

Judge, Don't Write: Scaling Golden Data Generation via Psychometric Standard-Setting

Scott Novotney, Youbing Yin and Erin Babinsky

Capital One

{scott.novotney, youbing.yin, erin.babinsky}@capitalone.com

Abstract

Human-written references are typically treated as the upper-bound gold standard for fine-tuning and evaluation, but writing them is an inefficient use of subject-matter expert (SME) time. To best use limited SME attention, we apply psychometric standard-setting to discover acceptance criteria for GenAI applications. These acceptance boundaries are themselves useful for calibrating an LLM-as-a-judge, model tuning, and defensible quality monitoring. We report a case study on summarization across two domains at a financial institution. While summarization is a classic GenAI use case, what counts as acceptable differs significantly across domains. 25 SMEs across three panels produced 2,500+ judgments and yielded calibrated thresholds in an estimated 10x less SME time than the standard playbook of writing gold references (or “goldens”) directly. In both domains, LLM-generated goldens are preferred over SME-written ones roughly 80% of the time, by an independent human annotation cohort. Limited SME attention is best used to judge candidate goldens rather than write them.

1 Introduction

Summarization of multi-party dialogues and correspondence in regulated industries is a near-commodity GenAI use case: a frontier model produces a fluent summary, a subject matter expert (SME) reads it, the system ships. Yet applied use needs a defensible threshold separating acceptable from unacceptable on every rubric dimension, plus goldens with expert ratings. These are used to calibrate an LLM-as-a-judge; generate fine-tuning or heldout evaluation data; and produce rubrics for manual annotation and quality monitoring. However, SME time is the limiting factor in producing these high quality artifacts for model development and monitoring.

Human-written outputs have long served as reference gold standards across NLP tasks, from transcription (Godfrey et al., 1992) to summarization (Over et al., 2007; Fabbri et al., 2021), with evaluation metrics computed against them rather than against an explicit threshold. Even on established summarization rubrics, expert annotators reach only $\kappa = 0.41$ in a first pass and require a second calibration round to recover $\kappa = 0.71$ (Fabbri et al., 2021; Bayerl and Paul, 2011). The literature has long established that judging is easier than producing under expert constraints (Polanyi, 1966; Dreyfus and Dreyfus, 1986); the standard playbook nonetheless asks SMEs to write gold references (or “goldens”) directly. While fields like active learning and LLM-assisted annotation address this problem through rapid iteration, we apply *psychometrics* to best optimize limited SME time to efficiently discover what matters to our experts.

Psychometrics measures latent traits (knowledge, skill, quality) from observable responses. Since the 1970s, its standard-setting branch has elicited defensible pass/fail thresholds from expert panels through structured accept/reject decisions (Cizek and Bunch, 2007; Mitzel et al., 2001), producing the cut-scores that underpin professional licensing (USMLE for medical practice, the bar exam for law) and K-12 accountability testing. Recent work applies psychometrics to NLP from the opposite direction. Item Response Theory has been used to compare benchmark difficulty (Vania et al., 2021), weight evaluation examples by informativeness (Rodriguez et al., 2021), and shrink leaderboards down to a fraction of their original size with minimal information loss (Polo et al., 2024). Psychometric profiling treats LLMs as test-takers, administering personality and cognitive instruments to model outputs (Serapio-García et al., 2025).

These applications measure properties of test items or model traits. Standard-setting addresses a different question: given a fixed rubric, where does

the line between acceptable and unacceptable sit, and which dimensions actually move it? This paper asks whether standard-setting can answer that question for LLM-generated text in a regulated applied setting, and whether the same SME elicitation budget can yield boundary-line goldens, a calibrated rubric, and training data for an automated judge alongside the thresholds. The case study covers two summarization domains at a financial institution and a controlled simulation that calibrates each elicitation method against known thresholds at empirically observed panelist disagreement.

Our contributions are: (i) a pipeline that adapts psychometric standard-setting to LLM evaluation, yielding calibrated per-dimension acceptance thresholds, a decision rule, and a human-judgment calibration set as its primary artifacts, with boundary-line goldens as a by-product, at an estimated $10\times$ reduction in SME time; (ii) a controlled simulation at empirically observed inter-annotator disagreement ($\sigma = 0.24$) showing structured elicitation beats random sampling at every decision budget, with Bookmark matching random's 40-decision floor in a single placement; (iii) an empirical case study across 2 summarization domains, 25 SMEs, and 2,500+ judgments, showing that threshold-constrained outputs beat SME-written goldens 79% of the time, and statistically significant domain-specific differences in thresholds motivate per-domain panels.

2 Related Work

Eliciting expert judgment. Eliciting expert evaluation signal is expensive, and a substantial line of work has tried to reduce that cost. RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023), and active preference learning (Das et al., 2025) elicit pairwise judgments and convert them into reward functions; the resulting scalar is ordinal and carries no natural zero or threshold, so teams pick an operational cutoff post-hoc. LLM-as-judge systems (Liu et al., 2023; Zheng et al., 2023; Kim et al., 2024) score outputs through chain-of-thought, crowdsourced comparison, or distillation from GPT-4 examples, but each ranks quality without locating a pass/fail boundary. LLM-Rubric (Hashemi et al., 2024) is the closest peer-reviewed approach: it calibrates a multidimensional judge against per-rater Likert weights, calibrating the scoring function, while standard-setting calibrates the threshold the judge applies on top of it. The broader active-learning

literature (Settles, 2009) optimizes for classifier accuracy rather than threshold posterior calibration.

Psychometric standard-setting. Standard-setting is the psychometric branch that elicits criterion-referenced cut-scores from expert panels. The Bookmark method (Mitzel et al., 2001) asks panelists to place a divider in a difficulty-ordered sequence; Contrasting Groups compares score distributions of known masters versus non-masters (Cizek and Bunch, 2007; Yousefi Afrashteh, 2021). Both underpin high-stakes professional licensing (medical boards, bar exams). For Contrastive elicitation we adopt QUEST (Watson and Pelli, 1983), the standard psychophysical algorithm for Bayesian threshold estimation from binary responses. Psychometrics has also entered NLP from a measurement direction: Item Response Theory has been used to compare benchmark difficulty (Lalor et al., 2019; Vania et al., 2021), weight evaluation examples by informativeness (Rodriguez et al., 2021), and shrink leaderboards with minimal information loss (Polo et al., 2024), while Serapio-García et al. (2025) administer validated personality inventories to LLM outputs across 18 models. These applications measure properties of items or models; we measure where on a rubric the acceptable/unacceptable line sits.

Generating reference data with LLMs. A growing line of work uses LLMs to produce evaluation artifacts. Self-Instruct (Wang et al., 2023), WizardLM (Xu et al., 2024), and Dynabench (Kielbaso et al., 2021) generate test items; CheckList (Ribeiro et al., 2020) produces minimal pairs for behavioral testing; FActScore (Min et al., 2023) verifies atomic facts in long-form output. DSPy (Khattab et al., 2024) compiles prompts and demonstrations toward a metric ceiling, treating evaluation as an objective to maximize. These systems iterate or train toward the best possible output, optimizing for the ceiling. We instead iterate toward a *threshold* elicited from experts: a generator LLM produces a candidate, an evaluator scores it on each rubric dimension, and the loop revises until every dimension clears its panel-elicited cut-score. The output is calibration data at the pass/fail boundary rather than reference items at the ceiling. Chawla et al. (2026) survey the development lifecycle of an applied dialogue summarization system, the broader context our case study sits within.

Each thread above optimizes for a different artifact in isolation: a reward scalar, a judge cali-

bration, a benchmark, a generator. We elicit the threshold itself from experts, target it in a generator-critic loop, and recover thresholds, goldens, and judge-training data from a single panel session.

3 Method

We assume four inputs: a corpus of source documents (e.g., call transcripts, written correspondence), evaluation dimensions with Likert definitions (e.g., coherence 1–5, completeness 1–5), a panel of 5–10 domain experts who can judge but need not write, and a small set of hand-written goldens from experts or existing systems. The pipeline produces three artifacts which could be used to develop an automated judge: 1) a calibrated acceptance threshold per dimension, 2) a set of goldens at the empirical pass/fail boundary, and 3) a decision rule mapping multidimensional Likert scores to a binary verdict.

The pipeline proceeds in four stages (Figure 1): (1) construct a pool of single-dimension degradations from example outputs, (2) score the pool with an LLM to establish a quality ordering, (3) elicit acceptance thresholds from experts using either Bookmark or Contrastive protocols, and (4) generate goldens constrained to the elicited thresholds via an iterative generation-evaluation loop.

3.1 Pool Construction

The elicitation protocols require a pool of candidates that vary in quality along individual dimensions. For each dimension, an LLM edits a hand-written golden to produce quality variants at each target level (1–4; level 5 is the verified gold) consistent with the rubric definition. We over-generate (typically 4x per level), relying on stochastic diversity rather than rule-based mutation operators (e.g., delete-a-fact, swap-pronouns) that would require hand-specifying defects per dimension. Generating defects one dimension at a time keeps the expert’s later judgment on that dimension clean: for example, a hallucination defect should not drag down the grammar rating.

3.2 Pairwise Ranking

The over-generated pool needs ordering before experts can review it for the Bookmark method. We use an LLM to obtain pairwise preferences and fit a Bradley-Terry model (Bradley and Terry, 1952) to the win/loss outcomes, producing a continuous quality score per item. From the ranked pool we

select 9 evenly-spaced items per dimension as the Bookmark spectrum; for Contrastive elicitation, the QUEST algorithm (Watson and Pelli, 1983) draws adaptively from the full ranked pool, concentrating trials near the predicted boundary.

3.3 Threshold Elicitation

Given the scored pool, we elicit acceptance thresholds from the panel of 5–10 experts.

3.3.1 Bookmark

The Bookmark method (Mitzel et al., 2001) presents items ordered by predicted quality for one dimension at a time. The procedure is:

1. Each expert reviews the ordered sequence and places a “bookmark” at the transition from unacceptable to acceptable.
2. Panelists reveal placements simultaneously to prevent anchoring.
3. The panel discusses disagreements with a focus on items near the boundary.
4. Experts may revise placement after discussion (optional second round).

We fit a Beta distribution to panelist placements (it respects the bounded 1–5 scale and captures asymmetric distributions or panelist clustering at multiple values that a normal distribution smooths over) and set the operational threshold at one standard deviation below the mean, requiring outputs to satisfy the stricter panelists (false acceptances in practice are costlier than false rejections). A representative Beta fit is shown in Appendix Figure 4.

3.3.2 Contrastive

The Contrastive protocol replaces the ordered-sequence review with targeted binary decisions: “Is this acceptable on [dimension]?” The QUEST adaptive algorithm (Watson and Pelli, 1983) is the field-standard method for Bayesian threshold estimation from binary responses; it selects each next item to maximize expected information gain about the threshold and returns a calibrated posterior with a principled stopping rule (95% credible interval below 0.3 score units, typically converging in 12–32 items per dimension). Generic active-learning alternatives optimize for classifier accuracy rather than threshold posterior calibration and would require a separate rank-to-threshold conversion step. Contrastive needs no pre-built spectrum and is suited

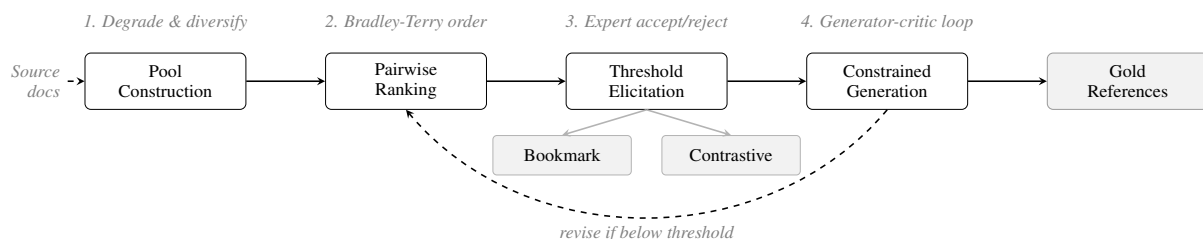


Figure 1: Overview of the four-stage pipeline. Experts make structured accept/reject decisions (Stage 3) rather than writing goldens; the generator-critic loop (Stage 4) produces outputs that clear every threshold.

to settings where ordering is difficult or expert sessions are short, at the cost of more individual judgments.

3.4 Constrained Generation

Once thresholds are set, a generator LLM produces a candidate summary, an evaluator LLM scores it on each dimension against the rubric, and the loop revises with dimension-specific feedback until every dimension clears its threshold or the iteration cap is reached. Both the generator and evaluator roles are filled by Claude Opus 4.6, accessed via Amazon Bedrock. The loop targets the acceptance *boundary* rather than the ceiling, preventing reward overoptimization, and the all-dimensions criterion does not allow dimensions to compensate for one another. Generated outputs that pass undergo a final lightweight expert review (accept/reject on the full output) as a quality gate rather than a calibration step.

4 Experimental Setup

4.1 Domains

We validate on two summarization domains at a financial institution.¹ Both domains are multi-party call summarization; they differ in business focus and in summary length and format (Banking outputs a longer structured summary). For both tasks, our SMEs were drawn from managers, experienced agents and quality assurance representatives.

The Auto domain panel comprised 8 SMEs and produced 10 hand-written goldens through a two-hour Bookmark elicitation session.

The Banking domain was a larger study with 50 hand-written goldens: 10 reserved as few-shot exemplars for the generation conditions and 40 used as the human baseline in the evaluation ablation.

¹We are unable to provide precise details about the source data used to ensure compliance with internal data governance policies. All source content was masked prior to use to remove personally identifiable information.

We collected threshold elicitation from two separate panels (8 SMEs in a Bookmark UI session, and 9 SMEs in a replication study)

4.2 Simulation Study

We validate elicitation efficiency in a controlled setting where ground-truth thresholds are known, allowing direct measurement of recovery error. The simulation uses spectra from our item pool (5 dimensions, items scored 1–5 by a 3-model ensemble) with known boundary locations derived from the pool structure.

Each trial samples a 5-person panel with per-panelist noise drawn from $\mathcal{N}(0, \sigma^2)$, representing disagreement about the true boundary. We apply uniform jitter (± 0.05) to pool items each trial to prevent methods from exploiting exact score coincidences. We sweep $\sigma \in \{0.1, 0.24, 0.5\}$, where $\sigma = 0.24$ matches empirically observed inter-annotator disagreement in our SME panels. We replicate each condition 100 times across all 5 dimensions (500 observations per cell). All three methods operate on the same item pool per trial:

Bookmark: Panelist reviews 9 items in score-sorted order (the spectrum from §3.2) and places a single boundary marker between items. This requires a pre-constructed spectrum but only 1 decision per panelist.

Contrastive (QUEST): Adaptive algorithm presents items sequentially; panelist makes a binary “above/below threshold” judgment on each. QUEST selects each next item to maximally reduce posterior uncertainty. We test budgets from 5 to 40 decisions (our maximum size of SME-generated goldens). This method requires an unordered pool with no spectrum construction.

Random: Items sampled uniformly from the pool; same binary judgment as Contrastive. Serves as the unstructured baseline.

All methods fit a Bayesian logistic to the accumulated binary responses and report the crossover

point (50% acceptance probability) as the estimated threshold. We measure mean absolute error (MAE) against the known boundary.

5 Results

5.1 Elicitation Efficiency

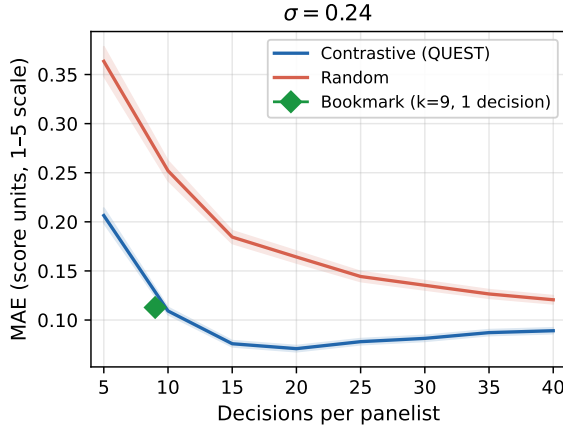


Figure 2: Threshold recovery MAE vs. decision budget at empirical noise ($\sigma = 0.24$). Bookmark (green diamond) achieves strong accuracy in 1 placement decision. Contrastive (QUEST, blue) converges faster than Random (red) and reaches lower asymptotic error.

Three patterns emerge across noise levels (Figure 2). At realistic noise ($\sigma = 0.24$), both structured methods average 0.11 MAE vs 0.20 for Random (45% lower); Bookmark in 1 decision (0.11) matches Random at 40 decisions (0.12), and Contrastive peaks at $k = 20$ (0.07) with diminishing returns beyond that. At low noise ($\sigma = 0.1$) structured methods improve over Random; at high noise ($\sigma = 0.5$) all methods converge near MAE 0.20, so protocol choice matters most in the moderate-agreement range we observe in practice.

The two structured methods differ in cost structure. Bookmark dominates at low budgets: 1 decision yields what Contrastive needs ~ 10 to match. But it requires a pre-sorted spectrum, meaning pool construction cost (§3.2) must be paid upfront. Contrastive requires no pre-ordering and achieves lower asymptotic error (0.071 vs. 0.113 at $\sigma = 0.24$), but at $20\times$ the per-panelist decision count.

5.2 Auto Domain Study

The Auto domain study validated the workflow before the larger Banking study: 7 evaluation dimensions, 8 SME panelists and 10 SME-written goldens. The SME-written goldens passed every threshold in only 60% of cases without the elicited rubric

and 100% with it. An independent SME (not on the panel) preferred the threshold-constrained output over SME-written goldens in 8 of 10 blind head-to-head pairs, citing completeness, grammar, and absence of superfluous detail. Finally, in a follow-up replication, two SME reviewers accepted 93% of 100 LLM-generated threshold-constrained goldens (71% with both reviewers agreeing), at a mean cost of 2.2 generation iterations. Writing these 100 goldens manually would take about 4,500 minutes (45 minutes each, self-reported by our SMEs). The elicitation session that produced them cost about 480 SME-minutes (8 SMEs for roughly an hour each), an estimated 10x reduction.

5.3 Banking Replication

Two Banking panels (17 SMEs total) elicited thresholds on the 8-dimension rubric for the Banking summarization task, with 40 SME-written goldens as the human baseline. We compared six generation methods on 1,020 pairwise A/B votes from a separate 6-SME cohort, and the Bank autotjudge (calibrated on elicited thresholds) scored every method’s output on each rubric dimension.

All four pipeline conditions beat both Reference System and Human SME by a margin wider than any bootstrap CI; Bookmark beats Human SME 79.4% of the time [67.8, 87.5] (Figure 3). The 5/5 ceiling is not always reachable; pushing one dimension to its ceiling destabilizes others (maximizing completeness makes coherence degrade), and Golden converges in 70% of cases while Bookmark clears the acceptance threshold 100% of the time. All-dimensions pass rates separate the methods sharply (Golden 92.5%, Rubric-only 75%, Bookmark 60%, Few-shot summary 60%, Reference System 0%; $n=40$ per method). Hallucination Rate and Referential Clarity bind: every method drops on these two while the remaining six dimensions saturate at $\geq 97.5\%$ across every pipeline-aware method.

The two Banking panels’ thresholds cluster. Across the Banking and Auto domains, seven shared dimensions split: on four (Repetition, Referential Clarity, Relevance, Hallucination Rate), across-domain SD exceeds within-domain SD by $2.3\text{--}4.2\times$ (Bonferroni $p < 0.001$); the remaining three (Coherence, Objectivity, Grammaticality) do not differ. Thresholds diverge across domains on exactly the dimensions that drive the pass/fail bar within each domain (Table 1).

For our second panel, we elicited thresholds

Dimension	Gap	Across/within	Bonf. p
Repetition	+1.74	4.17	$< 10^{-5}$
Referential Clarity	+1.22	3.38	$< 10^{-4}$
Relevance	+0.96	2.72	$< 10^{-3}$
Hallucination Rate	+1.17	2.31	$< 10^{-5}$
Coherence & Structure	-0.28	1.02	n.s.
Objectivity	-0.38	0.70	n.s.
Grammaticality	-0.01	0.01	n.s.

Table 1: Bank-vs-Auto threshold gap by dimension. Across/within SD ratios and Bonferroni-corrected p -values from independent two-sample comparison

on three shared dimensions under both protocols due to time constraints. Bookmark thresholds sat +0.60 score units above Contrastive (Wilcoxon $W = 1$, $p = 0.001$), consistent across all three dimensions. Panelists preferred Contrastive’s isolated binary judgments to Bookmark’s gradation across an ordered sequence (5-to-1 stated preference), suggesting the format itself, not just the resulting strictness, drives panelist preference.

5.4 Golden Generation Ablation

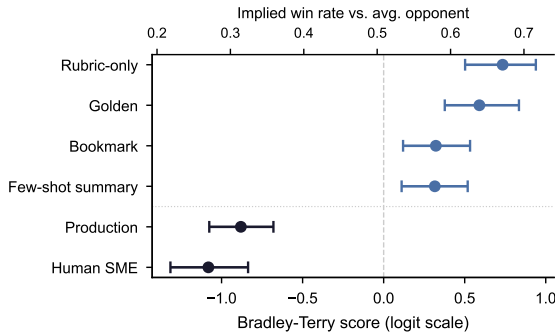


Figure 3: Bradley-Terry ranking of six conditions on 1,020 pairwise A/B votes (95% bootstrap CIs). SME-written and current reference summaries are easily beaten by any of the iterative generative process. By design, “Goldens” are more preferred than the “Bookmark” subset that stops at ‘acceptable’, vs ‘perfect’ thresholds.

To isolate the contribution of each pipeline component, we ran four generation conditions on the same backbone model across 40 source documents: *Bookmark* (rubric-aware loop, target = elicited threshold, 10-iteration cap); *Golden* (same loop, target = 5/5 on every dimension); *Rubric-only* (Golden’s prompt, single attempt, no critic loop); and *Few-shot summary* (10 summaries as style demonstrations, no rubric). The few-shot condition does not receive explicit acceptance criteria, isolating the rubric’s contribution from in-context

examples. Additionally, two baselines were included for comparison: *Human SME* summaries written from scratch, and *Reference System* (the existing institutional system).

A separate 6-SME annotation cohort cast 1,020 pairwise preference votes; we fit a Bradley-Terry model with 95% bootstrap CIs. The Bank auto-judge (calibrated on the elicited Bank thresholds) scored every pipeline-generated method on each rubric dimension to compute all-dimensions pass rates.

We do not claim that eliciting thresholds produces *better* goldens than the rubric alone: for generation, the rubric with iteration is competitive, and we report that as a finding. The Bookmark condition stops once it clears the elicited threshold rather than chasing 5/5, so measured against that threshold it passes less often than methods aiming higher; this reflects its target, not evidence that thresholds hurt generation.

6 Discussion

The expert-written reference set is the wrong artifact to optimize. The human baseline lost to every pipeline-aware condition in pairwise preference, and the rubric alone (without iteration) was already enough to clear that bar. What the SME budget actually buys is the threshold itself: where the line sits, which dimensions it binds, and how the line drifts across applications.

Only two of eight rubric dimensions (Hallucination Rate, Referential Clarity) drive the all-dimensions pass/fail bar; the other six saturate at $\geq 97.5\%$ across every pipeline-aware method. A team budgeting SME review uniformly across all eight is spending most of its budget on dimensions the model already passes. Standard-setting surfaces this asymmetry; the rubric alone does not.

The +0.60 score-unit gap between Bookmark and Contrastive on the same panelist on the same dimension complicates methodology choice. Both protocols are calibrated to elicit the same construct, yet they disagree by half a Likert unit. We read this as evidence that the construct itself is partly tacit: panelists hold it more strictly under gradation pressure (Bookmark’s ordered review) than under isolated binary judgment (Contrastive’s targeted accept/reject). Whichever protocol a panel uses, the resulting threshold is conditional on protocol choice, and downstream judges trained against it inherit that conditioning. Reporting the protocol

alongside the threshold is part of the threshold’s identity, not bookkeeping.

Thresholds replicate within a domain but split across domains on exactly the binding dimensions (Table 1); a team inheriting cutoffs from a sibling application calibrates the wrong boundary on the dimensions that matter most. Each new application therefore runs its own panel, and the elicitation’s audit trail keeps the per-application cost defensible.

7 Conclusion

Calibrated evaluation data requires knowing where the quality boundary is, not just what quality looks like. Standard-setting elicits that boundary from experts and sits between rubric design and judge calibration. Experts are good at judging quality; their time is better spent judging than writing.

Limitations

We validate on two summarization tasks at one organization. Generalization to other task types, languages, or organizations with different rubric structures remains untested. Panel sizes of 5–10 are adequate for Beta distribution fitting but may underestimate uncertainty at the tails. Cross-domain threshold transfer (Auto to Banking) is promising but rests on the assumption of comparable quality spectra; we do not yet have a defensible criterion for when transfer is safe. We do not report direct judge-human agreement metrics; the elicited thresholds are inputs to standard judge-training procedures (Kim et al., 2024; Hashemi et al., 2024; Chawla et al., 2026) and evaluating the trained judge against held-out human ratings is a separate study. The system described here is a research artifact; before any such pipeline is used in a real-world application, it requires further domain-specific validation, monitoring, and review against the requirements of the deploying organization.

Ethical Considerations

Our work was approved by an established internal review procedure. All source content was masked prior to use, removing personally identifiable information. We are not aware of any harmful applications of the methods described in this work. Given that LLMs are prone to hallucination and may reflect biases in their training data, we call for rigorous, domain-specific testing and validation before any such system is used for a real-world

application, alongside continuous monitoring of system outputs.

References

- Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. [What determines inter-coder agreement in manual annotations? A meta-analytic investigation](#). *Computational Linguistics*, 37(4):699–725.
- Ralph A. Bradley and Milton E. Terry. 1952. [Rank analysis of incomplete block designs: I. the method of paired comparisons](#). *Biometrika*, 39(3/4):324–345.
- Kushal Chawla, Chenyang Zhu, Pengshan Cai, Sangwoo Cho, Scott Novotney, Ayushman Singh, Jonah Lewis, Keasha Safewright, Alf Samuel, Erin Babin-sky, Shi-Xiong Zhang, and Sambit Sahu. 2026. [Lessons from the field: An adaptable lifecycle approach to applied dialogue summarization](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics: Industry Track*, pages 535–544. Association for Computational Linguistics.
- Gregory J. Cizek and Michael B. Bunch. 2007. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Sage Publications, Thousand Oaks, CA.
- Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2025. [Active preference optimization for sample efficient RLHF](#). In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*.
- Hubert L. Dreyfus and Stuart E. Dreyfus. 1986. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. The Free Press, New York.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- John J. Godfrey, Edward Holliman, and Jane McDaniel. 1992. [SWITCHBOARD: Telephone speech corpus for research and development](#). In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’92)*, pages 517–520, San Francisco, CA, USA. IEEE.
- Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. [LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. [DSPy: Compiling declarative language model calls into state-of-the-art pipelines](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- John P. Lalor, Hao Wu, and Hong Yu. 2019. [Learning latent parameters without human response patterns: Item response theory with artificial crowds](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Howard C. Mitzel, Daniel M. Lewis, Richard J. Patz, and Donald R. Green. 2001. The bookmark procedure: Psychological perspectives. In Gregory J. Cizek, editor, *Setting Performance Standards: Concepts, Methods, and Perspectives*, pages 249–281. Lawrence Erlbaum Associates, Mahwah, NJ.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35 (NeurIPS)*. Curran Associates, Inc.
- Paul Over, Hoa Dang, and Donna Harman. 2007. [DUC in context](#). *Information Processing & Management*, 43(6):1506–1520.
- Michael Polanyi. 1966. *The Tacit Dimension*. Doubleday and Co., Garden City, NY.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. [tinyBenchmarks: Evaluating LLMs with fewer examples](#). In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS)*. Curran Associates, Inc. NeurIPS 2023 oral.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should change NLP leaderboards?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. [A psychometric framework for evaluating and shaping personality traits in large language models](#). *Nature Machine Intelligence*, 7(12):1954–1968.
- Burr Settles. 2009. [Active learning literature survey](#). Technical Report TR1648, University of Wisconsin-Madison, Department of Computer Sciences.
- Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. [Comparing test sets with item response theory](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers). Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Andrew B. Watson and Denis G. Pelli. 1983. [QUEST: A Bayesian adaptive psychometric method](#). *Perception & Psychophysics*, 33(2):113–120.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. [WizardLM: Empowering large pre-trained language models to follow complex instructions](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.

Mahbobeh Yousefi Afrashteh. 2021. [Comparison of the validity of bookmark and Angoff standard setting methods in medical performance tests](#). *BMC Medical Education*, 21(1):57.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*. Curran Associates, Inc.

A Rubric Definitions

Table 2 lists the evaluation dimensions and their operational definitions. Six dimensions derive directly from DUC (Over et al., 2007) and SummEval (Fabbri et al., 2021); Objectivity was identified as binding during the Auto study, and Completeness was added for the Banking domain. Dimensions marked † appear only in Banking.

B Elicitation Protocol Details

Each Bookmark session proceeds in four stages. In the orientation stage (15–20 minutes), the facilitator explains the rubric, defines the scoring scale, and walks through two anchor examples at each extreme. In the individual placement stage (60–90 minutes), each panelist independently reviews the 9-item ordered spectrum for one dimension at a time and marks the transition from unacceptable to acceptable, working silently without discussion. In the reveal and discussion stage (30–45 minutes), placements are shown simultaneously; the facilitator leads structured discussion on items

where placements diverge by more than one position, without disclosing individual identities. In the optional revision stage (15 minutes), panelists may move their bookmark after discussion. Total session time is approximately 2–3 hours for 7–8 dimensions.

Contrastive sessions present items as isolated binary judgments: “Is this summary acceptable on [dimension]?” Each item appears with only the source document and the dimension definition visible; the score and rank are hidden from the panelist. QUEST selects each next item to maximize expected information gain about the threshold, concentrating trials near the predicted boundary after 5–8 warm-up items drawn uniformly. Typical convergence (95% credible interval < 0.3 score units) occurs in 12–32 items per dimension, taking approximately 2–4 minutes per dimension. The Banking live session covered three dimensions per panelist in a 45-minute block.

In the Banking live session, panelists completed Bookmark first, then Contrastive on the same three dimensions (Coherence, Grammaticality, Hallucination Rate), with a brief break between protocols. Order was fixed (Bookmark before Contrastive) to ensure the Bookmark spectrum was not influenced by Contrastive binary experience; this introduces a potential ordering confound acknowledged in the Limitations. After both protocols, panelists answered two questions: which protocol felt more natural, and whether the boundary felt consistent between protocols. Five of eight panelists reported Contrastive felt more natural; one preferred Bookmark; two reported no preference. Three panelists reported their mental boundary felt consistent between protocols; five noted the Bookmark gradation judgment felt more effortful and potentially stricter. The +0.60 Bookmark-vs-Contrastive offset reported in §5.3 held across all three live-session dimensions (Coherence +0.59, Grammaticality +0.82, Hallucination +0.39).

C Annotator Preference Reasons

Table 3 classifies the free-text rationales from 771 decisive A/B votes using an LLM classifier over ten quality dimensions. Themes here are not constrained to the eight rubric dimensions in Table 2; conciseness in particular reflects the implicit counter-pressure to Completeness rather than a separate rubric criterion. Completeness and conciseness account for 83% of all stated reasons, reflect-

Dimension	Definition
Hallucination Rate	The summary must contain only facts supported by the transcript. Information is supported if it is explicitly stated in the dialogue or is a direct, logical inference. Information from external sources outside the transcript is not supported.
Referential Clarity	All pronouns and references must be unambiguous. The reader must be able to tell who did what without re-reading. Use “the caller,” “the agent,” or specific role labels rather than ambiguous pronouns.
Coherence & Structure	The summary must flow logically and be well-organized. Information should follow a clear sequence (chronological or topical), with smooth transitions and no jarring jumps between topics.
Grammaticality	The summary must use correct grammar, proper sentence structure, and appropriate punctuation. No fragments, run-ons, subject-verb disagreements, or tense inconsistencies.
Objectivity	The summary must present facts neutrally without editorializing, expressing opinions, or using loaded language. No sentiment, judgment, or interpretation beyond what was explicitly stated.
Relevance	The summary must include only information relevant to the call’s purpose. Irrelevant details, tangential information, or filler content not pertinent to the primary issue should be excluded.
Repetition	The summary must not repeat the same information in multiple places. Each fact or detail should appear exactly once, in the most logical location.
Completeness [†]	The summary must capture all substantive information from the call: the reason for contact, verification steps, key details discussed, and outcomes or next steps.

Table 2: Operational definitions for the 8 evaluation dimensions used across the two domains. Six map directly to DUC/SummEval dimensions; Objectivity was identified as binding during the Auto study; Completeness ([†], Banking only) was added to address the longer structured output format of the Banking task.

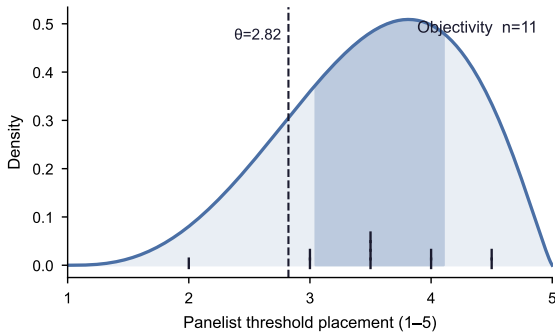


Figure 4: Beta-fitted threshold distribution for Objectivity from the Auto Bookmark session ($n=11$). Votes span 2.0 to 4.5, reflecting genuine disagreement about where acceptable objectivity begins. Ticks show individual placements (stacked where multiple panelists agree); shaded region is the central 50% credible interval. The Beta captures this left-skewed pattern that a Normal would smooth over.

ing the core trade-off the pipeline is designed to navigate. Conciseness is one side of this trade-off the panel is calibrating, not a confound: expert-written summaries win on conciseness (62% of their rationales) and the pipeline wins on completeness (76%), while stated accuracy is comparable on both sides ($\approx 7\%$). The pipeline therefore does not trade correctness for length.

D Threshold Votes: Bank vs. Auto

Figure 5 shows raw threshold votes on the seven shared rubric dimensions for the Bank panel and the Auto panel. Each dot is one panelist vote; the thick bar is the group mean. Distributions overlap substantially on Grammaticality and Coherence, while Hallucination Rate and Referential Clarity show the wider separation that drives the cross-domain threshold gap reported in Table 1.

E Computational Budget

Claude Opus 4.6 was accessed via Amazon Bedrock API.

Theme	Golden	Bookmark	Rubric-only	Few-shot	Human SME	Reference
Completeness	76%	60%	50%	41%	23%	46%
Conciseness	12%	22%	34%	42%	62%	19%
Accuracy	7%	7%	8%	10%	7%	17%
Coherence	4%	8%	3%	4%	4%	14%
Ref. Clarity	1%	3%	3%	1%	1%	4%
<i>n</i>	168	134	175	153	69	72

Table 3: Primary preference reason by winning method (% of notes, 771 decisive votes). Human SME wins primarily on conciseness; Golden wins primarily on completeness.

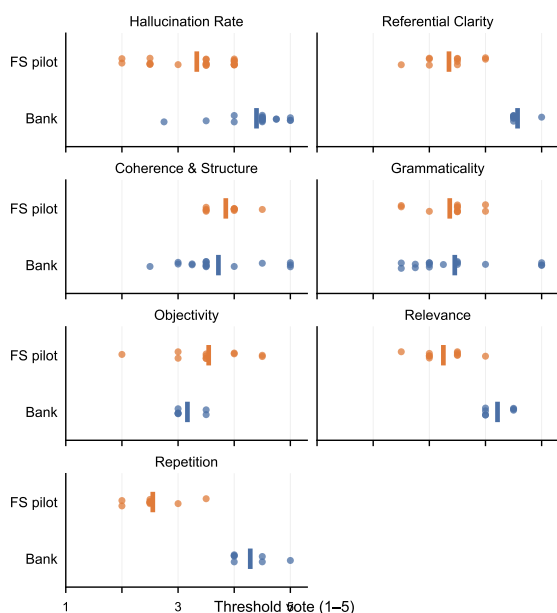


Figure 5: Raw threshold votes (1–5) on seven shared dimensions for Bank (blue) and Auto (orange). Thick bar = group mean. Four dimensions (Repetition, Referential Clarity, Relevance, Hallucination Rate) show clear separation; the remaining three overlap.