b**	**Test-Challenge τ_b**
Direct vs non-direct	0.778	0.822
Fully evasive vs rest	0.689	0.467
Quadratic-weighted κ	0.822	0.911
Negative ordinal MAE	0.854	0.764

This reproduces the boundary analysis: each split has 200 human-reference pairs, with Eva-4B evaluated on its 199 valid outputs in each split (all other systems: 200). Scoring schemes change the target; these correlations do not establish significance of adjacent ranks. Dev rankings were excluded from this historical analysis because its Dev references were model consensus; the current release supplies human-majority Dev references.

Table 23: Effect sizes and archived Holm-adjusted values. Δ is left-minus-right three-way Macro-F1; intervals are pointwise 95% paired-bootstrap intervals.

Comparison	**Δ F1**	**95% CI**	**McNemar p**	**Holm p**
Condition-D agreement: p tests paired correctness, full 24-test family				
Opus B vs Opus A	+0.218	[+0.182, +0.255]	$< 10^{-4}$	$< 10^{-4}$
GPT B vs GPT A	+0.162	[+0.131, +0.192]	$< 10^{-4}$	$< 10^{-4}$
Gemini B vs Gemini A	+0.102	[+0.076, +0.128]	0.0004	0.005
Opus B vs Eva-4B	+0.193	[+0.155, +0.231]	$< 10^{-4}$	$< 10^{-4}$
Test-Std 200 human-reference pairs: three-way paired-bootstrap tests, three-test family				
Comparison	**Δ F1**	**95% CI**	**Bootstrap p**	**Holm p**
Opus 4.6 B vs A	-0.076	[-0.217, +0.055]	0.244	0.244
GPT-5.2 B vs A	+0.110	[-0.009, +0.231]	0.068	0.205
Gemini 3 Pro B vs A	+0.093	[-0.011, +0.194]	0.079	0.205
Combined 400 human-reference pairs: three-way paired-bootstrap tests, four-test family				
Comparison	**Δ F1**	**95% CI**	**Bootstrap p**	**Holm p**
Eva-4B+FOMC vs GPT-5.2 B	-0.058	[-0.119, +0.002]	0.058	0.115
Opus 4.7 B vs GPT-5.2 B	-0.038	[-0.094, +0.017]	0.184	0.184
GPT-5.4 B vs GPT-5.2 B	-0.059	[-0.110, -0.009]	0.022	0.066
Gemini 3.1 Pro B vs GPT-5.2 B	-0.099	[-0.155, -0.045]	0.0002	0.0008

The upper block is recomputed from retained caches: F1 differences and intervals use common valid IDs ($n = 2154, 2154, 2153, 2142$ in row order), 10,000 paired resamples and seed 42. Its unchanged McNemar/Holm tests concern degree-label correctness against Condition D (Table 44), with $n = 2154, 2154, 2153, 2154$; Eva invalid outputs remain errors for these tests. The middle and lower blocks are recomputed from cached predictions with 10,000 paired resamples, seed 42, a two-sided sign-tail bootstrap test and a +1 Monte Carlo correction. The retrospective families are three B-versus-A comparisons on Test-Std and, separately, Eva-4B+FOMC and the three newer frontier systems versus GPT-5.2 on the pooled tests. Pointwise intervals are not multiple-testing-adjusted. In these archived independent-pair calculations, no Test-Std protocol adjusted value is below 0.05; only Gemini 3.1 Pro crosses that threshold in the pooled-test family. Meeting-cluster and post-selection error control are not established.

of 400 (17.5%) where the Condition A majority matches the reference and Condition B does not, including intermediate reference answers that Condition B labels direct. The selected examples do not identify why the configurations differ. No pair in the 400 exhibits a three-way disagreement under Condition B (across the three frontier LLM families), indicating that disagreement, when it occurs, is binary between two labels rather than scattered. These overlapping, curated diagnostic buckets illustrate both directions of protocol error; they are not a disjoint error census or evidence of a uniform Condition B improvement.

Table 24: Curated examples from the human-reference 400-pair test set, organized by prediction pattern. A/B votes show the majority prediction across Opus 4.6, GPT 5.2, Gemini 3 Pro under each prompt condition. Questions, answers, and Opus Condition B rationales are abridged verbatim; bracketed ellipses mark omissions. Rationales describe one model's judgment and need not match the B majority or human reference.

Pair	Question excerpt	Answer excerpt	Reference	A / B	Opus B rationale
(a) Direct reference — A disagrees, B agrees		(pool size = 33)			
M20191211_P009(Powell)	[. . .] what is it that gives you confidence that the Fed has the tools to get inflation back up to target or even overshoot [. . .]	[. . .] there is a relationship between resource utilization, by which I mean unemployment, and inflation. It's just relatively weak. [. . .] I would say there's more humility than there is confidence in this at this point. [. . .]	dir	inter / dir	[. . .] by citing the empirical relationship between unemployment and inflation, while also honestly acknowledging humility and the difficulty of the challenge. [. . .]
M20120620_P017(Bernanke)	[. . .] the automatic tax-hike aspect of that and the automatic spending cuts [. . .] How do you parse the relative importance of those two aspects? [. . .] how the Fed is going to conduct open market operations [. . .] they will essentially have no short-term securities to use.	[. . .] the dollar amount associated with the tax expirations—tax-cut expirations, including the payroll tax cut and so on, is larger than the spending cuts [. . .] We'll still be able to do open market operations with our securities, even if the amount of short-term debt is very low. [. . .]	dir	inter / dir	[. . .] Both parts are addressed. [. . .] he notes the dollar amount of tax expirations is larger but emphasizes the total matters. [. . .] he explains they can still operate [. . .]
M20170614_P010(Yellen)	[. . .] a group of economists sent the Fed a letter earlier this month disagreeing with your 2 percent inflation target [. . .] You say you're committed to the 2 percent target, but what do you say to them?	[. . .] It's one that we will be reconsidering at some future time. [. . .] It needs to be a balanced assessment. [. . .]	dir	inter / dir	[. . .] acknowledging that the neutral rate has declined making the zero lower bound constraint more relevant, conceding this is an important question they will reconsider in the future [. . .]
M20230503_P012(Powell)	[. . .] If the whole point of slowing down the pace was to see the effects of your moves [. . .] why did the Committee feel it was necessary to keep moving?	[. . .] We always have to balance the risk of not doing enough and, and not getting inflation under control against the risk of maybe slowing down economic activity too much. And we thought that this rate hike, along with the meaningful change in, in our policy statement, was the right way to balance that.	dir	inter / dir	[. . .] He explains they're trying to reach a sufficiently restrictive stance, that slowing down was right but continued tightening was warranted [. . .]
M20211103_P017(Powell)	[. . .] once you get the taper process complete, will you hold the balance sheet steady, or will you allow it to start passively winding down? [. . .] what is bond buying actually doing to help the economy?	[. . .] we haven't gone back to them. [. . .] we'll talk about reinvestment and all those things. [. . .] That will drive down longer-term rates and hold them lower [. . .] So lower rates encourage more borrowing, encourage more economic activity. [. . .]	dir	inter / dir	[. . .] they haven't yet discussed reinvestment/runoff but plan to in coming meetings, and he provides a thorough explanation of how QE works [. . .]
M20231213_P004(Powell)	[. . .] can you confidently say that the economy has avoided a recession and isn't heading for one now? And if the answer is "no," I'm curious about what you'd still be looking for. [. . .]	[. . .] there's little basis for thinking that the economy is in a recession now. [. . .] the economy could cool off in a way that enabled inflation to come down without the kind of large job losses [. . .] This result is not guaranteed. [. . .]	dir	inter / dir	[. . .] acknowledges a meaningful probability always exists, and describes the scenario he's been hoping for (cooling without large job losses). [. . .]

(continued on next page)

(Table 24 continued from previous page.)

Pair	Question excerpt	Answer excerpt	Reference	A / B	Opus B rationale
(b) Direct/intermediate boundary — model disagreement			(pool size = 218)		
M20241107_P022(Powell)	[. . .] many average Americans are still not feeling the strength of the economy in their wallets. So what's your message to them on when they might expect relief?	[. . .] it takes some years of real wage gains for people to feel better. [. . .] But it'll be some time before people, you know, regain their confidence and feel that. [. . .]	dir	inter / inter	[. . .] Powell acknowledges the pain, explains the mechanism (real wage gains over time), says the process is underway [. . .]
M20120913_P019(Bernanke)	[. . .] How long do you think that it might work that you have such strong growth and no inflation pressure? [. . .] you give some kind of carte blanche to the fiscal policy and to the Congress not to do enough on their side of the policy action?	[. . .] there's a lot of slack in the economy [. . .] So we don't anticipate the economy is going to be overheating anytime soon. [. . .] Our job is to use the tools we have to meet our mandate, which is maximum employment and price stability. [. . .]	dir	inter / dir	[. . .] explains the slack/unused capacity means faster growth without overheating [. . .] the Fed's obligation is to use its tools to meet its mandate regardless of what fiscal policymakers do.
M20140618_P013(Yellen)	[. . .] with only four SEPs and press conferences a year, how do you signal to the markets what the Fed is doing so that you don't run the risk of an event like last September when people were surprised [. . .]	[. . .] I would feel it appropriate for me to either have additional communications meetings with the press or to give speeches [. . .] to make clear what the Committee's thinking is [. . .]	inter	dir / inter	[. . .] but doesn't address the structural signaling mechanism problem or how the Fed will avoid surprises. The answer is vague on the 'how' of signaling.
M20120913_P020(Bernanke)	[. . .] as the economy picks up steam, whether the FOMC will really follow through and keep rates low or whether you will do as the Fed has always done and begin to raise the funds rate.	[. . .] the empirical evidence is that our announcements do have considerable credibility. [. . .] there is a consensus that even as the personnel change and so on going forward, that this is the appropriate approach [. . .]	dir	inter / dir	[. . .] cites empirical evidence that markets respond to Fed guidance, and argues there's a consensus within the Committee supporting this approach even as personnel change.
M20151216_P022(Yellen)	[. . .] Are you concerned about the negative impact that your decision could have on the emerging markets? Do you fear it could trigger some financial turbulences abroad that could eventually hit the U.S.?	[. . .] I think this move has been expected and well communicated [. . .] there can be negative spillovers through capital flows [. . .] there are also positive spillovers from a strong U.S. economy. [. . .]	dir	inter / dir	[. . .] argues this move was well-communicated and not a surprise, discusses both negative and positive spillovers, assesses emerging markets' resilience [. . .]
M20171213_P013(Yellen)	[. . .] What use do you see a cryptocurrency could have for the Fed, and do you see the potential of a cryptocurrency that might be considered legal tender?	[. . .] this is not something the Federal Reserve is seriously considering at this stage [. . .] limited benefits from introducing it, a limited need for it, and some substantial concerns. [. . .]	dir	inter / dir	[. . .] states the Fed is not seriously considering it, cites limited benefits and substantial concerns, and effectively addresses the question fully.
(c) Fully evasive reference — B agrees (positive control)			(pool size = 25)		
M20211103_P018(Powell)	[. . .] questions about the stock purchases that took place from some of the regional Fed presidents [. . .] what would you say—to the President and to senators [. . .] about this, specifically with regards to your future as Fed Chair for a possible second term?	[. . .] I'm not going to have any comment whatsoever on the renomination process at all. [. . .]	evas	evas / evas	[. . .] Powell explicitly refuses to comment on the renomination process at all.

(continued on next page)

(Table 24 continued from previous page.)

Pair	Question excerpt	Answer excerpt	Reference	A / B	Opus B rationale
M20110622_P013(Bernanke)	[. . .] what would your appropriate policy be to get there? Would it be an early gradual exit, a late steep exit, or no exit?	[. . .] I don't think it would be constructive for me to give you a tentative projection [. . .] there's simply no alternative but to watch the incoming data [. . .]	evas	evas / evas	[. . .] Bernanke explicitly declines: [. . .] and pivots to general statements about data dependence and uncertainty.
M20140917_P012(Yellen)	[. . .] Do you see any impact if Scotland does vote for independence on the European economy and potentially on the American economy? And, if so, is the Fed doing anything in preparation for that possibility? [. . .]	[. . .] Scottish voters are about to go to the polls tomorrow [. . .] I really don't want to weigh in on this today.	evas	evas / evas	[. . .] Yellen explicitly refuses to weigh in, providing no substantive content whatsoever.
M20190619_P015(Powell)	[. . .] you should push back on the President's criticisms rather than ignoring him? [. . .] do you think that you and the President have the same goals when it comes to monetary policy?	[. . .] I don't discuss elected officials publicly or privately, really. [. . .] our independence from direct political control we see as an important institutional feature [. . .]	evas	evas / evas	[. . .] Powell declines to discuss elected officials and deflects to the Fed's mission and independence, not answering either question directly. [. . .]
M20211103_P005(Powell)	[. . .] what are the economic conditions that would perhaps warrant a faster pace of tapering? [. . .] how you would also characterize the risks that the Fed may actually need to accelerate that process eventually. [. . .]	[. . .] we're prepared to deviate from that path if warranted by changes in the economic outlook. [. . .] I'm not going to give you a lot more detail on what that might be. [. . .]	evas	evas / evas	[. . .] Powell restates the statement language about being prepared to deviate but explicitly declines to provide detail on what conditions would trigger it.
M20250129_P021(Powell)	[. . .] when it comes to executive orders and OMB memos, do those always apply to the Fed, sometimes, never, or do you just often voluntarily comply? [. . .]	[. . .] work to align our policies to those that are mentioned in, in the executive orders. [. . .] I'm not going to go any deeper than that [. . .]	evas	evas / evas	[. . .] explicitly refuses to go deeper, not answering the legal question about applicability.
(d) Reverse pattern — A agrees with the reference, B disagrees			(pool size = 70)		
M20110622_P010(Bernanke)	[. . .] Does that mean you think the medium-term tradeoff between inflation and growth has got worse? [. . .] Has the unexpected rise in core inflation changed your understanding of the output gap? [. . .]	[. . .] the Committee still believes that the output gap is quite large. [. . .] some of the temporary factors affecting inflation, including core inflation, are likely to recede [. . .]	inter	inter / dir	[. . .] implicitly addressing that the tradeoff hasn't fundamentally worsened. Both parts are substantively engaged.
M20140618_P014(Yellen)	[. . .] Do you think the U.S. can be a lone engine of economic recovery globally? [. . .] is financial stability currently preventing the Fed from being more accommodative than it would like? [. . .] when do you expect that to happen, if at all? [. . .]	[. . .] I don't see them shaping monetary policy in an important way right now. [. . .] You expressed a lot of pessimism about emerging markets, and I see it more likely that we'll see moderate growth and a pickup there. [. . .]	inter	inter / dir	[. . .] On the global risks question, she only briefly addresses it [. . .] She doesn't directly address whether the U.S. can be a 'lone engine.'

(continued on next page)

(Table 24 continued from previous page.)

Pair	Question excerpt	Answer excerpt	Reference	A / B	Opus B rationale
M20190619_P005(Powell)	[. . .] I was hoping that you could clarify for us a little bit how you're thinking [. . .] about the risks of waiting too long to cut rates versus the risks of cutting rates prematurely—sort of, what the balance of risks are and how you talk about that.	[. . .] given the quite recent nature of some of the events [. . .] the right thing to do was to wait and see more. [. . .] I don't think the risk of waiting too long is prominent right now. [. . .]	dir	dir / inter	[. . .] doesn't substantively discuss the risk of cutting prematurely or how the two are weighed against each other.
M20111102_P015(Bernanke)	[. . .] a 9 percent Tier 1 capital ratio is appropriate. [. . .] do you think it's appropriate that the big banks in the U.S. should accept such a Tier 1 capital requirement, and how does that development stand at this point?	[. . .] implementing Basel III, which has a basic requirement including a buffer of 7 percent, and then additional surcharges for the largest institutions. [. . .] So there are a number of questions of comparability. [. . .]	inter	inter / dir	[. . .] Bernanke explains the U.S. Basel III framework (7% plus surcharges), raises comparability issues with the European 9%, and explains the phased implementation timeline. [. . .]
M20120620_P003(Bernanke)	[. . .] Given this weaker outlook, why such a modest program? [. . .] does that mean you are prepared to do a full-on new asset purchase program?	[. . .] The incoming data were somewhat disappointing—but not entirely clear how to read them. [. . .] additional asset purchases would be among the things that we would certainly consider [. . .]	inter	inter / dir	[. . .] The second part is answered directly [. . .] he doesn't squarely explain why the program is modest relative to the deteriorated outlook.
M20120913_P004(Bernanke)	[. . .] you would reverse course if inflation were to be above your target level, even given that statement?	[. . .] We bring inflation back to the target over time, but we do it in a way that takes into account the deviations of both of our objectives [. . .]	inter	inter / dir	[. . .] Bernanke directly explains the balanced approach: bringing inflation back to target over time while considering deviations in both mandates. [. . .]

Rationale-audit evidence. The rationale-audit rule selects 21 of 389 intermediate-to-direct errors (570 pooled errors) using strict answer-shortfall cues in cached rationales. The audit accepts 17 instances across 12 pairs and six systems and rejects four. The supplement's `failure_analysis/` contains the original errors, metrics, examples, and recorded audit, including all 21 decisions, excerpts, pair locators, source hashes, and a membership/count checker. This targeted lexical screen provides a lower-bound diagnostic, not exhaustive inconsistency measurement or independent reference validation.

D.7 Model-reference diagnostics

The per-class breakdowns (Table 27), per-chair results (Tables 28 and 30), and error patterns (Table 29) use the full dataset (2,154 pairs) evaluated against the Condition D reference (Opus 4.6 with reasoning). These numerical values measure agreement with the Opus diagnostic reference, not accuracy against the human-majority labels. They accompany the ablation study in Section D.4; comparisons against an Opus reference can favour related model families. Tables 25, 26, and 31 describe agreement within or against the specified model panels, whose composition differs across analyses.

Table 27 shows per-class F1 scores. Against the Condition D reference, Eva-4B agrees least on direct (0.711) and fully_evasive (0.595), while Opus B scores 0.904/0.888/0.835. These are reference-agreement scores, not independent accuracy estimates.

Table 28 shows higher agreement with the Condition D reference across all Fed chairs, with the largest difference for Bernanke (+25.9 pp).

Table 29 quantifies error correction: across models, the dominant error under the EvasionBench prompt is misclassifying direct answers as intermediate. In total, Opus B reverses 494 Condition A disagreements with the reference, with a 3.7:1 ratio of reversals toward versus away from the reference.

Table 25: LLM panel validation: agreement across 6 model configurations (3 model families $\times$ 2 prompt conditions) on Rasiah 3-level classification. Complete-case panel agreement uses 2,153 of 2,154 pairs (one invalid Gemini-B output); this historical diagnostic excludes that output, whereas primary scoring retains invalid outputs as errors. We report nominal Krippendorff's α (Krippendorff, 2004). Opus, GPT and Gemini denote Opus 4.6, GPT-5.2 and Gemini 3 Pro. This panel is distinct from the three-family Condition-D labels in the frozen training split.

Panel Agreement (6 model configurations)	
Fleiss' κ	0.5237
Nominal Krippendorff's α	0.5237
Mean pairwise κ	0.5265
Model Configuration vs. Panel Consensus	
Configuration	κ
Opus B	0.857
Opus D	0.8376
GPT B	0.6874
GPT D	0.624
Gemini B	0.5923
Gemini D	0.5912
Eva-4B vs. Panel Consensus	
κ	0.4953
Macro-F1	0.6984

Table 26: Pairwise agreement between model configurations (Cohen's κ). Opus 4.6 Conditions D and B have the highest agreement among the displayed pairs ($\kappa = 0.80$).

	Opus D	GPT D	Gem. D	Opus B	GPT B	Gem. B	Eva-4B
Opus D	–	0.56	0.49	**0.80**	0.61	0.49	0.48
GPT D	0.56	–	0.52	0.50	0.59	0.43	0.50
Gem. D	0.49	0.52	–	0.48	0.44	0.46	0.36
Opus B	**0.80**	0.50	0.48	–	0.58	0.48	0.44
GPT B	0.61	0.59	0.44	0.58	–	0.48	0.52
Gem. B	0.49	0.43	0.46	0.48	0.48	–	0.42
Eva-4B	0.48	0.50	0.36	0.44	0.52	0.42	–

The T4 chair-partition diagnostic compares cached predictions against the full-corpus Condition D reference. It does not perform the retraining with one chair excluded that a held-out generalisation protocol would require.

D.8 Mechanism, tactic, and calibration diagnostics

The retained full-corpus mechanism, tactic, calibration, temporal, and chair analyses use their original model-derived references and are descriptive diagnostics. The auxiliary 1–5 evasion score also lacks separate human validation; higher scores indicate more responsive answers.

Table 32 describes model-assigned mechanism frequencies by chair: omission is most common for Yellen (49.7%), and vagueness for Bernanke and Powell ($\sim$46%). These frequencies do not identify speaker strategies or intentions.

Table 33 reports confidence calibration. Against the model-derived reference, Opus has lower expected calibration error (ECE = 0.054) than GPT and Gemini (0.188, 0.237; Figure 5). Calibration to this reference is not calibration to human judgments.

Table 27: Per-class F1 agreement with the Opus 4.6 Condition D reference, recomputed from retained prediction caches. These are model-reference diagnostics.

Configuration	Direct	Intermediate	Evasive	Macro-F1
Eva-4B	0.711	0.743	0.595	0.683
Opus 4.6 (A)	0.727	0.748	0.500	0.658
Opus 4.6 (B)	0.904	0.888	0.835	0.876
Opus 4.6 (C)	0.793	0.735	0.498	0.675
GPT 5.2 (A)	0.622	0.693	0.405	0.573
GPT 5.2 (B)	0.801	0.792	0.613	0.735
GPT 5.2 (C)	0.661	0.692	0.417	0.590
GPT 5.2 (D)	0.794	0.753	0.577	0.708
Gemini 3 Pro (A)	0.797	0.620	0.289	0.569
Gemini 3 Pro (B)	0.758	0.704	0.549	0.670
Gemini 3 Pro (C)	0.784	0.623	0.414	0.607
Gemini 3 Pro (D)	0.789	0.649	0.581	0.673

Valid predictions only: $n = 2154$ per configuration, except Gemini B (2153) and Eva-4B (2142). One Gemini B output and 12 invalid Eva outputs are excluded from these diagnostic F1 scores.

Table 28: Macro-F1 (agreement with Condition D reference) by Fed Chair.

Configuration	Bernanke	Yellen	Powell
Eva-4B	0.643	0.713	0.685
Opus B	0.902	0.811	0.884
GPT B	0.807	0.757	0.726
GPT D	0.775	0.771	0.691
Gemini B	0.674	0.813	0.651

D.9 Cross-domain diagnostics and reference sensitivity

T3 is a proposed cross-domain agreement task on a 2,000-pair earnings subset (1,000 Chinese, 1,000 S&P 500), which is absent from the release and requires third-party redistribution permission. The following comparisons use different corpora, reference sources, and evaluation populations. Their captions identify larger earnings-call pools and a Mandarin diagnostic sample; their record overlap with the proposed T3 subset is unverified, and they are not controlled transfer estimates.

Table 29: Changes in degree-label agreement with the Opus 4.6 Condition D reference when switching from prompt A to B, recomputed from retained caches. The first three rows count selected A disagreements that become B agreements; totals include all patterns.

Reference label	A prediction	Opus	GPT	Gemini
direct	intermediate	349	334	99
intermediate	fully_evasive	80	91	164
intermediate	direct	62	34	119
B corrects A		494	466	389
A corrects B		134	165	295

Common valid A/B prediction IDs: $n = 2154$ for Opus 4.6 and GPT-5.2, and $n = 2153$ for Gemini 3 Pro. "Corrects" denotes agreement with the model reference, not a human accuracy judgment.

Table 30: Chair-stratified Macro-F1 agreement with the Opus 4.6 Condition D reference, recomputed from retained caches. Parentheses give differences from each system's full-dataset score. Predictions are partitioned by chair; no held-out-chair retraining was performed.

System	**Bernanke** ($n = 266$)	**Yellen** ($n = 296$)	**Powell** ($n = 1592$)	**Full** ($n = 2154$)
Eva-4B	0.643 (-0.040)	0.713 (+0.030)	0.685 (+0.002)	0.683
Opus A	0.680 (+0.022)	0.667 (+0.009)	0.656 (-0.002)	0.658
Opus B	0.902 (+0.026)	0.811 (-0.065)	0.884 (+0.008)	0.876
GPT B	0.807 (+0.071)	0.757 (+0.021)	0.726 (-0.009)	0.735
Gemini B	0.674 (+0.004)	0.813 (+0.143)	0.651 (-0.020)	0.670

Valid predictions only. Header counts apply except Eva-4B ($265/296/1581$ by chair; 2142 full) and Gemini B (1591 Powell; 2153 full). These scores measure model-reference agreement.

Table 31: Leave-one-family-out reference sensitivity, recomputed from the retained A/B prediction files. The panel contains three families and two prompt conditions, distinct from the B/D panel in Table 25.

System	**Full F1**	**LOO F1**	Δ	**Voters**	N
Opus B	0.697	0.615	-0.082	4/6	2154
GPT-5.2 B	0.788	0.729	-0.059	4/6	2154
Gemini 3 B	0.720	0.660	-0.061	4/6	2153

The reference is plurality voting; ties use fixed input order (Opus A, Opus B, GPT A, GPT B, Gemini A, Gemini B, after family exclusions). Invalid predictions are omitted. Values replace archived results that did not exactly reproduce from the retained inputs; historical and recomputed values are recorded in the verification artifact. These are model-reference agreement scores, not estimates against human references.

Table 32: Distribution of model-assigned Bavelas mechanisms by Fed Chair among responses labeled non-direct by Opus 4.6 Condition D. The rare Restatement category (3 of 1,092 responses; 0.3%) is omitted. These descriptive label frequencies do not establish speakers' intentions.

Mechanism	Overall	Bernanke	Yellen	Powell
Vagueness	45.0%	46.7%	37.7%	46.1%
Omission	41.0%	36.2%	49.7%	40.1%
Non-Sequitur	13.7%	17.1%	11.9%	13.6%
N (non-direct)	1092	105	151	836

Table 33: Confidence calibration against the Opus 4.6 Condition D reference, recomputed from retained A/C outputs. Mean confidence exceeds reference accuracy for every configuration. ECE is the sample-weighted mean absolute difference between bin accuracy and bin confidence over ten equal-width bins on $[0, 1]$, with 1 included in the last bin. These values do not measure calibration to human judgments.

Model	Cond.	Accuracy	Mean conf.	ECE↓	Overconf.	N
Opus	A	0.728	0.779	**0.054**	+0.051	2154
Opus	C	0.752	0.826	0.078	+0.074	2154
GPT	A	0.650	0.838	0.188	+0.188	2154
GPT	C	0.665	0.851	0.189	+0.186	2154
Gemini	A	0.680	0.917	0.237	+0.237	2148
Gemini	C	0.698	0.916	0.218	+0.217	2154

Six Gemini-A records lack confidence and are excluded; all other confidence values are finite and in $[0, 1]$. The final bin includes 105 confidence-1 outputs for Gemini A and 109 for Gemini C.

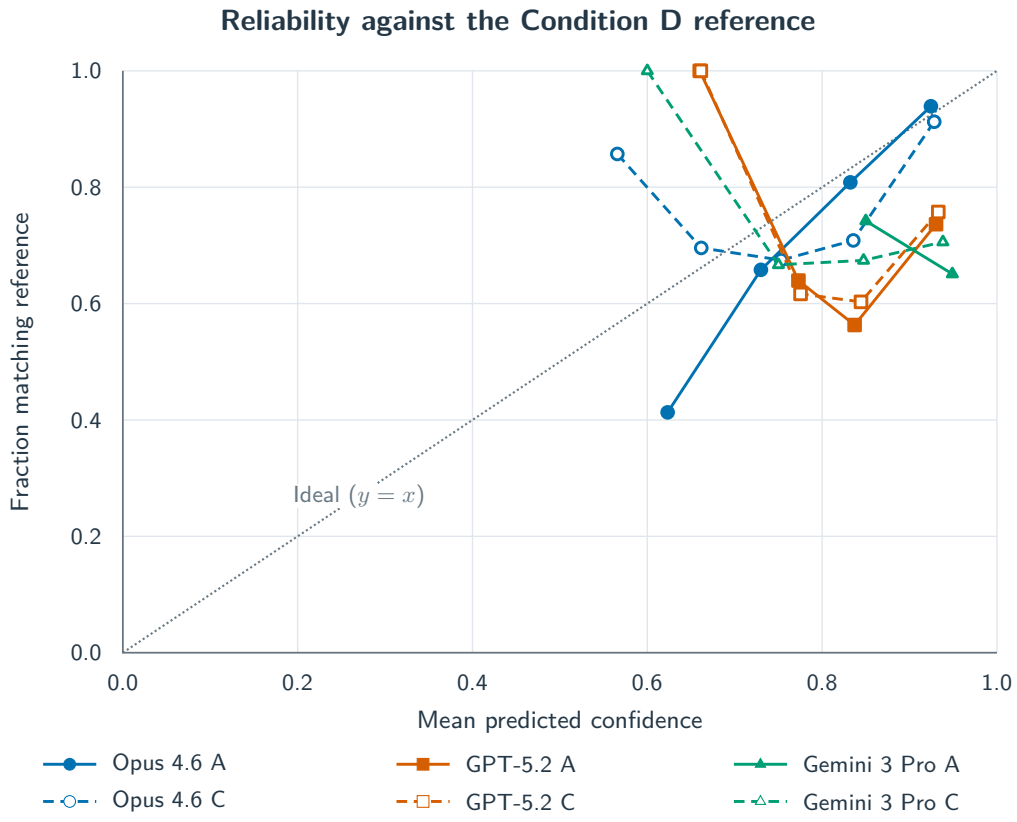


Figure 5: Reliability curves for Conditions A and C against the Opus diagnostic reference, recomputed from retained confidence outputs. Each point gives mean confidence and reference accuracy within a nonempty bin; lines connect points for readability. Ten equal-width bins cover $[0, 1]$, including confidence 1 in the last bin. The Opus curves use a reference from the same model family (Opus 4.6 D). Calibration to this model-derived reference is not calibration to human judgments.

Table 34: Task T2: Mechanism identification on non-direct pairs (agreement with Opus-derived Bavelas labels). Weighted-F1 [95% CI] with per-mechanism breakdown. The Opus 4.6 Condition-D reference has no independent human validation and may favour Opus-family predictions. Majority covers all 1,092 reference-non-direct pairs with mechanisms; Opus B is scored only on the 960 for which it also returns a non-direct label and valid mechanism. These different supports preclude a matched baseline comparison.

System	N	W-F1	Vague.	Omis.	Non-seq.	Restat.
Majority	1092	$0.279_{[0.25,0.31]}$	0.620	0.000	0.000	0.000
Opus B	960	$\mathbf{0.827}_{[0.80,0.85]}$	0.866	0.819	0.734	0.800

Table 35: The benchmark's 12 operational tactics adapt the earlier financial-Q&A scheme (Al Nuaimi et al., 2025b). These model-derived categories are assigned to non-direct responses alongside the mechanism field. The definitions below are specific to this benchmark, not an identical historical codebook.

Tactic	Definition
Ignore	Fails to address the question entirely
Acknowledge without answer	Acknowledges the question but provides no substantive response
Challenge premise	Questions the validity or framing of the question itself
Attack question	Criticises the question or questioner rather than responding
Refusal to answer	Explicitly declines to provide an answer
Agenda shifting	Redirects to a preferred topic or talking point
Claim ignorance	States lack of knowledge or information
Literal interpretation	Addresses an overly narrow reading of the question
Partial answer	Responds to only one part of a multi-part question
Repeat previous answer	Restates a prior response without adding information
External blame	Attributes responsibility to external factors or parties
Prepared remark injection	Inserts a pre-planned statement unrelated to the question

Table 36: Descriptive cross-corpus metrics on different reference populations. FOMC scores measure agreement with the Opus Condition-D reference. The earnings-call (EC) metric is a source-reported benchmark result for Eva-4B and cross-system agreement for the Opus row. The 15,828-pair agreement population differs from the 16,726-record and 5,574,452-pair corpus summaries elsewhere; their exact record overlap is unverified and their totals or label shares are not interchangeable. None is the proposed 2,000-pair T3 population or a controlled cross-domain accuracy comparison.

	FOMC (2,154)		Earnings Calls	
Method	F1	κ	EC metric	N
Eva-4B (fine-tuned)	0.683	0.485	0.849[†]	1,000
Our prompt + Opus 4.6	**0.876**	**0.799**	0.624	15,828
Our prompt + GPT 5.2	0.735	0.606	—[§]	—
Our prompt + Gemini 3 Pro	0.670	0.490	—[§]	—

[†]Source-reported result from Ma et al. (2026), not cross-system agreement and not directly comparable to the next row. [‡]Cross-system agreement (F1 between our labels and Eva-4B labels, $\kappa = 0.519$); no shared reference exists for that paired earnings-call population, so the Opus earnings-call entry measures inter-system agreement, not accuracy. Our prompt classifies 65.5% as direct vs. Eva-4B's 52.3%. [§]Not evaluated on earnings calls.

Table 37: Cross-domain comparison: FOMC press conferences vs. corporate earnings calls. FOMC label percentages use the Opus 4.6 Condition-D diagnostic reference, not the current human-majority degree labels. FOMC answer length is recomputed from released text using whitespace tokens; earnings figures retain their historical analysis source.

Metric	FOMC	Earnings Calls
Q&A pairs	2,154	5,574,452
Transcripts	89	297,773
Median answer length	189.5	107
Direct (%)	49.3	54.1
Intermediate (%)	48.1	37.3
Fully evasive (%)	2.6	8.6
Vagueness (%)	45.0	78.0
Omission (%)	41.0	13.8
Non-Sequitur (%)	13.7	7.8
Restatement (%)	0.3	0.4

Table 38: Evasion label distribution across domains. FOMC uses the Opus 4.6 Condition-D diagnostic reference; the current human-majority degree distribution is reported separately in Table 6.

Label	FOMC	EC–EvasionBench	EC–Ours
direct	49.3%	52.3%	54.1%
intermediate	48.1%	44.0%	37.3%
fully_evasive	2.6%	3.7%	8.6%
N	2,154	16,726	5,574,452

Table 39: Appendix-only Mandarin prompt-sensitivity sanity check on $\sim$10K Q&A pairs, whose record overlap with the proposed 1,000-pair Chinese component of T3 is unverified. Per-pair prompt (A) vs. meeting-level prompt (B) using the same model (Opus 4.6). One pair excluded from Condition A due to a parse error ($N_A = 9{,}999$; $N_B = 10{,}000$). The intermediate shares are close: 52.2% under A and 51.8% under B.

	Per-class F1			
Metric	Direct	Intermediate	Evasive	Macro-F1
A vs. B agreement	0.861	0.859	0.802	0.841
	Label distribution (%)			
	Direct	Intermediate	Evasive	N
Cond. A (per-pair)	36.4	52.2	11.4	9,999
Cond. B (meeting)	36.0	51.8	12.1	10,000
	Directional label disagreement			
	$P(\text{A=inter} \mid \text{B=direct}) = 13.4\%$			

E Temporal Dynamics

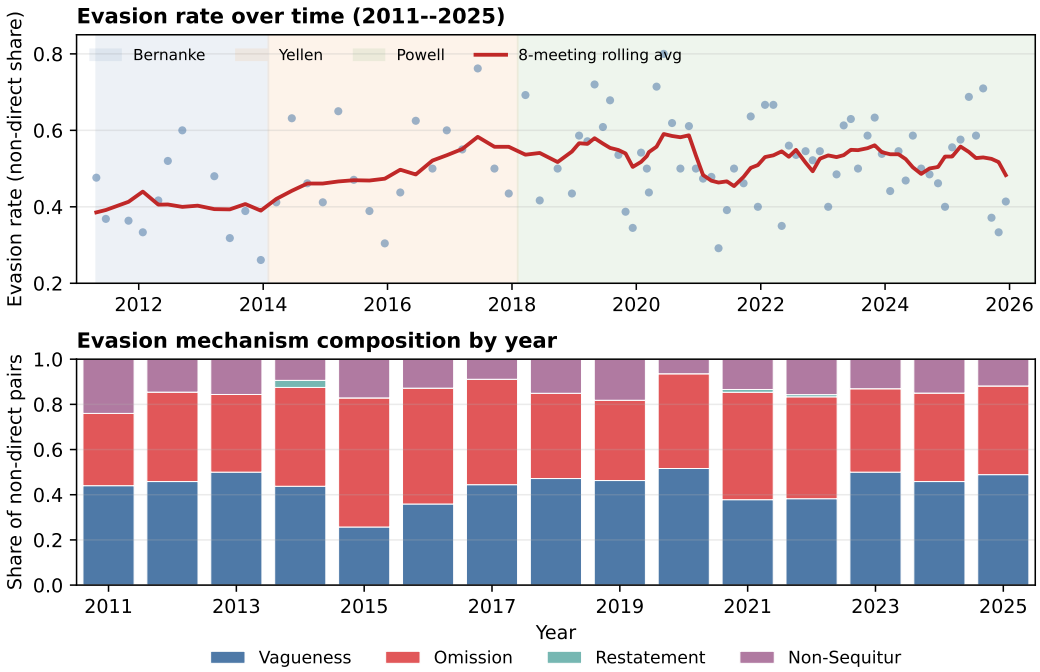


Figure 6: **Model-derived non-direct share and mechanism composition from Opus 4.6 Condition D (2011–2025). (Top)** Meeting-level model-derived non-direct share; the graphic labels this share as evasion rate. The red line is the 8-meeting rolling average; shaded regions mark chair tenures. The rolling curve describes changes in the model-assigned non-direct share; event-specific causes are not identified. **(Bottom)** Annual model-derived Bavelas mechanism composition among pairs labelled non-direct by Opus 4.6 D: vagueness and omission dominate (∼85% combined share in most years); non-sequitur is near 10%; restatement is rare. These associations are descriptive and confounded by topic and period; they do not measure changes in honesty or intent.

F Detailed Ablation Results

The comparisons in this section use the model-derived Condition D reference unless a table identifies another reference population. They measure configuration agreement with that diagnostic reference, not human-reference accuracy or isolated causal effects of prompting, reasoning, or context. The ablation scores and paired contrasts are computed from the identified prediction files. These diagnostic F1 scores condition on valid outputs: 2,154 pairs per configuration except Gemini B (2,153) and Eva-4B (2,142). Paired contrasts use their common valid pairs; this differs from the full-split invalid-as-error scoring of the primary human-reference benchmark.

F.1 Prompt, context, and reasoning configurations

The G/H/I labels in this subsection refer to the API ablation family (Appendix D.1), not the open-weight benchmark prompt. G uses per-pair context with the project's three-axis framework; H uses meeting-level context with degree-only (Rasiah) instructions; I uses the project's three-axis framework with meeting-level Q&A records in shuffled order. Their exact configuration descriptions are retained with the source runs; full wire/native equivalence is unverified.

Table 42 reports agreement with the Condition D reference within model-assigned categories. Eva-4B matches 61.6% of direct reference labels, 84.1% of vagueness labels, and 81.9% of omission labels; Opus B matches 92.7% of direct labels. The systems differ in model, prompt, taxonomy, and context,

Table 40: Prompt and reasoning ablations: three-way Macro-F1 and nominal κ agreement with the Opus 4.6 Condition D reference, recomputed from retained caches. A uses the EvasionBench per-pair prompt; B uses our meeting-level prompt; A and B disable reasoning, while C adds reasoning to the EvasionBench prompt. Condition D supplies the reference.

Model	Condition	Macro-F1	95% CI	κ	n
Opus 4.6	A	0.658	[0.624, 0.690]	0.499	2154
Opus 4.6	B	0.876	[0.845, 0.902]	0.799	2154
Opus 4.6	C	0.675	[0.642, 0.706]	0.544	2154
GPT-5.2	A	0.573	[0.542, 0.604]	0.369	2154
GPT-5.2	B	0.735	[0.699, 0.767]	0.606	2154
GPT-5.2	C	0.590	[0.559, 0.621]	0.393	2154
Gemini 3 Pro	A	0.569	[0.542, 0.593]	0.449	2154
Gemini 3 Pro	B	0.670	[0.636, 0.702]	0.490	2153
Gemini 3 Pro	C	0.607	[0.576, 0.636]	0.454	2154
Eva-4B	—	0.683	[0.647, 0.716]	0.485	2142

Valid predictions only: one Gemini B output and 12 Eva outputs are excluded. Pointwise intervals use 10,000 QA-pair bootstrap resamples, RandomState seed 42, and linear 2.5th/97.5th percentiles. These fixed-cache diagnostics measure model-reference agreement.

Table 41: Prompt-configuration contrasts against the model-derived Opus Condition D reference. Values are Macro-F1 point estimates with pointwise 95% bootstrap intervals. Parentheses report differences from Condition B in percentage points. The intervals describe individual scores, not paired differences; no significance claim about Macro-F1 differences is made. The configurations change prompt format, available context, taxonomy instructions, or exchange order.

Cond.	**Configuration**	**Opus 4.6 Macro-F1**	**GPT-5.2 Macro-F1**	**Gemini 3 Pro Macro-F1**
B	Meeting-level, full framework	$0.876_{[0.84,0.90]}$	$0.735_{[0.70,0.77]}$	$0.670_{[0.64,0.70]}$
G	Per-pair, full framework	$0.708_{[0.67,0.74]}$ (-16.8)	$0.676_{[0.64,0.71]}$ (-5.9)	$0.717_{[0.68,0.75]}$ (+4.7)
H	Meeting-level, Rasiah-only	$0.790_{[0.76,0.82]}$ (-8.5)	$0.659_{[0.63,0.69]}$ (-7.6)	$0.592_{[0.56,0.62]}$ (-7.8)
I	Meeting-level, shuffled order	$0.810_{[0.77,0.84]}$ (-6.5)	$0.753_{[0.72,0.78]}$ (+1.8)	$0.658_{[0.62,0.69]}$ (-1.3)

so these differences do not isolate which component accounts for agreement with the Opus-family reference.

Agreement also varies by meeting size (Figure 7). On meetings with more than 28 pairs, Opus B reaches 0.914 Macro-F1 against the reference, 0.246 above Eva-4B. The observed difference is +6.0 pp for small meetings and +24.6 pp for very large ones (Table 43). Meeting-size groups differ in content and period; this comparison does not establish that adding context improves human-reference performance.

Reasoning-enabled and non-reasoning configurations produce different class-level agreement patterns on their common valid pairs. The Opus C−A contrast gives +16.7 pp recall for reference-direct labels; the GPT and Gemini D−B contrasts give −9.2/−15.8 pp recall for reference-intermediate labels. Recall is the fraction of examples in a reference class predicted as that class. These examples exclude comparisons with the reference configuration itself. The configurations also differ in generation settings; equality of effective sampling parameters is not established. We therefore report these as configuration contrasts against a model-derived reference, with retrospective pair-level test calculations in Table 44. Those calculations do not establish meeting-aware or post-selection error control, or a general causal benefit of reasoning.

Prompt-configuration comparisons. Table 41 reports Macro-F1 point estimates and pointwise intervals against the model-derived Condition D reference. The stored paired-bootstrap tests compare per-pair correctness indicators, testing accuracy differences rather than Macro-F1 differences; they therefore do not supply significance markers for the displayed Macro-F1 changes. The reported Cohen's d values for the comparison with Condition G are likewise computed from correctness

Table 42: Degree-label agreement with the Opus 4.6 Condition D reference within four of its model-assigned categories, recomputed from retained caches. Direct is a degree label; the other categories are Bavelas mechanisms. Restatement has three reference instances and is omitted. These categories have no separate human validation.

Config	Direct *(n=1062)*	Vague. *(n=491)*	Omiss. *(n=448)*	Non-Seq. *(n=150)*
Eva-4B	0.616	0.841	0.819	0.760
Opus A	0.617	0.851	0.844	0.773
Opus B	0.927	0.855	0.850	0.927
GPT A	0.480	0.845	0.828	0.687
GPT B	0.753	0.855	0.812	0.760
Gemini A	0.812	0.558	0.609	0.367
Gemini B	0.757	0.715	0.680	0.647

Header counts are the reference-category sizes and scoring denominators, except Gemini B Omission ($n = 447$; one output is excluded). Eva-4B's unknown outputs remain disagreements in the full category denominators. All three A caches contain valid predictions for all 2,154 pairs.

Table 43: Descriptive agreement with the Condition-D model reference, grouped by meeting size. Opus B's Macro-F1 gap over Eva-4B rises from +6.0 pp to +24.6 pp across the displayed bins; Eva-4B's scores are non-monotonic. The fraction of Eva-4B reference disagreements on which Opus B matches the reference is highest for the large bin (92%), followed by the very-large bin (89%). These associations do not isolate a causal effect of context length.

Meeting size	**Pairs**	**Mtgs**	**Eva-4B**	**Opus B**	**Δ**	**Fix rate**
Small (≤20)	436	25	0.732	0.792	+0.060	76%
Medium (21–25)	721	31	0.675	0.858	+0.184	83%
Large (26–28)	243	9	0.647	0.890	+0.243	92%
Very large (>28)	754	24	0.667	0.914	+0.246	89%
Overall	2,154	89	0.683	0.876	+0.193	—

indicators: 0.32 (Opus), 0.16 (GPT), and −0.10 (Gemini). Condition G has a higher Gemini Macro-F1 point estimate against this reference. The configurations also change prompt format, taxonomy instructions, or exchange ordering; these comparisons do not establish a general causal effect of discourse context. Configured output budgets do not measure context-window sizes; Appendix J distinguishes saved settings from unverified cached requests.

F.2 Replicate stability and efficiency

The retained analysis reports three replicate scores per model for reasoning-enabled Conditions C and D (Table 45). Its script evaluates model-consensus degree references, rather than the Opus diagnostic reference used for the ablation tables. The recorded sample standard deviations range from 0.006 to 0.047. Original replicate outputs and input manifests are not available, so their per-run populations and independent execution cannot be verified. These summaries do not establish robustness to inference variation; paired comparisons elsewhere condition on their identified cached predictions.

F.3 Training-label consensus variants

Table 47 distinguishes the model-consensus Dev selection criterion, human-reference Test-Std performance, and model-derived training labels. The compared supervised systems were trained or selected using those documented model-label variants, not the human references.

Table 44: Exploratory McNemar tests of paired agreement with the Condition-D Opus reference. Entries are Holm-adjusted p-values over 24 tests; parentheses give aligned pair counts. The tests treat pairs as independent and do not account for shared meetings; they do not establish meeting-generalization effects.

Comparison	Opus 4.6	GPT-5.2	Gemini 3 Pro
B vs A	$< 10^{-4}$ (2154)	$< 10^{-4}$ (2154)	0.005 (2153)
D vs C	$< 10^{-4}$ (2154)†	$< 10^{-4}$ (2154)	0.031 (2154)
D vs B	$< 10^{-4}$ (2154)†	0.221 (2154)	1.000 (2153)
C vs A	0.095 (2154)	0.481 (2154)	0.357 (2154)
B vs Eva-4B	$< 10^{-4}$ (2154)	$< 10^{-4}$ (2154)	1.000 (2153)
D vs Eva-4B	$< 10^{-4}$ (2154)†	0.0005 (2154)	1.000 (2154)
A vs Eva-4B	1.000 (2154)	$< 10^{-4}$ (2154)	0.013 (2154)
C vs Eva-4B	0.030 (2154)	$< 10^{-4}$ (2154)	0.481 (2154)

†The left system is the reference itself; these comparisons are not independent evidence of superiority. The tests use continuity-corrected χ^2_1 McNemar statistics. These are two-sided differences: a small p does not imply that the left system is better. One invalid Gemini-B output is excluded by the original condition loader; Eva-4B invalid outputs are retained as errors. This historical loader policy differs from the current primary scorer, which retains every invalid output as an error. These are descriptive model-reference agreement tests, not tests against human references. The family and correction were reconstructed retrospectively.

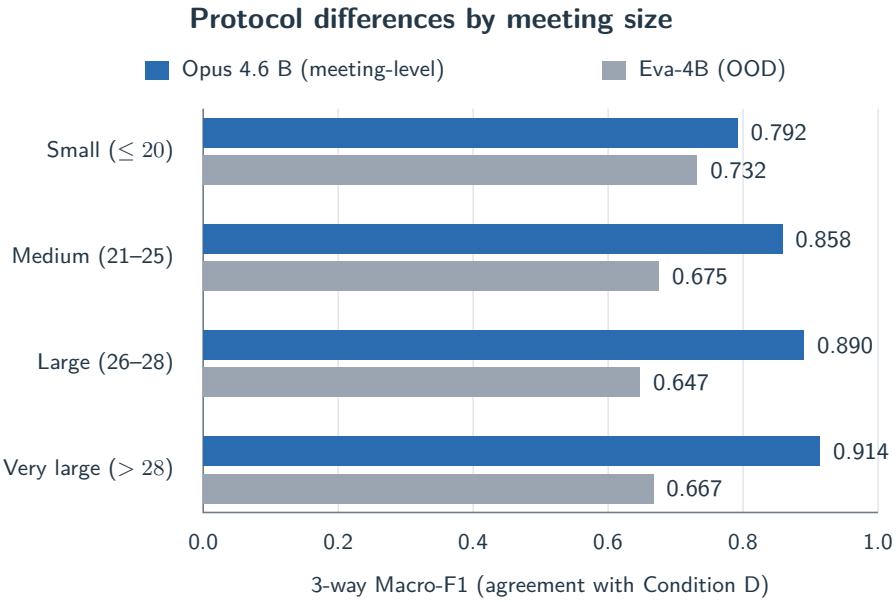


Figure 7: Macro-F1 agreement with the Opus 4.6 Condition D reference by meeting size. Opus B shares the reference's model family and meeting-level input. These groups differ in content and period as well as size, and the compared systems differ in model and prompting; the pattern does not identify a causal effect of additional context.

Table 45: Retained historical summaries for reasoning-enabled Conditions C and D: mean±sample SD over three recorded Macro-F1 scores. The analysis script uses model-consensus degree references, distinct from the human test references and the single-model Opus diagnostic reference. Original replicate outputs and input manifests are not available; historical per-run populations and independent execution remain unverified.

Model	Cond. C	Cond. D
Opus 4.6	0.732 ± 0.009	0.812 ± 0.018
GPT-5.2	0.643 ± 0.006	0.837 ± 0.028
Gemini 3 Pro	0.687 ± 0.006	0.787 ± 0.047

Table 46: Recorded inference requests and token counts: 89 meeting-level requests versus 2,154 per-pair requests. Macro-F1 measures agreement with the Opus 4.6 Condition D reference. Baseline token counts are unavailable, so total token usage cannot be compared.

Method	Requests	Input Tokens	Output Tokens	Macro-F1
Eva-4B (per-pair)	2,154	—	—	0.683
Our prompt + Opus 4.6	89	993K	433K	0.876
Our prompt + GPT 5.2	89	927K	323K	0.735
Our prompt + Gemini 3 Pro	89	∼1.0M	414K	0.670

Table 47: Consensus-variant summaries for DeBERTa-v3-base trained on model-labelled Train with eight label strategies. Scores are three-class F1 on the historical model-consensus Dev references and human-reference Test-Std. "All 15" has the highest Dev point estimate; "reasoning_only" has the highest Test-Std point estimate, without an interval. These archived summaries do not establish superiority or justify selection on test scores; the checkpoint and per-pair predictions are unavailable (Appendix J). The retained model-labelled Train reference follows the three-family Condition-D consensus rather than all_15. G in this table is the API framework condition defined in Appendix D.1.

Variant	Description	Dev F1	Test F1
all_15	A/B/C/D/G × 3 families	**0.739**	0.599
reasoning_only	D × 3 families	0.710	**0.605**
perpair_their	A × 3 families	0.702	0.534
mmc6_bd	B+D × 3 families (6 votes)	0.473	0.424
meeting_only	B × 3 families	0.471	0.419
gpt_only	GPT-5.2, B+D	0.478	0.384
opus_only	Opus 4.6, B+D	0.466	0.389
perpair_ours	G × 3 families	0.386	0.409

G Dataset Documentation

Relationship to prior resources. The earlier FOMC study (Al Nuaimi et al., 2025a) reports 1,728 pairs from 83 conferences and the same degree/mechanism/tactic framework. The related FinNLP paper (Al Nuaimi et al., 2025b) studies earnings-call Q&A and treats FOMC transfer as future work; it is a taxonomy and task predecessor rather than an earlier FOMC recordset. The FOMC predecessor is a non-archival workshop paper; the FinNLP study is published in workshop proceedings. Table 48 separates inherited components from the present benchmark's evaluation and release design. The current 12-tactic definitions are specified explicitly rather than assumed to be an identical historical codebook. A publication-bound pair/meeting manifest for the earlier FOMC release is not available; exact inherited, added, removed, or resegmented records therefore remain unverified. In particular, subtracting the published corpus totals does not establish 426 new examples or the identities of six additional meetings.

Informed by the datasheet framework of Gebru et al. (2021) and the data statement guidelines of Bender & Friedman (2018), we document FOMCBench below.

Auxiliary category coverage. The released Train fields contain no Attack Question, External Blame, or Prepared Remark Injection labels; Ignore and Literal Interpretation occur once each, and Restatement occurs twice among mechanisms. These are counts of the retained Opus-D auxiliary fields, whose model-assigned degree can differ from the current human-majority degree. The schema supplies diagnostic vocabulary; its presence does not establish sufficient examples for supervised learning of every category or validate any category against human annotations.

Dataset summary. FOMCBench contains 2,154 question–answer pairs from 89 FOMC press conferences (April 2011–December 2025). Each pair includes: the reporter's question text, the Fed chair's answer text, reporter name and media organisation, multi-taxonomy evasion labels (Rasiah degree, Bavelas mechanism, Bull tactic), and a 1–5 evasion score, where higher values indicate more responsive answers.

Source and collection. Transcripts are official Federal Reserve PDFs. Reporter organisations were extracted via LLM with cross-meeting gap filling (99.6% coverage of reporter–meeting records). Audio timestamps derive from Whisper-based transcription aligned to official text via fuzzy matching.

Preprocessing. Transcripts were parsed, cleaned (1,098 non-substantive speech removals, 43 merges), and aligned with zero overlapping timestamps. Speaker names were normalised using 27 rules covering nicknames, spelling corrections, and PDF typos. All 89 meetings maintain question–answer count parity after cleaning.

Cuts. The primary release contains 2,154 Q&A pairs with evasion annotations. Per-speech (4,397 rows) and per-section (178 rows) representations were used in construction, but are absent from the Q&A supplement. The manuscript does not count them as released benchmark cuts.

Known biases. The dataset is English-only and U.S.-centric. Jerome Powell chairs 61 of 89 conferences, creating a recency bias. Degree references derive from three human judgments per pair across the corpus (Appendix B.2). The 400 test triplets are 73% unanimous with three-class Fleiss' $\kappa = 0.676$; full-corpus agreement on the retained ratings before professional review is 0.7911. Train/Dev selection and tuning exposure limit interpretation.

All Bavelas, Bull, and score fields remain model-derived from Opus 4.6 under Condition D and lack independent human validation. Scores are integers from 1 to 5, with higher values indicating more responsive answers. For the diagnostic degree view—model-consensus Train/Dev labels and human test labels—mapping scores 4–5 to direct and 1–3 to non-direct disagrees on 256/2,154 pairs (11.9%): 153 direct labels have scores below 4 and 103 non-direct labels have scores above 3. This mixed-source consistency diagnostic is distinct from comparison with the full human-reference set. The recovered prompt constrains model-direct outputs to scores 4–5 but does not impose the converse; the auxiliary score is not a substitute for the human degree reference.

Table 48: Relationship to the earlier FOMC study (Al Nuaimi et al., 2025a). "Not reported" concerns the inspected publication, not the possible existence of unpublished artifacts. Differences in corpus totals do not establish counts of newly added records.

Component	Earlier FOMC study	Current benchmark
Coverage	83 conferences; 1,728 Q&A pairs; three Fed chairs	89 conferences; 2,154 Q&A pairs; same institution and chairs
Framework	Degree, mechanism and tactic axes	Inherited axes; current operational tactic definitions in Table 35
Model references	Claude 4 Sonnet annotation	Original three-family labels retained for lineage and historical analyses; auxiliary fields remain model-derived
Splits	Present four-way protocol not reported	Fixed Train/Dev/Test-Std/Test-Challenge: 1,554/200/200/200
Human references	Stratified human validation reported without its sample size or detailed agreement	Three professional degree judgments for all 2,154 pairs; original 400 test references unchanged; historical rater-identity limits documented
Release	Annotations and prompts linked through Kaggle	Offline scorer, cached predictions, checksums, documentation and recovered B/D prompt
Analysis	Linguistic comparisons and market associations	Human–model reference audit, eight open-weight G/B evaluations, historical API panel, uncertainty and failure analysis

Artifact contents and licensing. The release provides all 2,154 processed Q&A texts, fixed splits, three human votes per item, and complete binary/three-class majority references. Split files are under `data/splits/`, with `all.jsonl` containing the complete corpus and `test.jsonl` the combined test convenience view. Anonymous votes are in `data/human_annotations/votes.jsonl`; versioned annotation metadata and change records are retained alongside them. Model-consensus Train/Dev labels are provided under `data/legacy_model_references/` for explicitly sourced comparisons. Private annotator exports, notes, and journals are omitted from the anonymous package. The release also includes cached predictions, model prompts, documentation, reproducible scoring/statistics code, and file checksums. Data are licensed under CC BY 4.0 and code under MIT.

Release limits. The source-text erratum in Appendix A.1 accompanies the frozen records; corrected transcripts would require reassessment of labels and dependent inputs. Cached API results support reproducible score calculations, but the package does not reconstruct complete inference or training pipelines. Recovered B/D prompts and additional G/H caches cover their documented analyses; exact request equivalence for every API configuration is unverified. Other corpus cuts, the proposed earnings-call subset, and multimodal media are absent. No permanent DOI or completed public archival deposit is claimed.

Use and maintenance scope. The benchmark measures discourse-level responsiveness. Its labels do not establish deception, private beliefs, motives, speaker credibility, or institutional trustworthiness. New conferences, changed taxonomies, and corrected labels would require a separately versioned artifact; the present analyses retain the existing pairs and split membership. No fixed refresh cadence, private held-out service, or maintenance response time is established by the current release.

Contamination. The transcripts are public and can appear in model training data. Pair dates permit model-specific temporal analyses only when a model's training cutoff is known; they do not certify absence of exposure. Neither a public text search nor the fact that classification uses a rubric establishes decontamination, and no model in this study is certified uncontaminated.

Sample size and benchmark refresh. Under a simulation's gold-tilt disagreement scenario, approximately 930 independent pairs were estimated to provide 80% power for a five-point binary Macro-F1 difference. This scenario-specific estimate is interpolated from the simulated power curve

and concerns one comparison without multiplicity adjustment. This exceeds the 200 pairs per test split. The benchmark therefore supports a validity and failure-mode study more directly than fine-grained annual progress tracking.

A fully new, post-cutoff slice would accumulate slowly: the observed 24.2 usable pairs per conference imply about 194 pairs annually at eight conferences per year. The resulting multi-year requirement applies to an independently refreshed, adequately powered small-gap comparison, not to the present analyses. New labels, rare-class enrichment, and controlled cross-institution comparisons remain possible extensions.

H Illustrative Examples

Table 49: **Traceable examples of agreement and disagreement.** Questions are abridged; answer excerpts retain the recorded wording, and pair IDs identify the complete Q&A records. All displayed references are the benchmark's final human-majority degrees; split names identify their evaluation populations. Eva-4B denotes the out-of-domain checkpoint, and Opus B denotes Opus 4.6 under Condition B.

Q&A context	Degree labels
Powell (2018) `M20180321_P014` *Question (abridged):* What trade-policy developments would concern the Committee's outlook? What is the path of productivity growth, and have recent policy changes altered that assessment? *Answer excerpt:* "…more widespread retaliation and more widespread actions back-and-forth kind of thing. I can't be any more specific than that. In terms of productivity…productivity, as I'm sure you know, has been very weak since the financial crisis."	**Train** Reference: direct Eva-4B: intermediate Opus B: direct
Yellen (2014) `M20140618_P016` *Question (abridged):* What wage growth is expected this year and next? Would inflation outpacing wages delay a rate increase, and would wages merely matching inflation be sufficient? *Answer excerpt:* "I see compensation growth, broadly speaking, as having been very well contained… My own expectation is that as the labor market begins to tighten, we will see wage growth pick up".	**Train** Reference: intermediate Eva-4B: intermediate Opus B: direct
Bernanke (2011) `M20110427_P011` *Question (abridged):* What positive effects can be attributed to the second round of quantitative easing, and can it end in June with unemployment around 9%? *Answer excerpt:* "We saw increases in stock prices. We saw reduced spreads in credit markets. We saw reduced volatility… The tradeoffs are getting…are getting less attractive at this point. Inflation has gotten higher."	**Dev** Reference: direct Eva-4B: intermediate Opus B: direct
Bernanke (2011) `M20110622_P013` *Question (abridged):* Which policy path would achieve the projected 2013 conditions: an early gradual exit, a late steep exit, or no exit? *Answer excerpt:* "I don't think it would be constructive for me to give you a tentative projection… there's simply no alternative but to watch the incoming data".	**Test-Challenge** Reference: fully evasive Eva-4B: fully evasive Opus B: fully evasive

Table 49 provides the question context and current human reference for four retained prediction comparisons. For the Powell and Bernanke QE2 examples, Opus B agrees with the human-majority direct reference, whereas Eva-4B predicts intermediate. The Yellen example reverses this pattern: Eva-4B agrees with the human-majority intermediate reference, whereas Opus B predicts direct. These three cases illustrate disagreement at the direct/intermediate boundary rather than a uniform advantage for either system.

For the Bernanke exit-projection example, both systems predict fully evasive, matching the recorded Test-Challenge degree reference. The corresponding *Omission* and *Refusal to Answer* fields are model-derived diagnostic annotations, not human-validated mechanism or tactic labels. The Eva-4B cache records degree outputs without rationales for these cases; the excerpts do not establish which textual features caused its predictions.

I PROMPT TEMPLATE

The following schematic records the meeting-level prompt structure. The auxiliary 1–5 evasion score runs from less to more responsive answers. Bracketed fields are placeholders, not the complete annotation rubric or executable prompt:

```
You are a financial discourse classification expert
performing machine-labelling of FOMC press conference Q&A
exchanges.

Below is the COMPLETE Q&A section of the FOMC press
conference on [DATE], chaired by [CHAIR]. Read the entire
transcript first to understand the full context.  Then
classify EACH exchange individually using the
Rasiah-Bavelas-Bull evasion framework.

[CLASSIFICATION FRAMEWORK: multi-taxonomy definitions,
constraints]

[6 FIXED FEW-SHOT EXAMPLES (covering all three Rasiah
degrees and common Bavelas mechanisms)]

[FULL Q&A TRANSCRIPT (N exchanges)]

TASK: Return a JSON array with exactly N objects, one per
exchange, in order.  Keys:  pair_id, rasiah, bavelas,
subtype, evasion_score, rationale, evidence_spans.
```

The supplement includes the recovered framework, six fixed few-shot examples, and meeting-prompt assembly in `prompts/meeting_condition_B_D_template.txt` and `prompts/evasion_prompt_source.py`. The excerpt above remains a schematic. These files preserve the authored prompt text for Conditions B and D; they do not establish coverage of every later model configuration or every ablation condition. The package README documents citation-string errors and attribution limits in these preserved inputs. The prompt's word "complete" describes the supplied processed records, not independent verification of source-transcript completeness (Appendix A.1).

J REPRODUCIBILITY DETAILS

Table 50 records the three-family API prompt-ablation settings. The open-weight panel has separate source-bound profiles in Appendix C. API calls used a LiteLLM proxy with an OpenAI-compatible interface; a routing layer does not fix vendor-side model revisions. Fine-tuning recipes are reported separately in Table 52.

Table 50: Recorded configurations for the three-family API prompt ablations. Token values are configured output limits.

Parameter	Value
Consensus label models	Claude Opus 4.6, GPT-5.2, Gemini 3 Pro
Ablation models	Claude Opus 4.6, GPT-5.2 (OpenAI), Gemini 3 Pro (Google)
Fine-tuned baseline	Eva-4B (Qwen3-4B, EvasionBench)
API routing	LiteLLM proxy (unified OpenAI-compatible endpoint)
Temperature evidence	B/D saved configs: unspecified; A/C driver settings: see text
Output-budget evidence	B/D saved configs: 64,000; A/C run-specific limits unrecorded
Reasoning effort	`high` (Conditions C, D); `none` (Conditions A, B)
Prompt format	Meeting-level JSON (B, D); per-pair JSON (A, C)
Label constraint	`rasiah=direct` $\Rightarrow$ `bavelas="--"`, `subtype="--"`, score ≥ 4

The saved B/D configurations leave temperature unspecified; the client omits that request parameter. The recovered A/C driver requests 0.1 for A and, for C, 1.0 for Opus and 0.1 for GPT/Gemini. A/C run summaries omit temperatures and output limits, so the driver alone does not authenticate every cached request. API defaults and provider-side behaviour remain unverified. The retained three-score

summaries in Table 45 describe a separate run-variation analysis with incomplete input provenance. Bootstrap intervals condition on the cached runs and do not estimate variation across API calls, training seeds, or provider updates.

Scoring human-majority references. The accompanying offline scorer evaluates T1b (binary) by default and T1 with an explicit three-way option. From the supplement root, the following command scores a packaged prediction file:

```
PYTHONPATH=code/src python3 -m fomcbench \
  --predictions predictions/claude_opus_4_7_condition_B.jsonl \
  --split test_std --task binary
```

Use `--task threeway` for T1. Binary scoring maps intermediate and fully_evasive to non-direct; three-way scoring retains all three fixed classes. Missing or invalid model outputs count as errors on the complete reference split rather than being silently dropped. Table 22 separately identifies the earlier valid-output-only sensitivity calculation. The package README gives exact commands. Re-scoring cached outputs does not rerun inference or recover missing model artifacts.

For a full-corpus diagnostic, `--split all` uses all 2,154 final human references under either `--task binary` or `--task threeway`; no current reference requires tie exclusion. Missing or invalid predictions on this population still count as errors. Train/Dev selection and tuning exposure limit full-corpus predictive analyses to descriptive audits.

Reproducing the main cached API G/B comparison. The three API systems' fixed-test records and source policy are bound by `analysis/closed_api_gb/manifest.json`; expected scores, paired differences, and method metadata are in the same directory. G predictions are stored under `analysis/extended_protocol_predictions/`, and B predictions under `predictions/`. These are frozen exported degrees, separate from the open-weight required-field recovery and adaptive views. Reproduce all 24 arm scores and 12 paired contrasts from the package root with:

```
python3 -B code/reproduce_closed_api_gb.py \
  --output-dir ../closed-api-points --bootstrap
```

Omitting `--bootstrap` reproduces point estimates only. The output directory must be new and outside the package. The same released scorer, fixed reference classes, complete 200-item denominators, and shared whole-meeting draws apply to both main-table groups; exact cached API wire inputs and native equivalence remain unverified. The additional comparisons below retain their own documented resampling designs.

Reproducing additional API and supervised-baseline uncertainty estimates. With NumPy installed, `code/reproduce_statistics.py` reproduces binary and three-way point estimates and 95% intervals for the ten cached systems on both test splits, the Test-Std binary Opus 4.7 versus GPT-5.2 comparison, three Test-Std protocol comparisons using the additional Condition-A caches, and the four pooled-test comparisons in the lower block of Table 23. Run `python3 code/reproduce_statistics.py --output-dir ../statistics-output` from the supplement root. The script preserves 10,000 paired resamples, seed 42, fixed target classes, invalid-as-error scoring, percentile intervals, the sign-tail test and Holm correction. `analysis/README.md` documents the conventions, expected results and coverage limits. It does not regenerate the full-corpus diagnostic intervals or the 24-test McNemar family, and pair resampling does not capture meeting dependence, training variation or reference-label uncertainty.

The additional `code/reproduce_protocol_sensitivity.py` reports both targets on both test splits for all three A/B comparisons, with a joint retrospective 12-test Holm family (Table 17). It also reports the separate Opus 4.6 A versus Opus 4.7 B binary comparison as exploratory. The original statistics output remains available for the earlier three-test family; its adjusted values should not be confused with those from the expanded family.

Reproducing full-corpus human-reference statistics. The release separately identifies the retained-rating snapshot used for agreement and the reviewed reference version used for evalu-

Table 51: Run and scoring provenance. Earlier three-way summaries are shown separately from the retained predictions used in the revised leaderboard. Values in the last two columns come from the same cached outputs.

System	Earlier T1 summary	Cached T1	Cached T1b
Qwen3-4B	0.4604	0.4047	0.4071
Qwen3-FOMC	0.6764	0.5735	0.6339
Eva-4B	0.783	0.7763	0.8133
DeBERTa-v3	0.664	—	—

Qwen summary-only evaluations are distinct inference runs; their per-pair outputs are unavailable, so the difference cannot be assigned to decoding or checkpoint changes. We make no run-equivalence claim. The earlier Eva-4B value 0.783 is not reproduced by its retained cache; excluding its invalid output yields T1=0.7775 and T1b=0.8151 on 199 pairs, compared with the all-200 error convention above. DeBERTa-v3 remains an archived three-way result.

ation. Versioned inputs, expected results, and change records are retained under `data/human_annotations/` and `analysis/`; their manifests specify the applicable populations and reference rules. These records remain available for auditing the relationship between the two inputs.

From the supplement root, reproduce agreement on the retained ratings with:

```
python3 -B code/reproduce_human_statistics.py \
  --stage pre_review --bootstrap \
  --output ../human-agreement-bootstrap.json
```

The `final` stage selects the reviewed reference version. Both stages require NumPy 1.22 or later and preserve their version-specific inputs. Each reconstructs the reported point estimates and, with `--bootstrap`, the 10,000 whole-meeting resamples per cohort (seed 20260926), interval endpoints, and undefined-replicate counts. Output paths must not already exist. The stages identify data versions, not independent replications or untouched first-pass responses. Private annotator notes and journals remain outside the anonymous package.

In both stages, the unchanged Dev fully-evasive class has five items, all unchanged between the original model and human labels. For each conditional unchanged-class fraction, 58/10,000 meeting-bootstrap replicates contain no eligible example and remain undefined; the interval uses the other 9,942. Its conditional $[100\%, 100\%]$ interval is not evidence of certain population agreement. The analysis retains fixed taxonomies and reports undefined fractions rather than substituting zero.

Run provenance and availability. The API and supervised-baseline prediction exports cover each of the six API Condition B systems and Eva-4B (OOD) on all 2,154 pairs, plus the two Qwen systems and Eva-4B+FOMC on the combined 400 test pairs. These minimal exports contain pair identifiers and predicted labels and support score verification. Appendix D.6 describes the separate rationale-audit package.

The Qwen3-4B and Qwen3-FOMC rows use their cached per-pair run, which differs from the earlier summary-only run (Table 51). Their cached three-class Macro-F1 scores are 40.5% and 57.3%, respectively, with binary scores of 40.7% and 63.4%. Earlier three-class summaries of 46.0% and 67.6% came from different inference artifacts and are excluded from the matched-cache evaluation.

The DeBERTa-v3 result is summary-only: its checkpoint and pair-level predictions are unavailable, preventing verified binary scoring or paired reanalysis. Eva-4B and Eva-4B+FOMC are distinct systems; the FOMC-adapted baselines used model-derived training-label variants. Their cached test predictions are evaluated against human majorities. The package does not supply the full-corpus Opus 4.6 Condition-D reference or per-system mechanism/tactic predictions, so diagnostics requiring those inputs are not reproduced by its scorer; descriptive auxiliary-field frequencies remain reproducible. Anonymous human triplets support aggregate agreement statistics. Reproducing scores from caches does not reproduce API inference or unavailable training artifacts.

Supervised-baseline recipes. Table 52 records the adapted baselines' recipes and model-derived training-label provenance, including gaps in checkpoint-to-label linkage.

Table 52: Training recipes for fine-tuned baselines. The historical training summaries use 1,554 model-labeled training pairs and model-consensus Dev for selection; these models were not trained on the newly completed human references. Label variants are distinguished in Table 47; summary-only and cached inference runs are distinguished in Table 51. These recipes do not establish that every retained checkpoint was trained on the exact frozen release labels.

System	Base	Epochs	LR	Batch	Max Len	Hardware
DeBERTa-v3	DeBERTa-v3-base	10	2e-5	16	512	H200 (Nebius)
Qwen3-FOMC	Qwen3-4B-Instruct	3	2e-5	16	2048	H200 (Nebius)
Eva-4B+FOMC	Eva-4B (84K earnings)	3	2e-5	16	2048	H200 (Nebius)
TF-IDF+LR	—	—	C=1.0	—	50K feat.	CPU