

When to Trust: Margin-Calibrated Routing Between LLM-as-a-Judge and Human Experts

Xinran Zhang, Ping Nie, Michelle Lok-ye Chan, Joshua L. Mason
Arash Sadati, Tate Sterling Avery

Thomson Reuters Labs

{Xinran.Zhang, Arash.Sadati}@thomsonreuters.com

Abstract

LLM-as-a-judge has become a common approach for comparing candidate answers at scale, but human experts who define what counts as “better” remain a scarce and finite resource. In practice, especially in professional domains, the key question is not whether to use the LLM judge or humans, but **how to combine them into a more efficient evaluation ecosystem**. We study five datasets spanning multiple domains and diverse LLM tasks, and find a consistent pattern: judge–human agreement increases as the margin between two answers grows, while model-produced factuality signals weaken as the margin narrows. This pattern is robust across datasets, different margin signals, and judge model families. We further find that the fact winner is not always the overall winner, and fact-decidable questions are not necessarily decided on factual grounds. Together, these results support a *fact-grounded but not fact-only* interpretation of comparison evaluation. Motivated by these findings, we propose and evaluate a practical *route-by-margin* evaluation strategy: choose a target agreement level T ; trust the LLM judge above the corresponding margin threshold; and defer the remaining low-margin cases to human experts, thereby using each resource where it is strongest.

1 Introduction

With the development of large language models, an increasing number of applications in different forms have been developed (Zhao et al., 2026): long-text Q&A, generated spreadsheets (Chen et al., 2025), and code (Chen et al., 2021). An important question in the development process, then, is how to compare answers in order to make iteration decisions. Given a question and two candidate answers, *which one is better?* Should we ultimately release changes? Currently, LLM-as-a-judge methods have been widely applied in such evaluation (Zheng et al., 2023; Li et al., 2025), especially for large-scale tasks (Chiang et al., 2024).

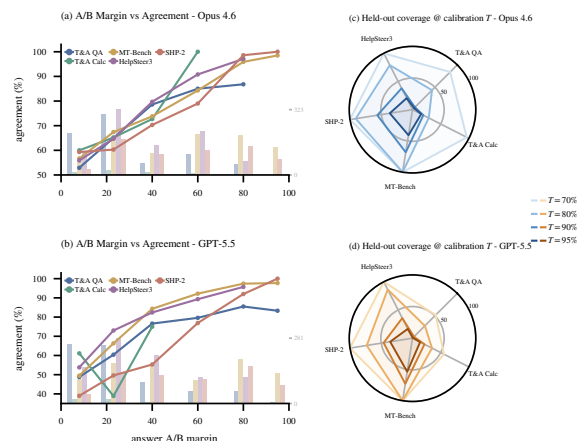


Figure 1: **The margin-agreement trend holds across datasets and judges, and supports held-out routing.** (a), (b) Judge-human agreement rises with the answer A/B margin on every dataset, under Opus 4.6 (a) and GPT-5.5 (b) (margin in six buckets, §3.3; buckets with fewer than five pairs omitted; color distinguishes the datasets; faint bars give the per-bucket sample size n , drawn in light gray with only two ticks, 0 and its max). (c), (d) Pooled held-out coverage at calibration target T , under Opus 4.6 (c) and GPT-5.5 (d). For each held-out fold, the other four folds select the highest-coverage margin threshold whose agreement is at least T ; the threshold is then applied unchanged to the held-out fold. Each polygon is one calibration target (lighter = lenient, darker = strict; shared legend, right).

Yet even with an LLM judge, the humans whose preferences define “better” (Ouyang et al., 2022) are expensive and finite (Zheng et al., 2023): in professional domains they are subject-matter experts (human experts) (Szymanski et al., 2025), and on open-ended tasks they are trained annotators or a voting crowd (Chiang et al., 2024). The practical problem is therefore not the LLM judge or humans, but **how to better leverage both the LLM judge and humans so that scarce human attention goes only where it is actually needed**, and a useful answer must say, per comparison, when the LLM judge can be trusted to decide alone and when a

human must be brought in (Shankar et al., 2024).

This paper studies how the *margin* (distance between two candidate answers, 0 means equally good, 100 means one is dramatically better than the other) affects the performance of the LLM judge, and how it can serve as a routing signal for judge-human collaboration. We conduct experiments on **five** natively pairwise datasets (Table 1) spanning diverse domains and LLM tasks: three public datasets and two expert-annotated datasets in the Tax & Accounting domain. We analyze the margin, its decomposition into fact-share, and its separate fact-margin; we also perform ablations over different margin signals (pairwise A/B margin, human margin, and rubric score margin), and replicate results using two separate models. The evaluation unit is the same throughout: a question paired with two candidate answers, with *agreement with human preference* as the evaluation metric.

We make three contributions. **First**, we show that judge-human agreement rises with the answer margin across five datasets, alternative margin signals, and two judge families. **Second**, model-based analyses suggest that low-margin comparisons are less often fact-separated; we characterize this pattern as *fact-grounded but not fact-only*. **Third**, we propose and validate *route-by-margin* with held-out threshold transfer and compare it with fixed uncertainty rules and a learned text-only router.

2 Related Work

Using a strong LLM to score or compare candidate outputs is now a standard evaluation tool (Liu et al., 2023; Li et al., 2025). LLM judges remain susceptible to the order in which the answers are presented (Wang et al., 2024), a preference for longer answers (Dubois et al., 2024), and a tendency to favor its own generations (Panickssery et al., 2024). Pairwise preference is one of the most common signals for both evaluating and aligning models (Ouyang et al., 2022), since a “which is better” judgment is easier to elicit reliably than an absolute score.

Expert domains such as tax, accounting, and law sharply raise the cost of the ground-truth signal (Aguda et al., 2024; Tseng et al., 2025), and the LLM judge’s agreement with the experts who define “better” is markedly lower there than on general instruction following (Szymanski et al., 2025). Aligning an automated evaluator with those expert preferences is itself an iterative, human-in-the-loop process (Shankar et al., 2024). The efficient alloca-

tion of expert attention then becomes the binding constraint on how fast a product can be evaluated.

Deferring uncertain decisions to humans is the central idea behind selective prediction and the reject option (Geifman and El-Yaniv, 2017), and behind learning-to-defer (Madras et al., 2018; Mozannar and Sontag, 2020). Routing between a model and a human is a promising way to leverage both.

3 Setup

3.1 Datasets and Pairwise Construction

We experiment on five native pairwise datasets spanning multiple domains and diverse LLM tasks (Table 1). Three are public: MT-Bench (Zheng et al., 2023), HelpSteer3 (Wang et al., 2025), and SHP-2 (Ethayarajh et al., 2022) (restricted to finance/legal/business domains). The other two are expert-annotated datasets we construct in the Tax & Accounting domain: T&A QA (long-text Q&A collected over six months) and T&A Calc (with spreadsheet artifacts as outputs). We cast every dataset into one common row schema (Q, v_1, v_2, p) : a question Q , two candidate answers v_1 and v_2 , and the ground-truth human preference $p \in \{v_1, v_2\}$.

All five sets are native pairwise, but each elicits and encodes the human preference differently. **MT-Bench** records a human majority vote over several annotators per pair; **HelpSteer3** gives three-annotator graded preferences, each a signed integer from -3 to $+3$ (excluding 0) whose sign marks the better answer and whose magnitude marks the strength of the preference; **SHP-2** provides a community upvote preference together with an upvote score ratio; and **T&A QA** and **T&A Calc** carry human expert pairwise preferences, the latter additionally with a four-dimensional rubric.

From each dataset we derive a binary ground-truth label and a native margin signal where available. When a set already supplies a direct pairwise label (**MT-Bench**, **SHP-2**, and both **T&A** sets), we take it as the ground-truth preference; for **HelpSteer3** we instead recompute the label from the raw annotator scores, taking the majority sign across a pair’s available scores and dropping the rare pairs whose scores split evenly. The native margin is then dataset-specific: for **HelpSteer3**, the mean of the annotators’ absolute scores (a *human margin*); for **SHP-2**, the dataset-provided ratio of the two comments’ upvote scores (also a *human margin*); and for **T&A Calc**, the *rubric score margin*, the absolute difference between the two answers’ overall

Dataset	Queries / pairs	Domain	Answer length	Preference source
T&A QA	293 / 720	tax/acct/audit Q&A	long, ~ 9.6 k ch.	human expert pairwise
T&A Calc	20 / 60	tax/acct calculation spreadsheets	long, ~ 11.8 k ch.	human expert pairwise (+ rubric)
MT-Bench	79 / 873	general chat/instructions (8 cat.)	short, ~ 0.9 k ch.	human majority vote
HelpSteer3	877 / 877	general open-ended Q&A	medium, ~ 1.8 k ch.	3-annotator graded
SHP-2	383 / 641	finance/legal/business Q&A	short, ~ 0.7 k ch.	community upvote ratio

Table 1: The five evaluation datasets used in the main analysis. Each row is converted into the common pairwise schema (Q, v_1, v_2, p) ; “Queries / pairs” reports unique prompts and answer comparisons, domain shows the main domain of the dataset, answer length is the mean character count per answer, and preference source shows how the original human preference is collected.

rubric scores ($s \in [0, 1]$; §3.2) rescaled to 0–100. These native margins are the signals used in our robustness analysis (§4.2). Dataset-construction details are in Appendix A.

3.2 Experimental Design

Across all five datasets, we apply the same pairwise evaluation pipeline: a simple LLM judge, a margin-and-decomposition step, an independent two-stage fact-margin extraction pipeline, and fact/non-fact labeling. On T&A Calc, we additionally run a pointwise rubric evaluation as an ablation.

Simple Pairwise LLM-as-a-judge: Given only a simple prompt, the baseline judge sees just (Q, v_1, v_2) and chooses the better one.

Margin & Decomposition: A separate call rates the overall answer A/B margin $m \in [0, 100]$ and decomposes it into fact and non-fact shares, also on a 0–100 scale and summing to 100. We use m as the main answer A/B margin (Figure 1).

Fact-Margin & Fact Winner: We run an independent two-stage, three-call fact-margin extraction pipeline. In the first stage, two separate calls take (Q, v_1) and (Q, v_2) and extract the atomic factual claims from each. The second stage consists of a final comparator call: it takes the question and the two fact-claim lists, aligns them into shared and unique claims, judges each distinct claim as correct, incorrect, or unverifiable, and returns seven counts: shared-correct s^+ , shared-incorrect s^- , per-answer unique-correct u_i^+ and unique-wrong u_i^- for $i \in \{1, 2\}$, and unverifiable. Write $U_i = u_i^+ + u_i^-$ for the claims unique to answer i , $n_i = u_i^+ - u_i^-$ for its net unique-fact score, and $D = s^+ + s^- + U_1 + U_2$ for the number of distinct claims. We then derive the *fact-margin*

FM, *fact-divergence* FD, and the *fact winner*:

$$\text{FM} = \frac{|n_1 - n_2|}{D} \times 100, \quad \text{FD} = \frac{U_1 + U_2}{D} \times 100,$$

$$\text{FW} = \begin{cases} v_1, & n_1 > n_2, \\ v_2, & n_2 > n_1, \\ \text{tie}, & n_1 = n_2. \end{cases}$$

Thus FM=0 means the two answers are equally correct on the facts, while FM=100 means one is far more correct; FW supplies the direction. If $D = 0$, we set FM = FD = 0 and FW = tie.

Fact/Non-fact Labels: We label factuality at both the query and pair levels. The query-level label sees Q alone and decides whether it is fact-decidable in principle. The pair-level label sees (Q, v_1, v_2) , the human preference (along with a comment if available), and decides whether that preference is *fact-driven* (one answer gets a verifiable fact right or wrong) or *non-fact/judgment-driven* (scope fit, completeness, clarity, tone, etc.).

Rubric Pointwise Evaluation: On T&A Calc only, we add a pointwise rubric evaluation as a pairwise-vs-pointwise ablation. Each question has a human-reviewed four-dimensional rubric: *Computational Accuracy*, *Input Source Correctness*, *Task Completeness*, and *Methodology Correctness*. For each answer, the LLM judge evaluates every rubric item as pass, partial, or fail, corresponding to scores of 1, 0.5, and 0. An item’s score is computed by multiplying this verdict value by the item’s maximum point value. The score for each dimension is then calculated as the total points earned in that dimension divided by the maximum possible points for that dimension. The answer’s overall rubric score $s \in [0, 1]$ is the equal-weight mean of the four dimension scores. The *rubric score margin* is $|s_1 - s_2| \times 100$, and the rubric winner is $\text{sign}(s_1 - s_2)$. Flagged errors are additionally validated by human experts. (Appendix F).

Signal	Source and use
Judge preference $\hat{p}$	Simple pairwise judge over (Q, v_1, v_2) ; compared with the recorded human preference to compute agreement.
Answer A/B margin m	Margin-and-decomposition call; main bucket and routing axis for agreement analyses.
Fact share	Same margin-and-decomposition call; share of m attributed to factual differences.
Fact-margin FM	Independent fact-claim pipeline; fact-level gap magnitude between v_1 and v_2 .
Fact winner FW	Same fact-claim pipeline; gives the more factually correct answer, or tie.
Human margin	Dataset-provided strength signal; used to test robustness to non-judge margin signals.
Rubric score margin	pointwise rubric score gap $ s_1 - s_2 \times 100$; used for the pointwise-vs-pairwise ablation.
Fact/non-fact labels	Query-level and pair-level labels; used to distinguish fact-decidable questions from fact-driven preferences.

Table 2: Analysis signals used throughout the experiments. The table maps each quantity to its source in the pipeline and to the analysis where it is used. Detailed definitions are in §3.2.

We independently experiment with two models from different families, *Claude Opus 4.6* and *GPT-5.5* (Table 6), to test cross-family robustness. We run with both models and report them side by side. Because T&A Calc is small ($n=60$ pairs), we repeat selected preference and structured-label components three times and aggregate. Detailed prompts and the full run inventory are in Appendix C.

3.3 Margin Buckets and Metric

The metric throughout is the LLM judge’s *agreement* with the human preference, on strict preferences only (ties and blanks excluded). We summarize the relationship in two ways.

As a **bin-free** summary we report the *Spearman rank correlation* ρ between a pair’s margin and whether the LLM judge agrees with the human; $\rho > 0$ means agreement rises with the margin, and being rank-based it requires no bucket edges.

For readability we also bucket the margin using edges at 0, 16, 31, 51, 71, 91, 101 for every dataset, defining left-closed intervals labeled ≤ 15 , 16–30, 31–50, 51–70, 71–90, > 90 ; buckets containing fewer than five pairs are omitted because their agreement estimates are unstable. Per-bucket 95% *Wilson intervals* and the *question-clustered bootstrap interval* on the overall agreement are reported in Appendix G.

For **HelpSteer3** and **SHP-2** we additionally study agreement against each set’s human margin (annotator-score strength and upvote ratio, respectively; §3.1), and for **T&A Calc** against the rubric score margin (§4.2).

3.4 Route-by-Margin and Evaluation

Route-by-margin retains high-margin comparisons for the LLM judge and routes the rest to a human expert. On labeled calibration set D , let $C_D(\tau)$ and

$A_D(\tau)$ denote coverage and judge–human agreement among cases with margin at least τ . For required agreement T , we select

$$\tau^*(T) \in \arg \max_{\tau} C_D(\tau), \text{ s.t. } A_D(\tau) \geq T.$$

Future cases with margin at least $\tau^*(T)$ remain with the LLM judge; the rest go to an expert.

We compare against three fixed rules and one learned router. *Self-consistency* retains unanimous same-model repeats; *cross-model agreement* retains cases where judge models agree; and *majority-low confidence* routes cases when most runs report low confidence. A per-dataset logistic regression predicts judge–human agreement from (Q, v_1, v_2) .

We report *coverage* (the fraction retained for the LLM judge), *selective agreement* (judge–human agreement on that set), and *selective risk* (one minus agreement). Fixed rules yield one coverage–agreement point. Continuous scores are compared by AURC (lower is better), and learned-router discrimination by AUROC.

To test threshold transfer for *Route-by-margin*, we use five deterministic, query-group-preserving folds. For each T , four folds calibrate $\tau^*(T)$, which is applied unchanged to the fifth. Thresholds range over distinct margins, with ties retained together; rotating the held-out fold yields pooled results. The learned router uses nested, group-preserving folds to prevent leakage; fixed rules require no fitting and are evaluated at predefined operating points. Details are in Appendix D.

4 Results

4.1 Agreement Rises With the Margin

On every dataset, the wider the gap between the two answers, the more often the LLM judge agrees with the human. A simple LLM judge that sees only (Q, v_1, v_2) agrees with the recorded

human preference on 61.5–82.5% of pairs overall (Table 3), but that single number hides structure: once pairs are bucketed by their margin, agreement climbs from near chance at near-ties to close to 100% at the widest gaps, nearly monotonically on every dataset and for both judges (Figure 1).

A bin-free rank correlation between margin and agreement, which avoids bucket choices, is positive throughout (Opus ρ from 0.16 to 0.39, GPT-5.5 from 0.25 to 0.45). The replicated association across the distinct architectures and training data of Opus 4.6 and GPT-5.5 demonstrates cross-family robustness.

Dataset	Opus 4.6		GPT-5.5	
	agree	ρ	agree	ρ
T&A QA	67.1	0.28	61.9	0.27
T&A Calc	71.2	0.16	61.5	0.25
MT-Bench	82.5	0.35	82.4	0.41
HelpSteer3	75.1	0.33	77.0	0.31
SHP-2	78.6	0.39	71.5	0.45

Table 3: Overall judge–human agreement (%) and the bin-free margin-agreement rank correlation ρ (§3.3) for all five datasets and both judges. ρ is positive on every dataset for both judges.

4.2 The Trend Is Robust Across Margin Signals

The margin trend generalizes across alternative margin signals. Using each dataset’s own human margin where available, HelpSteer3’s mean annotator preference strength and SHP-2’s community upvote ratio, agreement rises with the human margin just as it does with the A/B margin, under both judges (Figure 2a,b; $\rho \approx 0.25$).

T&A Calc supplies a sharper test through pointwise rubric scoring (Figure 2c). Each answer is graded on a human expert-authored four-dimensional rubric, and the answer with the higher overall rubric score is taken as the pointwise winner; this winner matches the human expert pairwise preference on 69.2% of pairs (67.3% under GPT-5.5). Bucketed by the rubric score margin, agreement again climbs monotonically from 50% for near ties to 100% at the widest gaps ($\rho = 0.44$ for Opus and 0.39 for GPT-5.5). On this small set, the A/B margin yields a weak trend ($n=52$, $\rho=0.16$ for Opus and 0.25 for GPT-5.5), whereas the pointwise rubric margin recovers a sharper, fully monotone rise on the same pairs.

The pattern survives a change of both the margin

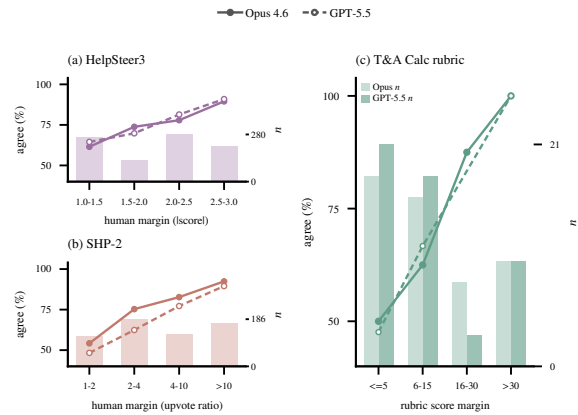


Figure 2: **The margin trend is robust to different margin signals** (both judges: Opus solid, GPT-5.5 dashed). Judge-human agreement rises with each dataset’s own *native margin signals*: Human margin (a) HelpSteer3’s mean annotator |score| and (b) SHP-2’s upvote ratio; and (c) T&A Calc’s pointwise rubric score margin. In every panel the two lines are the judges’ agreement and the light bars give each bucket’s size n ; in (a), (b) n is identical for both judges (the bins are on the human margin), in (c) the two shades are the per-judge n .

signal and the evaluation paradigm. Human experts confirm the rubric judge’s flagged errors on 85.2% of 1,216 validated items (Appendix F).

4.3 A Smaller Margin Indicates a Less Factual Comparison

Smaller margins have a lower fact-share and carry a smaller fact-margin. As A/B answer margin narrows, factual share of that margin declines, as does the fact-margin computed by an independent fact-margin extraction pipeline, consistently across all datasets and for both judges (Figure 3). These two independent signals suggest that, as two answers become more similar, factual differences explain less of their gap and the answers increasingly resemble each other at the factual level.

4.4 Margin Thresholds Transfer to Held-Out Query Groups

Margin thresholds selected on calibration data transfer to held-out groups. Across target agreement levels, thresholds calibrated as described in §3.4 yield pooled held-out agreement close to the target and coverage close to calibration coverage on the four larger datasets (Figure 4). Transfer is less stable on T&A Calc because its 52 strict comparisons leave only 10–11 examples per held-out fold. These results support threshold transfer to unseen query groups.

Dataset	No route Agree.	Self-consistency C / Agree.	Cross-model C / Agree.	Majority-low C / Agree.	LR AUROC	AURC LR / Margin
T&A QA	67.1	98.5 / 67.4	80.6 / 68.0	68.8 / 72.6	.587	.272 / .213
T&A Calc	71.2	100.0 / 71.2	71.2 / 73.0	69.2 / 69.4	.440	.297 / .197
MT-Bench	82.5	97.0 / 83.1	87.7 / 86.9	86.0 / 86.3	.518	.166 / .070
HelpSteer3	75.1	97.1 / 76.4	79.5 / 82.8	78.4 / 81.1	.532	.230 / .125
SHP-2	78.6	98.8 / 78.8	87.2 / 78.7	92.7 / 80.3	.713	.108 / .081

Table 4: Routing baselines for Opus 4.6 judgments. *C / Agree.* is retained-set coverage / judge–human agreement in percent. The three hard rules have fixed operating points. LR is evaluated with nested, query-group-preserving folds; higher AUROC is better. AURC summarizes selective risk across coverage; lower is better.

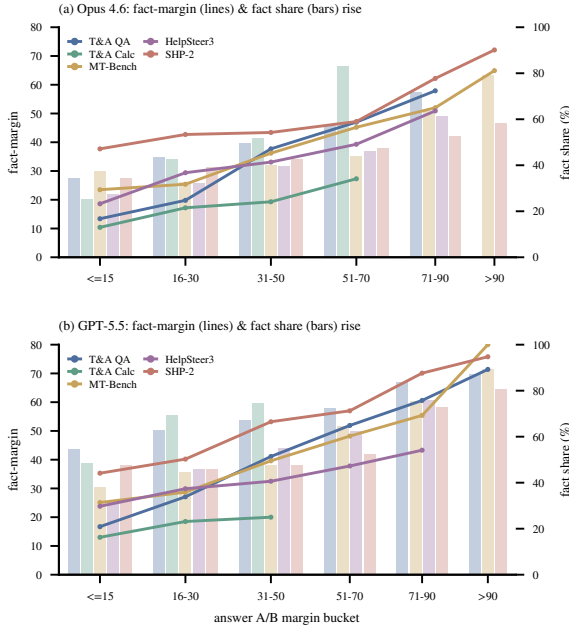


Figure 3: **Model-produced factuality signals rise with the A/B margin under both judges.** As the answer A/B margin grows, the per-dataset fact-margin (lines, left axis) and the fact-share of the margin (bars, right axis) both rise, under Opus 4.6 (a) and GPT-5.5 (b). Color denotes the dataset.

4.5 Alternative Routing Signals

Alternative signals yield dataset-specific gains, while margin remains consistently useful. Self-consistency routes only 0–3% of comparisons and changes agreement little. Cross-model agreement helps most on MT-Bench and HelpSteer3, while majority-low confidence improves retained-set agreement on four datasets, with T&A Calc as the exception. The learned router discriminates clearly on SHP-2; margin has lower AURC on all five datasets (Table 4).

Operationally, margin adds one estimation call and provides a continuum of risk–coverage operating points. Cross-model routing adds a second judge-family call; self-consistency and majority-low confidence use three Opus preference calls.

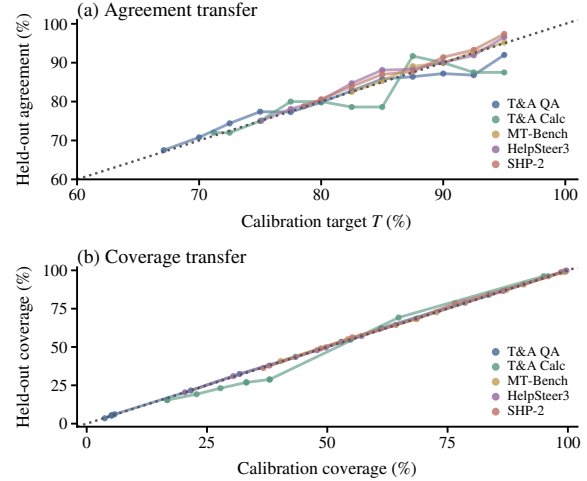


Figure 4: **Five-fold held-out transfer for route-by-margin using Opus 4.6.** Each point pools five held-out folds for one dataset–target pair. (a) Held-out agreement versus T ; (b) pooled held-out versus mean calibration-fold coverage. Dotted diagonals mark exact transfer. For each target, four folds select a threshold, which is applied unchanged to the fifth; held-out labels are used only for evaluation.

Overall, route-by-margin is an empirically validated operational strategy within selective prediction and learning to defer.

5 Discussion

5.1 A Model-Based Fact-Grounded but Not Fact-Only Interpretation

The fact winner and overall winner frequently diverge. An independent fact winner (FW), the answer with more net-correct checkable claims, predicts the human preference on at most 80.1% of pairs and on roughly two-thirds for most datasets (Table 5), below the full simple LLM-as-a-judge baseline (Table 3). The full judge’s higher agreement shows that qualities beyond factual differences contribute to preference.

Fact-decidable questions are not necessarily fact-decided. On expert-annotated T&A sets most

questions are fact-decidable (72.4% of T&A QA and 60.0% of T&A Calc; Table 5), yet far fewer pairwise differences are labeled fact-driven (18.1% and 28.3%): both answers clear the factual bar, so the comparison turns on non-fact considerations.

Comparison evaluation is entering a *fact-grounded but not fact-only* stage. Factual correctness remains a floor, but completeness, scope, clarity, and usefulness can decide between answers that are difficult to separate factually.

Dataset	Opus 4.6			GPT-5.5		
	%qf	%pf	fw	%qf	%pf	fw
T&A QA	72.4	18.1	63.9	78.5	25.4	64.7
T&A Calc	60.0	28.3	66.7	70.0	38.3	64.0
MT-Bench	49.4	52.0	76.7	50.6	60.6	80.1
HelpSteer3	15.6	27.4	63.2	14.1	44.5	62.0
SHP-2	2.6	41.7	66.1	4.4	50.2	65.1

Table 5: **The comparison is fact-grounded but not fact-only.** %qf: share of queries fact-decidable in principle; %pf: share of pairwise *preferences* that are fact-driven; fw: how often an independent fact winner predicts the human preference. Many questions are fact-decidable yet far fewer differences are fact-driven, and across every set the fact winner predicts the preference on only 62–80% of pairs, well below the full judge.

5.2 Applying Route-by-Margin in Practice

In deployment, historical labeled comparisons form the calibration set. For minimum agreement T , route-by-margin selects and freezes the highest-coverage feasible threshold for future comparisons.

Coverage and the target agreement involve a trade-off. A higher T raises the threshold, lowers LLM coverage, and sends more comparisons to an SME; a lower T retains more, including harder, lower-margin cases. There is an inherent tension between a higher T and higher coverage. Figure 1(c, d) shows held-out coverage across targets.

The optimal operating point depends on the application. Each distinct threshold defines one point on the risk–coverage frontier (Figure 5). The appropriate T depends on automated-error cost and available SME capacity. High-stakes applications may favor greater agreement and expert review; high-volume, lower-risk settings may favor greater coverage. A deployment should report its selected T , threshold, calibration agreement, and coverage, and recalibrate when the judge, margin prompt, domain, or answer distribution changes.

This perspective also clarifies where the LLM judge should improve. Higher overall agreement

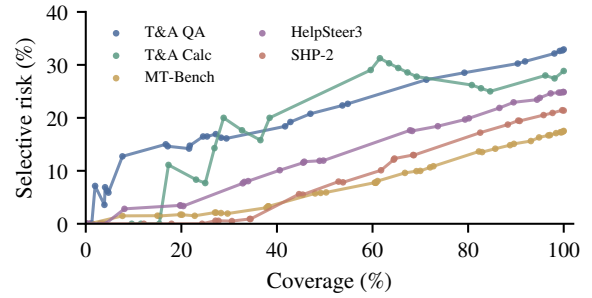


Figure 5: **Applying route-by-margin with historical labeled data.** Using Opus 4.6 results, each point represents a distinct deployable margin threshold; equal-margin comparisons enter together. Moving to higher coverage retains more LLM-judged comparisons but may increase selective risk. The curve shows operating points for application-specific calibration.

need not reduce SME demand if gains concentrate on easy, high-margin comparisons. Progress means pushing the coverage–agreement frontier outward: retaining more comparisons at fixed agreement, or higher agreement at fixed coverage. This matters as candidate answers converge: margins decrease across successive collection batches (Appendix H), shifting more comparisons toward the low-margin region. **The central challenge is therefore to judge near-ties reliably.** Improving accuracy on small-margin pairs keeps more with the LLM judge, reduces SME load, and sustains route-by-margin as candidate answers become harder to distinguish.

6 Conclusion

This paper asks: **how can LLM judges and human experts be combined in a more efficient evaluation ecosystem?** Across five pairwise datasets spanning domains and tasks, judge–human agreement consistently increases with answer margin. This relationship is robust across datasets, margin signals, and two judge families.

Model-based factuality analyses further suggest that close comparisons are less often fact-separated: fact winners do not always match overall winners, and fact-decidable questions are not always decided factually. This supports a *fact-grounded but not fact-only* view of comparison evaluation. We therefore propose route-by-margin: calibrate a margin threshold τ for target agreement T , retain higher-margin comparisons for the LLM judge, and defer the low-margin tail to human experts. Held-out transfer and routing-baseline comparisons support this strategy as a practical way to focus scarce human attention.

Limitations

Small T&A Calc set. The T&A Calc dataset is small, with only 60 pairs from 20 questions (52 strict; preference and selected structured labels use three runs to limit noise). The per-bucket agreements, the rubric-margin trend, and any routing threshold calibrated on this set therefore rest on few pairs and carry wide intervals, and the weak A/B-margin trend on this dataset ($\rho=0.16$) is partly a small-sample effect; a deployment should calibrate its threshold on a larger labeled sample.

Model-produced signals. The A/B margin, its fact/non-fact decomposition, the fact-margin, and the fact/non-fact labels are produced by the LLM judges themselves (Claude Opus 4.6 and GPT-5.5) and may inherit their biases and run-to-run non-determinism. Independent SME factuality annotations are unavailable, leaving these signals without direct factuality validation. Robustness checks substitute each dataset’s native human margin (**HelpSteer3** and **SHP-2**) and the expert-authored rubric score margin (**T&A Calc**); all recover the same trend (§4.2). These judge-independent signals reduce this dependence on the datasets where they are available. A fully independent margin signal is an important direction for future work.

Single longitudinal set and retrospective routing. The margin-erosion-over-time trend is observed only on the one set with collection timestamps (T&A QA); the external sets are single snapshots.

Scope of repeated-run evaluation. The panel-based routing baselines use three Opus preference runs. GPT-5.5 supplies one preference run on four main datasets and three on T&A Calc, while most margin and factuality components are evaluated once per judge. Thus, our repeated-run analysis focuses on Opus preference routing; extending it across judge models and intermediate labels is future work.

Baseline scope. The three fixed routing baselines use predefined rules without dataset-specific tuning. The learned router is a lightweight logistic regression over frozen all-MiniLM-L6-v2 embeddings and simple length/similarity features; only its regularization is selected on the inner folds. Future work may compare stronger, task-adapted learned routers and more extensively tuned routing alternatives.

Acknowledgments

We thank our expert team for their annotations, domain guidance, and feedback throughout the process.

References

- Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. [Large language models as financial data annotators: A study on effectiveness and efficiency](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). Preprint, arXiv:2107.03374.
- Yibin Chen, Yifu Yuan, Zeyu Zhang, Yan Zheng, Jinyi Liu, Fei Ni, Jianye Hao, Hangyu Mao, and Fuzheng Zhang. 2025. [SheetAgent: Towards a generalist agent for spreadsheet reasoning and manipulation via large language models](#). In *Proceedings of the ACM on Web Conference 2025, WWW '25*, pages 158–177, New York, NY, USA. Association for Computing Machinery.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot Arena: An open platform for evaluating LLMs by human preference](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. [Length-controlled AlpacaEval: A simple de-biasing of automatic evaluators](#). In *First Conference on Language Modeling*.
- Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. [Understanding dataset difficulty with \$\mathcal{V}\$ -usable information](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.
- Yonatan Geifman and Ran El-Yaniv. 2017. [Selective classification for deep neural networks](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang 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.
- David Madras, Toni Pitassi, and Richard Zemel. 2018. [Predict responsibly: Improving fairness and accuracy by learning to defer](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Hussein Mozannar and David Sontag. 2020. [Consistent estimators for learning to defer to an expert](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.
- Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. [Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences](#). In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST ’24, New York, NY, USA. Association for Computing Machinery.
- Annalisa Szymanski, Noah Ziemis, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. [Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks](#). In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI ’25, pages 952–966, New York, NY, USA. Association for Computing Machinery.
- Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. [Evaluating large language models as expert annotators](#). In *Second Conference on Language Modeling*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025. [HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, and 4 others. 2026. [A survey of large language models](#). *Frontiers of Computer Science*, 20(12):2012627.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Appendix overview

This appendix gives the full construction, configuration, prompts, scoring rules, and per-dataset detail behind the main results; each section is self-contained.

- A Dataset construction
- B Judge configuration
- C Experiment pipeline & prompts
- D Routing evaluation details
- E Rubric scoring
- F Error validation
- G Confidence & reliability
- H Margins erode over time
- I Results on additional datasets

A Dataset Construction

Every dataset is reduced to the common schema ($Q, v_1, v_2, \text{preference}$) and we keep strict preferences only (ties and blanks dropped).

A.1 Expert-Annotated T&A Datasets

T&A QA. Answer pairs are collected from a Q&A task in the Tax & Accounting domain. Annotations come from eight evaluation batches over six months (from Nov 2025 to May 2026); a batch date derived from each label (a trailing MM-DD) drives the longitudinal trend (§H).

T&A Calc. Twenty calculation questions, each answered four times, give 80 answers, 60 pairs, 52 with a strict human expert preference. Each answer is its written response *plus* its output spreadsheet serialized to text, so the LLM judge grades the same two-part artifact. Human expert preferences, free-text comments, and per-rubric-item error validation were all collected for this set, which is what lets it carry the rubric and error-validation analyses (§E–F).

A.2 External Public Sets

The three external sets in the main five are **MT-Bench** (Zheng et al., 2023), model-vs-model answers to an MT-Bench prompt with a human majority vote (we keep pairs with a strict majority); **HelpSteer3**, each pair rated by three annotators on a $-3..+3$ scale, the preference being the majority sign; and **SHP-2**, two comments under one Reddit post, the preference being the more-upvoted comment.

Four further public sets (LLMBar, Reward-Bench, CaseHOLD, and CFA), run under Opus 4.6 only, are introduced and analyzed together in Appendix I.1.

A.3 Human Margins

Three datasets carry a *human margin*, a human-side magnitude of preference we substitute for the LLM judge’s A/B margin in the robustness check (§4.2). **HelpSteer3**: the mean of the three annotators’ $|\text{score}|$ (distance from neutral), so a pair that all three annotators rate ± 3 has a larger margin than one rated ± 1 . **SHP-2**: the upvote ratio (higher / lower comment), a crowd-scale confidence in the preferred side. **T&A Calc rubric score margin**: the absolute difference between the two answers’ pointwise rubric scores (§E), a model-graded but *pointwise* quantity that evaluates the two answers separately and is therefore independent of the A/B call. Each is bucketed on interpretable fixed edges: HelpSteer3 in Figure 2a, SHP-2 in 2b, and the rubric score margin in 2c.

B Judge Configuration

We run two judges through one identical pipeline: **Claude Opus 4.6** and **GPT-5.5**. Their distinct architectures and training data make the reproduced margin trend evidence of a comparison-level effect across model families. Both process inputs without truncation and use a 128k-token output cap; structured tasks return strict JSON and the rubric judge uses a forced tool call. Decoding is each vendor’s default (the reasoning models reject an explicit temperature / top- p , so none is set). Content is cached by prompt, so every result is reproducible, and the two judges write to separate result trees so each number stays traceable to its LLM judge. Replication varies by component as detailed below.

Judge	Cutoff	Config (default)
Claude Opus 4.6	Aug 2025	no truncation; 128k output; strict JSON / forced tool
GPT-5.5	Dec 2025	identical

Table 6: The two judges. Knowledge-cutoff entries are the vendor-published training cutoffs (month granularity). Both run the identical prompts, input handling, and decoding; only the underlying model differs.

Run inventory. The five main datasets have three identical-prompt Opus preference runs for the self-consistency and majority-low baselines. GPT-5.5 has one preference run per dataset except T&A Calc, which has three. Margin decomposition and the query-/pair-factuality labels are single-run per judge except on T&A Calc, where they are repeated three times and aggregated. The fact-margin and rubric pipelines are single-run per judge.

Persona configuration. Every main experiment uses a fixed persona matched to the dataset. The same role is used for Opus and GPT-5.5 within each dataset; preference and structured-label prompts differ only in their task instructions.

Dataset	Persona used in the main experiments
T&A QA, T&A Calc	Senior tax and accounting expert answering and assessing complex professional questions.
MT-Bench	Senior expert answering and assessing complex questions across many domains.
HelpSteer3	Senior expert answering and assessing complex questions across many domains.
SHP-2	Senior expert in finance, law, and business answering and assessing questions in those domains.

Table 7: Dataset-matched personas used for the five main datasets. Persona text remains fixed within each dataset; routing uses the margin signal.

Persona robustness. We compare a detailed dataset-matched persona with a shorter, domain-appropriate persona on MT-Bench and two additional public benchmarks: LLMBar, which tests instruction following under adversarial comparisons, and RewardBench, which spans chat, safety, and reasoning preferences. Table 8 reports judge–human agreement and the item-level Spearman correlation between margin and agreement for one Opus run per condition.

Dataset	Persona	Agree. / ρ
MT-Bench	Detailed: expert across writing, roleplay, reasoning, math, coding, extraction, STEM, and humanities	82.0 / .346
	Shorter: senior expert across many domains	82.5 / .345
LLMBar	Detailed: instruction-following and response-quality expert	82.3 / .262
	Shorter: senior expert answering professional and complex questions	77.1 / .340
RewardBench	Detailed: helpful, harmless, and accurate response expert across chat, safety, and reasoning	91.9 / .337
	Shorter: helpful, harmless, and accurate response-assessment expert	91.4 / .346

Table 8: Persona robustness. Entries are judge–human agreement (%) and Spearman ρ between margin and the agreement indicator. Each persona condition is one Opus run.

The two personas yield similar agreement and margin–agreement association on MT-Bench and

RewardBench, while LLMBar agreement differs by 5.2 points. The association remains positive in all six conditions ($\rho = .262\text{--}.346$). These results show that the qualitative margin relationship persists across the tested persona variants, although persona choice can affect absolute agreement. The comparison is single-run and does not include HelpSteer3 or SHP-2.

C Experiment Pipeline and Prompts

The pipeline has six LLM components per dataset, each a separate cached call: the **baseline judge** (the A/B preference and binary confidence), the **margin decomposition** (overall margin split into fact / non-fact shares), the **query-fact** and **pair-fact** labels, and the two-stage **fact-margin pipeline** (per-answer claim extraction, then a comparator). T&A Calc additionally carries the rubric judge (§E). We quote each prompt’s operative text below.

Persona and controls. Two controls keep the cross-dataset comparison fair. First, *position*: v_1/v_2 order is randomized on the public sets, yielding balanced position preferences (§3). Second, *persona*: the expert-annotated T&A sets (T&A QA, T&A Calc) keep the original tax-and-accounting persona; the public sets swap in a domain-matched persona (e.g. “a senior legal expert” for CaseHOLD and “a senior finance expert” for CFA), extending the cross-dataset analysis beyond the tax framing. The fact/non-fact label bodies are identical across datasets, with only the persona line and the fact *examples* generalized (“an IRC/ASC statute” becomes “an established rule or citation”); the fact/non-fact concept itself is held fixed.

Baseline judge. The headline LLM judge uses (Q, v_1, v_2) as its complete input:

You are a senior tax & accounting expert who is good at answering complex Tax and Accounting questions.

You will be shown a user question and two candidate answers, V1 and V2. You need to apply your expertise and knowledge to decide which answer is better as an answer to the question.

Output STRICT JSON: a single object {"preference": "V1"|"V2", "confidence": "high"|"low", "reason": "<short>"}. No prose.

The listing shows the T&A role; public datasets substitute the corresponding persona in Table 7 without changing the decision task.

Margin decomposition. One call rates the overall gap and splits it into fact / non-fact shares (the source of the A/B margin and fact-share in Figure 3):

You will be shown a Question and two candidate answers (V1, V2). Rate the OVERALL gap between them as answers to this question -- weighing everything that matters together.

winner: "V1" | "V2"

margin: integer 0-100 (0 = the two are equally good overall, 100 = one is dramatically better)

Then decompose the margin into two shares that sum to 100:

fact_share : % of the gap driven by FACT factors -- factual/numerical/computational correctness, the right rule, citation, threshold, date or amount, a correct yes/no determination, non-fabrication, being responsive vs off-topic. These have a determinate, checkable right answer.

nonfact_share : % driven by NON-FACT factors -- completeness /scope trade-offs, clarity, organization, concision, tone, formatting, citation style, practical usefulness -- judgment calls where competent experts could differ.

fact_share + nonfact_share MUST equal 100.

Output STRICT JSON: {"winner","margin","fact_share","nonfact_share","basis"}. No prose.

Example	Public MT-Bench comparison	Observed
Low margin	Both answers correctly compute 8,000 + 4,000 = 12,000; V1 is more step-by-step and V2 more concise.	$m=2$; human V1; judge V2
High margin	A grammar task contains many errors; V1 says none exist, while V2 supplies a corrected paragraph.	$m=99$; human V2; judge V2

Table 9: Low- and high-margin examples. A small presentational difference yields a near-tie and judge-human disagreement; a decisive defect yields a large margin and agreement.

Query-fact and pair-fact. Two independent labels ask whether the *question* is fact-decidable in principle, and whether the *recorded preference* actually turned on a verifiable fact:

[query-fact] Decide whether the question is fact-decidable IN PRINCIPLE:

"fact" -- determinate, checkable answer: a specific number/threshold/date, a statute or rule, a yes/no, or a single correct procedure. Two competent experts who looked it up would agree.

"non-fact" -- open-ended/advisory: analysis, strategy, trade-offs, planning, explanation, with no single correct answer.

[pair-fact] Given Q, V1, V2, the human expert's preferred answer and (often) the human expert comment, decide WHY the human expert preferred one answer:

"fact" -- the preference is driven by which answer gets a verifiable fact RIGHT (the other is wrong, fabricates, cites the wrong rule, uses a stale threshold, or is non-responsive)

"non-fact" -- both answers clear the core facts; the preference turns on judgment: scope, completeness, concision, clarity, usefulness, tone. The human expert comment is the strongest signal and is weighted heavily when present.

Fact-margin pipeline (independent of the decomposition). Stage 1 extracts each answer's checkable claims (per-answer, cached); stage 2 aligns and grades them:

[extract] Extract the answer's atomic, CHECKABLE factual claims -- each a statement verifiable true or false (a number/value/date, an established rule/standard/citation, a yes/no, a definition, or a computed result). No opinions, hedging, or restatements.

Output: {"claims": ["<claim>", ...]}.

[compare] Given the question and the two claim lists (A1, A2):

1. Align: a claim is SHARED if both lists assert the same fact, else UNIQUE to its side.
2. Judge each distinct claim: correct / incorrect / unverifiable.
3. Count into: shared_correct, shared_incorrect, A1_correct_unique, A1_wrong_unique, A2_correct_unique, A2_wrong_unique, unverifiable.

From these counts we derive two magnitudes: the *fact-margin*, the absolute net-correct fact gap (with the fact winner supplying direction), and *fact-divergence*, the share of distinct claims the two answers disagree on. Both rise with the A/B margin (e.g. fact_divergence 44% → 100% on T&A QA, 43% → 100% on HelpSteer3).

D Routing Evaluation Details

Held-out reporting. We pool retained and correct counts across the five held-out folds and compute coverage and agreement from these pooled counts. Figure 4 evaluates Opus from each dataset's no-routing agreement at 2.5-point targets through 95%. Figure 1(c,d) shows the 70, 80, 90, and 95% subset and applies the same protocol separately to both judge families.

Fixed rules and vote entropy. Self-consistency retains an item only when all three Opus preferences are identical. For three binary votes, vote entropy is zero exactly in this unanimous case, so routing nonzero-entropy items is the same rule. Cross-model agreement retains an item only when Opus and GPT-5.5 agree; with two binary votes, nonzero vote entropy is exactly cross-model disagreement. Majority-low confidence routes an item when at least two of three Opus calls report low confidence. These rules use predefined operating points. The additional calls supply routing signals, while retained-set agreement uses the primary Opus preference.

Learned router. We train one logistic regression per dataset to predict whether the primary Opus preference agrees with the human, using Q , v_1 , and v_2 as inputs. All-MiniLM-L6-v2 separately embeds the three texts; long inputs are split into non-overlapping 256-token chunks whose mean-pooled embeddings are averaged and normalized. Symmetric pair features and nine deterministic length-/similarity features form the classifier input. In nested query-group-preserving five-fold evaluation, inner folds choose regularization by log loss and outer folds produce final predictions. Sweeping the predicted probability traces its risk-coverage curve.

AURC averages selective risk over coverage with ties handled by expected random ordering. Margin AURC is computed from the corresponding Opus margin ranking.

Inference cost. Margin adds one margin-estimation call. Cross-model agreement adds one GPT-5.5 preference call. Self-consistency and majority-low confidence share two additional Opus calls, for three Opus preference runs total. The learned router adds training and local encoder inference but no routing-time LLM call.

E Rubric Scoring

Each example in the T&A Calc dataset carries an expert-reviewed four-dimensional rubric, which scores each answer *pointwise*. The rubric spans four dimensions:

- Computational Accuracy (*pass / fail*)
- Input-Source Correctness (*pass / fail*)
- Task Completeness (*pass / partial / fail*)
- Methodology (*pass / partial / fail*)

The fact / non-fact distinction is made per rubric item: as it grades each item, the LLM judge also labels it as a checkable-*fact* criterion (a specific number, threshold, date, statute, citation, yes/no, or single correct procedure) or a *non-fact* judgment criterion (completeness, scope, clarity, organization, tone, or usefulness), and every dimension carries a mix of both.

Each item additionally returns the LLM judge’s confidence (0–100) and a severity (1–4). One LLM call grades all items of a single (answer, dimension); the grading prompt is the production per-dimension prompt with this addendum:

```
For EACH rubric item, give "verdict", a brief "explanation",
and three more fields:
"fact_nonfact" ("fact"|"non-fact") -- is this a checkable-
fact criterion (a specific number/threshold/date, a statute/
citation, a yes/no, a single correct procedure) or a judgment
criterion (completeness/scope, clarity, organization, tone,
usefulness)?
"confidence" (0-100) -- confidence in THIS item's verdict.
"severe" (1-4) -- importance of this criterion to the
OVERALL answer.
```

Scoring rules. A verdict maps to a coefficient (*pass*=1, *partial*=0.5, *fail*=0). The pointwise dimension score is a points-weighted pass rate, $\sum_i \text{verdict}_i p_i / \sum_i p_i$, averaged over the four dimensions; the pairwise winner is the higher-scoring answer. We compare the plain points-weighted score ($\mathbf{v} \times \mathbf{p}$) with a severity-weighted

Judge	all items		by item type	
	$\mathbf{v} \times \mathbf{p}$	$\mathbf{v} \times \mathbf{p} \times \mathbf{s}$	fact	non-fact
Opus 4.6	69.2	69.2	70.0	69.4
GPT-5.5	67.3	67.3	71.2	63.2

Table 10: T&A Calc rubric pointwise agreement (%) with the human expert pairwise preference, by scoring rule. $\mathbf{v} \times \mathbf{p}$: verdict coefficient $\times$ points; $\mathbf{v} \times \mathbf{p} \times \mathbf{s}$: additionally $\times$ severity. Severity weighting leaves the pairwise winner, and hence the agreement, unchanged for both judges (no confidence term is used). Scoring from only the items the judge labels *fact* versus only the *non-fact* items shows the fact items predict the preference at least as well, and markedly better for GPT-5.5 (71.2 vs. 63.2); $n=38\text{--}52$ pairs are decided (ties on a score excluded).

variant ($\mathbf{v} \times \mathbf{p} \times \mathbf{s}$); both exclude confidence. Severity weighting preserves the pairwise winner and agreement (69.2% Opus, 67.3% GPT-5.5; Table 10); we therefore report the simple $\mathbf{v} \times \mathbf{p}$ score in the body. Scoring from only the items that the LLM judge labels *fact* predicts the human expert preference at least as well as the non-fact items, and clearly better for GPT-5.5, consistent with the fact-grounded but not fact-only stage of the main text.

The LLM judge is more confident on fact items.

On the T&A Calc rubric items, where each graded item carries both a fact/non-fact label and a 0–100 self-confidence, both judges report markedly higher confidence on fact items than on non-fact items: Opus averages 93.8 vs. 88.4 (a high-confidence rate, ≥ 90 , of 82.9% vs. 59.7%) and GPT-5.5 97.4 vs. 95.7 (93.5% vs. 90.0%). The bulk of rubric items are fact criteria ($\approx 89\text{--}92\%$), while overall preferences still depend on judgment items (§5.1). The LLM judge is surest on checkable items and least sure on the judgment items that decide close calls.

F Error Validation

The error-validation result (§4.2) is a *precision* check on the errors the rubric judge flags, separate from the pairwise-agreement number. We draw 20 questions from the LLM judge runs; for each, the LLM judge’s flagged rubric issues on two answers are listed, giving 1,226 listed issues. A human expert determines whether each issue is genuine.

Human experts confirm the flagged error on $1,036/1,216 = 85.2\%$ of the issues carrying a verdict; the remaining 10 issues were left unmarked.

Dataset	≤ 15	16–30	31–50	51–70	71–90	>90
Opus 4.6						
T&A QA	52.9 (46,60)	65.0 (59,70)	78.6 (66,87)	85.0 (77,91)	86.8 (75,94)	—
T&A Calc	60.0 (31,83)	65.2 (45,81)	72.7 (43,90)	100 (68,100)	—	—
MT-Bench	56.7 (44,68)	67.4 (60,74)	73.8 (65,81)	84.3 (79,89)	95.9 (92,98)	98.5 (95,100)
HelpSteer3	55.9 (47,65)	64.7 (59,70)	79.7 (72,85)	90.8 (86,94)	97.1 (90,99)	—
SHP-2	59.3 (41,76)	60.3 (53,67)	70.3 (61,78)	79.0 (71,85)	98.6 (95,100)	100 (95,100)
GPT-5.5						
T&A QA	48.8 (43,55)	60.4 (54,66)	76.6 (67,84)	79.6 (67,88)	85.5 (74,92)	83.3 (44,97)
T&A Calc	61.1 (39,80)	38.9 (20,61)	75.0 (41,93)	—	—	—
MT-Bench	49.5 (40,59)	66.3 (59,73)	84.3 (78,89)	92.2 (85,96)	97.4 (94,99)	97.7 (94,99)
HelpSteer3	53.8 (46,62)	73.0 (68,78)	82.4 (77,87)	89.3 (82,94)	95.7 (90,98)	—
SHP-2	39.0 (26,54)	49.6 (41,58)	55.3 (46,64)	76.9 (68,84)	92.0 (87,95)	100 (95,100)

Table 11: Per-bucket judge–human agreement % with 95% Wilson intervals ($_{(lo,hi)}$), for both judges (omitted from Fig. 1 for legibility). Buckets with fewer than five pairs are left empty (§3.3). Masking is per judge: a sparse high-margin cell can be populated for one judge and not the other (e.g. the >90 bucket on T&A QA, $n=6$ under GPT-5.5 but $n=2$ under Opus 4.6), and the small T&A Calc set leaves several high-margin cells unpopulated under both.

Residual disagreements concentrate in a few questions and predominantly reflect rubric-quality issues, such as over-strict or mistaken gold expectations. The high confirmation rate supports the item-level reliability of flagged errors, including hard pairwise cases.

G Per-Bucket Reliability

Per-bucket intervals. Table 11 gives the 95% Wilson interval on each margin bucket’s agreement for both judges. Overall agreements additionally carry a question-clustered bootstrap interval where pairs share a question (T&A QA, T&A Calc, SHP-2); HelpSteer3 has one pair per question, so no clustering applies.

H Margins Erode Over Time

On T&A QA, whose pairs carry collection timestamps, the mean answer A/B margin more than halves over roughly six months, and its fact-margin and the LLM judge’s agreement fall in step (Figure 6). As the underlying models improve, the two candidate answers to a question become harder to tell apart, pushing more comparisons into the low-margin tail where the LLM judge is least reliable. The hard-case slice grows over time. This longitudinal evidence comes from T&A QA, the dataset with timestamps; the external sets are single snapshots.

I Results on Additional Datasets

Beyond the main five, we also ran the experiment and analysis pipeline on four additional public

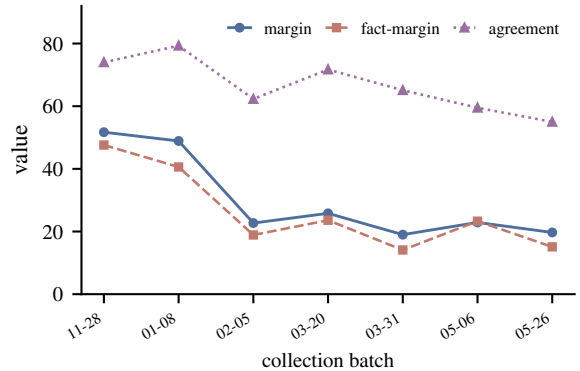


Figure 6: T&A QA over successive collection batches (as base models improve): the answer A/B margin, its fact-margin, and judge–human expert agreement all decline together.

datasets. We report the full per-dataset numbers for these additional public sets here.

I.1 Additional Public Sets

These four further public datasets use Opus 4.6 to map how far the margin trend reaches across domains, answer formats, and difficulty (Table 12). Each is reduced to the common schema (Q, v_1, v_2, p) with a native preference and, for the multiple-choice sets, a randomized gold side, making a fixed-position judge score near 50%. **LLMBar** is an adversarial meta-evaluation set for LLM judges (one output is objectively better, with splits where the worse one is crafted to look superficially better). **RewardBench** is a chosen-vs-rejected assistant-preference benchmark over chat, safety, and reasoning (we drop its `llmbar-*` subsets, which duplicate LLMBar). **CaseHOLD**

Dataset	Queries / pairs	Domain	Answer length	Preference source
LLMBar	419 / 419	instruction-following (adversarial)	short, ~ 0.5 k ch.	gold better output
RewardBench	382 / 382	assistant: chat / safety / reasoning	medium, ~ 0.9 k ch.	chosen vs. rejected
CaseHOLD	600 / 600	legal case holdings	short, ~ 0.2 k ch.	correct holding vs. distractor
CFA	436 / 436	finance exam (CFA L1/L2)	short, ~ 0.1 k ch.	correct option vs. distractor

Table 12: The four additional public sets (Appendix I.1), judged by Opus 4.6 only, in the format of Table 1. Each is reduced to the common pairwise schema (Q, v_1, v_2, p) with a native preference; the multiple-choice sets (CaseHOLD, CFA) pair the gold option against one distractor and randomize the side.

turns legal multiple-choice case *holdings* into one correct-vs-distractor pair each, and **CFA** does the same for CFA-exam questions (correct option vs. the longest distractor).

I.2 Margin Trend and Routing

Dataset	agr.	ρ	cov@85
LLMBar	77.1	0.34	61
RewardBench	91.4	0.35	100
CaseHOLD	92.7	0.35	100
CFA	95.0	0.11	100

Table 13: Additional public sets, Opus 4.6: overall judge–human agreement (%), the bin-free margin–agreement rank correlation ρ , and the coverage routing retains at the 85% target. Mirrors Table 3; $\rho > 0$ on every set.

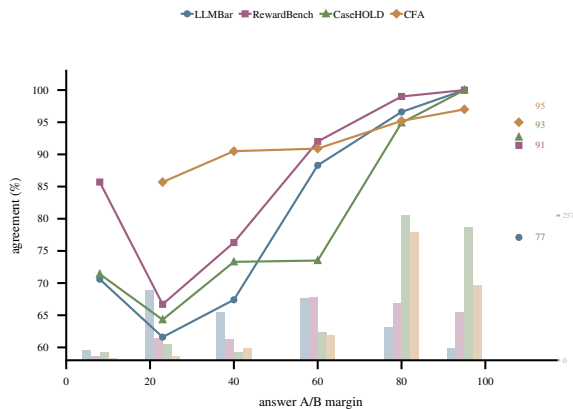


Figure 7: Answer A/B margin and agreement on the four additional public datasets. Agreement rises with the A/B margin; light bars show pairs per bucket and right-edge dots show overall agreement.

Agreement rises with the A/B margin on all four additional public datasets (Figure 7), and the bin-free correlation ρ is positive throughout (Table 13). RewardBench, CaseHOLD, and CFA sit near a high-agreement ceiling; CFA’s limited range compresses its positive ρ . The held-out routing evaluation uses the five main datasets.

I.3 Fact Content

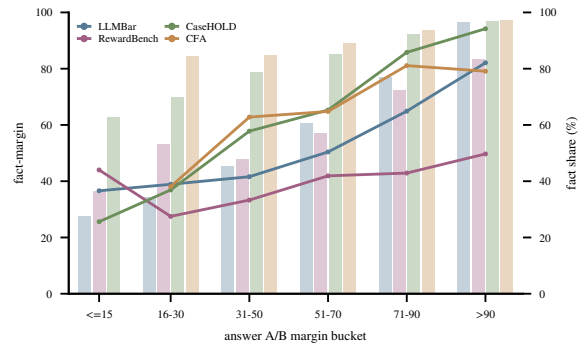


Figure 8: Answer A/B margin and fact-share/fact-margin on the four additional public datasets: the per-dataset fact-share (bars) and fact-margin (lines) both rise with the A/B margin.

The fact pattern of the main analysis also holds here (Figure 8): both the fact-share and the independent fact-margin rise with the A/B margin on every set.