

# THE SAME-FAMILY HALO: A GOLD-FREE AUDIT OF SOURCE-DEPENDENT AGREEMENT IN LLM SILVER LABELING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Scalable analysis of long-form human–human, human–agent, and agent–agent interactions requires reliable supervision. Large language models (LLMs) provide a practical source of silver labels, but agreement among models can reflect shared labeling preferences rather than independent confirmation. We investigate this dependence in an enterprise pipeline that uses 76,000 silver-labeled customer-service interactions to train classifiers operating over tens of millions of conversations, with a separate human-labeled holdout of 924 examples. We introduce a gold-free audit: holding each labeler’s predictions fixed, we vary the model supplying the reference labels and measure changes in agreement without consulting human annotations. Across 48 model–prompt–rendering configurations, we observe source-dependent agreement that extends beyond exact self-comparisons to sibling models. In a representative comparison, agreement with sibling-model labels exceeds agreement with three cross-family sources by 3.7–3.9 percentage points. To mitigate this effect, we propose combinatorial silver-label construction, assigning each example to a randomly selected model–prompt tuple — randomizing the prompt as well, since prompt choice is itself a first-order driver of label quality. Under this construction, the source-specific agreement advantage seen with fixed-source labels is directionally reduced in our evaluation, though not to statistical significance at our sample size. The method distributes supervision across configurations while preserving exactly one inference call per example. Our central — and demonstrated — finding is that same-family consensus can reflect a source-dependent agreement pattern that resembles independent confirmation while not providing it. The gold-free audit makes this dependence measurable even where human reference labels are scarce, and randomized construction offers a practical, single-call route toward mitigating it in scalable conversation analytics.

## 1 INTRODUCTION

The conversational data an operator must make sense of — human–human call transcripts, and increasingly human–agent and agent–agent interactions — now arrives at a scale no annotation team can label directly, yet the classifiers that analyze it still need supervision. When gold-standard annotation from humans is too slow or too expensive to cover the data a model needs, a common substitute is to let one or more large language models (LLMs) label the data instead, producing what are often called silver labels. Those silver labels then train a downstream classifier, benchmark a system, or decide which examples a human ever reviews. As this practice spreads, a single question governs whether the resulting system is trustworthy: how does a practitioner know the silver labels are right, when there is, by definition, little or no gold to check them against?

The field’s default answer is agreement. If several models, prompts, or runs converge on the same label, that convergence is read as a sign of correctness, and disagreement is read as a sign of difficulty or ambiguity. This intuition borrows from a long tradition in which individual voters, when aggregated, are more reliable than any single voter. The intuition has a hidden premise: that the voters are, in fact, independent. When the voters are large language models trained on overlapping corpora with overlapping objectives, that premise is exactly what is in doubt.

This paper asks whether agreement among LLM labelers means what practitioners assume it means, and finds that it does not — at least not uniformly. Agreement is inflated in a structured, predictable way that depends on the *relationship between a labeler and the source of the silver label its output is checked against*. When a model’s label output is compared to a silver label produced by a model of its own family, the two agree more than that same labeler’s agreement with cross-family silver sources would imply under an independence interpretation. The effect is not merely that strong models resemble each other; it is that the resemblance is organized along family lines — most clearly between sibling models — and that it contaminates the very agreement signal practitioners rely on to certify silver labels.

Establishing this cleanly is harder than it sounds. The obvious way to measure labeler bias is to compare a labeler’s agreement with model-written labels against its agreement with gold labels. But if the gold labels are themselves wrong on some fraction of cases — and we find that they are, sometimes flatly so — then this comparison is made against a broken ruler, and a labeler that “beats” the gold could be either sharing a human-invisible blind spot or simply being more correct than a fallible annotator. No amount of care with that comparison can separate the two. Our central move is not to discard the human comparison but to make the *bias claim* independent of it: we hold a single labeler fixed and vary only whose labels it is scored against, so that within this contrast the labeler’s own competence — and any dependence on gold — cancels. The residual therefore measures *source-dependent agreement*: how the same labeler’s agreement changes with the identity and relationship of the silver-label source. We still report each labeler’s agreement with gold labels, but treat it as motivation, not as evidence of bias. That residual — the observed excess agreement between a labeler and a distinct same-family silver source relative to cross-family sources — is what we call the same-family halo.

**Claims and non-claims.** We make two positive claims — (i) a gold-free same-family halo exists and is significant, and (ii) in our setting, model-family diversity provides more error decorrelation than prompt diversity — with a direct and, we hope, immediately usable consequence: a team assembling a panel of LLM labelers to vet silver labels should allocate more of its diversity budget toward model families than toward prompt engineering on a single model. Prompt variations of one model tend to fail the same calls, adding less *independent* signal than variation across families even when they differ in overall accuracy, while same-family agreement carries a halo precisely where a practitioner might otherwise interpret consensus as independent confirmation. We do *not* claim that the halo’s cause is lineage *per se* rather than the capability proximity siblings tend to share, that the general silver–gold divergence is bias (the broken-ruler problem), or that a de-biased panel raises downstream accuracy (untested). Nor is this a claim that prompt quality does not matter: prompt choice is a first-order determinant of a single labeler’s accuracy (Finding 3), and the family-over-prompt guidance concerns only how to spend a decorrelation budget — not whether getting the prompt right is worth the effort.

## 2 BACKGROUND & MOTIVATION

**The label taxonomy.** Our setting is the categorization of inbound customer-service calls in a telecommunications business. Call reasons are organized as a multi-layer nested hierarchy, running from a few coarse intent families down to fine-grained leaf categories that name the specific reason a customer called. The finest level contains **hundreds of fine-grained call-reason categories**, and it is where operational decisions are made and where the task studied here operates. This cardinality matters for everything that follows. With hundreds of near-neighboring categories, many of them semantically adjacent (e.g., a *payment* vs. a *payment extension*), the labeling problem is genuinely hard for humans and models alike, disagreement is common, and consensus is scarce enough that practitioners are tempted to trust it wherever it appears.

**Why silver labels, and why their trustworthiness is the bottleneck.** Human annotation at this granularity is slow and expensive, and the label space drifts as products change, so covering production volume with gold labels is infeasible. LLM-generated silver labels are the practical alternative: in the pipeline we study, roughly 76,000 silver-labeled interactions train a downstream classifier that then operates over tens of millions of conversations, with only a small human-labeled holdout

( $N = 924$ ) available to check the silver against. The downstream system inherits any bias in the agreement signal used to accept its silver labels.

**Related work.** The closest prior result establishes that, across a very large model population, stronger language models make more *correlated* errors, and that they do so even across distinct architectures and providers — shared vendor lineage alone does not explain the effect (Kim et al., 2025). We build directly on that finding while distinguishing our contribution from it: that work explains the *cause* of error correlation in general model behavior, whereas we identify a *specific, actionable consequence* of correlation in the silver-labeling workflow — the same-family halo — and, crucially, we detect it *without reference to gold*, which its strength-driven account (measured against reference answers) does not require and our broken-ruler analysis shows is unsafe in our setting. The same-family halo is a concrete mechanism by which algorithmic monoculture surfaces in a labeling pipeline (Kleinberg & Raghavan, 2021; Bommasani et al., 2022; Hedden & Raghavan, 2026). Ensembles benefit from decorrelated errors rather than member count (Krogh & Vedelsby, 1994; Breiman, 2001). We operationalize this principle with  $M_{\text{eff}}$  and test directly whether family or prompt diversity supplies more independence.

A parallel literature on LLM-as-a-judge documents that models prefer their own or their family’s generations (Panickssery et al., 2024; Wataoka et al., 2024), part of a broader account of biases in LLM evaluators (Zheng et al., 2023; Liu et al., 2023; Koo et al., 2023; Wang et al., 2024b). Our setting is adjacent but differs in a way worth stating: rather than prompting a model to rate a candidate answer, we have each model label the call independently and read its agreement with a silver label as an implicit endorsement — so the models we study act as *labelers* whose concordance we interpret, not as judges issuing an elicited verdict. Our halo is thus that self-preference relocated to the silver-labeling task and identified gold-free.

Finally, our finding is a caution for methods that consume LLM labels. Weak supervision and consensus label models estimate latent truth from multiple noisy sources, classically under conditional-independence assumptions (Dawid & Skene, 1979; Ratner et al., 2016; 2017); because LLM sources violate independence in a structured, family-indexed way, a same-family consensus is over-confident by a quantifiable margin. The practice of training downstream models on LLM-generated labels (Gillardi et al., 2023; Hinton et al., 2015; West et al., 2022; Burns et al., 2024; Bansal et al., 2025) inherits any bias in those labels, and cost-aware and encoder–LLM cascade labeling (Valdes Gonzalez, 2026; Chen et al., 2023) optimizes the price of producing them; we sit upstream, asking whether the labels such pipelines agree on can be trusted at all. For more detail on each cited work, see Appendix B.

**Problem statement.** We study a panel of LLM labelers evaluating silver labels for a categorization task over hundreds of fine-grained call reasons. We ask three questions. (1) *Is model agreement a trustworthy signal of silver-label correctness, or is it biased by the relationship between the labeler and the label’s source?* (2) *If biased, can that bias be identified in a way that survives the fact that the gold labels are themselves imperfect?* (3) *What should a practitioner do differently when constructing a panel to vet silver labels?*

### 3 PROBLEM FORMULATION

**Task and data.** Each call  $x$  has a latent true leaf category  $y^* \in \mathcal{Y}$ , where  $\mathcal{Y}$  is the finest level of the taxonomy and  $|\mathcal{Y}| = K$  runs to the hundreds. We observe a human “gold” label  $g(x)$ , an imperfect estimate of  $y^*$ , on  $N = 924$  calls. A *labeler*  $L = (m, p, r)$  is specified by a model  $m$ , a prompt  $p$ , and a rendering  $r$  (see Appendix C and §4), mapping a call to a predicted category  $L(x) \in \mathcal{Y}$ . Labelers belong to model *families*; write  $\text{fam}(L)$  for the family of  $L$ .

**Silver sources and the answer-key relation.** To score a labeler’s accuracy we compare its predicted labels against a reference set of labels treated as correct — an *answer key*, in the grading sense. When that answer key is the gold labels, we recover the usual accuracy; when it is instead a set of model-generated labels, we call it a *silver answer key*. Formally, both kinds of reference are label-assigning functions: gold  $g$  maps each call  $x$  to its human label  $g(x)$ , and a *silver source*  $S$  — a single model under one prompt, or a panel vote — maps  $x$  to a model-generated label  $S(x)$ . A single reference function  $R \in \{g\} \cup \{S\}$  — gold, or one silver source — supplies the answer key

$\{R(x)\}$ , and a labeler  $L$ 's accuracy against it is

$$\text{acc}_R(L) = \frac{1}{N} \sum_x \mathbb{1}[L(x) = R(x)]. \quad (1)$$

$\text{acc}_g$  measures agreement with  $g(x)$ ;  $\text{acc}_S$  measures agreement with another model's output. Thus  $L(x)$  is the prediction under audit, while  $S(x)$  is only the silver answer key used to score it.

**The silver–gold gap, and why zero is not the neutral point.** Define  $\Delta(L, S) = \text{acc}_S(L) - \text{acc}_g(L)$ . Partitioning on whether the labeler matches gold gives the exact decomposition

$$\Delta(L, S) = \underbrace{P(L = S \neq g)}_{\text{correlated error}} - \underbrace{P(L = g, S \neq g)}_{\text{silver-only error}}, \quad (2)$$

verified to residual  $1.3 \times 10^{-16}$  on our data. The first term — labeler and silver make the *same* departure from gold — is what a bias account cares about. But  $\Delta = 0$  arises whenever the two terms merely balance, so the sign of the gap says nothing directly about bias; and because  $g$  is imperfect,  $P(L = S \neq g)$  conflates shared model bias with cases where the model pair is right and  $g$  is wrong. This is the broken-ruler problem:  $\Delta$  measured against gold can motivate a bias hypothesis but cannot establish it.

**The halo: a within-labeler, gold-free contrast.** Fix a labeler  $L$  and compare its accuracy against a silver source  $S_{\text{same}}$ , where  $\text{fam}(S_{\text{same}}) = \text{fam}(L)$ , versus a cross-family source  $S_{\text{cross}}$ :

$$H(L) = \text{acc}_{S_{\text{same}}}(L) - \text{acc}_{S_{\text{cross}}}(L). \quad (3)$$

Both terms use the *same* labeler, so its competence against  $y^*$  and any dependence on  $g$  are differenced away. Instead,  $H(L)$  measures *source-dependent agreement*: the same labeler agrees differently depending on the source of the silver answer key, and a positive same-family contrast establishes that this dependence is organized along family lines. Accordingly, the quantity we report with “show” is the existence of *source-dependent agreement associated with model family*, not a claim about its underlying cause.

**Decorrelation as a second lens.** Independent of accuracy, we quantify how *jointly* labelers err. For a set of labelers let  $\bar{\rho}$  be the mean pairwise error correlation and

$$M_{\text{eff}} = \frac{n}{1 + (n-1)\bar{\rho}} \quad (4)$$

a design-effect proxy for the effective number of independent labelers in an  $n$ -labeler panel (an approximation for correlated binary error indicators, not an exact count; we lean on it for intuition). A halo predicts same-family pairs have higher  $\bar{\rho}$  (lower  $M_{\text{eff}}$ ) than cross-family pairs — a prediction on a *different* statistic than  $H$ , so agreement between the two is corroboration, not restatement.

## 4 METHODOLOGY & ANALYSIS

**Combinatorial labeler grid.** We evaluate a grid of (model, prompt, rendering) configurations on the same 924 gold-labeled calls. The grid spans 6 model families and 5 prompts under 2 rendering settings (full vs. no category definitions; see Appendix C). The prompt $\times$ rendering crossing is incomplete: 3 definition-based prompts (best-explains, own-words, taxonomic) take both renderings (6 configurations), while 2 name-only prompts (clue-hunt, distinctive-ev.) have no definitions to render and take a single setting (2 configurations), giving **8 configurations per family** and **48 labelers** in total. A separately served arm contributes three further model families that appear only as answer-key *sources* in the bias grid (below), not as labelers in the panel.

**Bias grid (the halo design).** Beyond scoring labelers against gold, each model uses a prompt that performs best when compared against the gold labels to generate output that is used in turn as the silver answer key, and every labeler's accuracy is measured against every such key (Figure 1). Because the halo  $H(L)$  is a within-labeler contrast — the same labeler graded against different silver sources — this choice fixes the reference set but is not what creates the family-indexed pattern

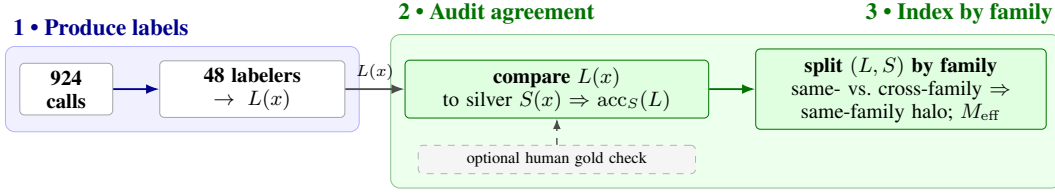


Figure 1: **The gold-free audit pipeline.** Each of 48 labelers produces  $L(x)$  on the same  $N = 924$  calls; each is then scored against a silver key  $S(x)$  from another labeler to give agreement  $\text{acc}_S(L)$ , with three further families acting only as sources. Splitting the  $(L, S)$  pairs by their model-family relation gives the same-family halo ( $M_{\text{eff}}$  summarizes panel decorrelation). Human gold  $g(x)$  serves only as an optional check during agreement auditing.

the contrast isolates; any decent prompt would serve. The diagonal of the resulting model  $\times$  labeler matrix (labeler graded on its own family’s silver) versus its off-diagonal entries (cross-family silver) instantiates the within-labeler contrast  $H(L)$  above. Two leave-out rules keep the contrast honest: the trivial self-cell — a labeler scored against its own exact output, which is 100% by construction — is dropped, and when the silver answer key is a panel the labeler is removed from that panel’s vote before grading. The same-family term is therefore a *sibling* comparison: a labeler is graded against a same-family but distinct model’s silver (e.g., Haiku against Sonnet’s answer key), never against itself. What it does not remove is the capability distance between a labeler and the silver source it is scored against: because same- and cross-family sources here are not matched on strength, this design instantiates the lineage-vs-capability confound rather than resolving it (§7).

#### Statistical machinery.

- *Paired significance.* Because labelers share calls,  $H$  is tested with the paired **McNemar** test (two-tailed); the headline Haiku contrast is reported with its exact  $p$ -value.
- *Multiplicity.* Across the full grid of family contrasts we control family-wise error with **Bonferroni** for the confirmatory headline and the **Benjamini–Hochberg** FDR at  $q = 0.05$  for the exploratory grid.
- *Uncertainty.* All accuracies and the churn/stability metrics carry **bootstrap** confidence intervals (500 draws).
- *Decorrelation.* Panels are characterized by mean pairwise error correlation  $\bar{\rho}$ , double-fault rate, Yule’s  $Q$ , ensemble ambiguity, and the derived  $M_{\text{eff}}$ .
- *Gap decomposition.* The exact split of  $\Delta$  into correlated-error and silver-only-error mass (§3) is computed per labeler–source pair.

## 5 RESULTS & FINDINGS

**Finding 1 — The same-family halo (headline, gold-free).** Holding the labeler fixed, a labeler scores higher against its own family’s silver than against another family’s. For the Haiku labeler the contrast is  $H \approx +3.8\text{pp}$  (own-family vs. cross-family silver), McNemar  $p \approx 3 \times 10^{-4}$ , surviving Bonferroni correction (Table 1). Across the exploratory grid, 17 of 20 family contrasts are significant and 16 are positive (Figure 2) — the halo is directional, not noise. These 20 are not independent replications — they share the same 924 calls and overlapping models and prompts — so we read them as repeated internal contrasts that corroborate the headline, not as twenty separate confirmations of it.

**Finding 2 — Decorrelation corroborates the halo (gold-free, orthogonal statistic).** Same-family labeler pairs fail the same calls more than cross-family pairs: the Sonnet $\leftrightarrow$ Haiku pair has mean pairwise error correlation  $\bar{\rho} = 0.616$  ( $M_{\text{eff}} = 1.24$ ), versus  $\bar{\rho} = 0.532$  ( $M_{\text{eff}} = 1.31$ ) for the average Sonnet $\leftrightarrow$ cross-family pair. Because this is a different statistic than Finding 1’s accuracy contrast — though computed from the same models and 924 calls — the agreement between them is orthogonal corroboration.

Table 1: The same-family halo, isolated in a single labeler (gold-free). Holding Haiku-4.5 as a fixed anchor, its agreement with the same-family (Sonnet-4.6) silver key is compared against its agreement with three cross-family keys. The halo  $H$  (own-family – cross-family agreement) reflects the answer key’s source, not the labeler’s own competence. Each  $H$  is a paired within-item contrast;  $p$ -values are McNemar’s test, with the Bonferroni-corrected value in parentheses.

| Silver answer key | Family (rel. to labeler) | Haiku agreement | $H$ (own – cross) | McNemar $p$ (Bonf.)                           |
|-------------------|--------------------------|-----------------|-------------------|-----------------------------------------------|
| Sonnet-4.6        | Anthropic (same)         | 0.838           | —                 | —                                             |
| Gemini-2.5-Pro    | Google (cross)           | 0.799           | +0.039            | $2.6 \times 10^{-4}$ ( $7.9 \times 10^{-4}$ ) |
| GPT-5.5           | OpenAI (cross)           | 0.799           | +0.039            | $2.6 \times 10^{-4}$ ( $7.9 \times 10^{-4}$ ) |
| Grok-4.6          | xAI (cross)              | 0.801           | +0.037            | $7.6 \times 10^{-4}$ ( $2.3 \times 10^{-3}$ ) |

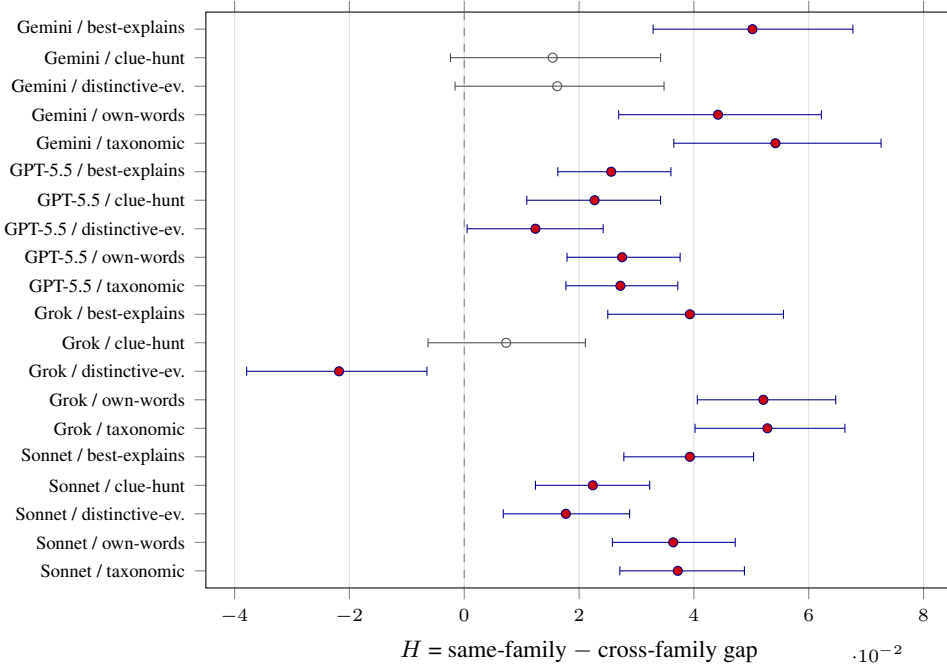


Figure 2: **The halo generalizes across 20 (silver, prompt) contrasts.** Each marker shows  $H$  with its 95% bootstrap CI for each frontier silver answer key (Gemini, GPT-5.5, Grok, Sonnet) and prompt. The dashed line marks  $H = 0$ . Under Benjamini–Hochberg control at  $q = 0.05$ , 17 of 20 contrasts are significant (filled markers) and 16 are positive (the halo, up to +5.4 points). The single significant *negative* contrast (Grok / distinctive-evidence,  $H = -2.2$  points) is one prompt where same-family labelers judge each other more harshly, consistent with the two-tailed framing. Every  $H$  holds the labeler fixed and cancels the gold dependence, making this figure gold-free.  $N = 924$ .

**Finding 3 — Family diversity decorrelates errors more than prompt diversity.** Five prompts on a single model give  $M_{\text{eff}} \approx 1.19$ –1.41 (Sonnet the most redundant at 1.19 — its five labelers act like roughly one vote), whereas one prompt across six families gives  $M_{\text{eff}} \approx 1.44$ –1.58. The family range sits entirely above the prompt range, so family diversity is the better of the two levers; but the margin is narrow, and neither buys much in absolute terms — the full grid of 48 labelers still caps at  $M_{\text{eff}} \approx 1.67$  (Table 2). For a panel’s *independence* budget, families are the better spend.

This does not make prompt choice dispensable. It barely moves a strong labeler but is a first-order driver of label quality for a weak one: across the five prompts the gold-accuracy swing runs from 3.2 points on the strongest model to 18.0 on the weakest (Table 3), and across all nine models mean accuracy and prompt-induced spread correlate at  $r = -0.97$ . The model  $\times$  prompt interaction accounts for  $\approx 5.1\%$  of accuracy variance, so a prompt’s pull is not uniform but concentrates where the model is weakest. Prompt choice thus governs quality — most of all when the panel leans on

Table 2: Error decorrelation across panel constructions (gold-free).  $\bar{\rho}$  is the mean pairwise error correlation among a panel’s labelers; the effective number of independent labelers is  $M_{\text{eff}} = n/(1 + (n - 1)\bar{\rho})$ ;  $M_{\text{eff}}$  near 1 means the panel votes like a single labeler. Same-family pairs are more redundant than cross-family pairs, and prompt diversity buys somewhat less independence than family diversity; the full grid of 48 labelers still behaves like fewer than two. Correlations are computed on shared items and are independent of the accuracy contrast in Table 1.

| Panel construction                       | $\bar{\rho}$ | $M_{\text{eff}}$ |
|------------------------------------------|--------------|------------------|
| Same-family pair (Sonnet, Haiku)         | 0.616        | 1.24 (of 2)      |
| Cross-family pair (Sonnet, other family) | 0.532        | 1.31 (of 2)      |
| One model, five prompts                  | —            | 1.19–1.41 (of 5) |
| One prompt, six families                 | —            | 1.44–1.58 (of 6) |
| Full grid (48 labelers)                  | —            | 1.67             |

Table 3: Prompt sensitivity by model, ordered by strength. Each model takes five prompts, each at a single rendering (the definition-consuming `full` rendering where the prompt uses it, else the bare `none` rendering); the range reported isolates the prompt axis rather than mixing in rendering. *Spread* is the resulting range of gold accuracy (best prompt minus worst); *mean acc.* averages the model’s five prompt variants.

| Model   | Mean acc. | Prompt spread (pt) |
|---------|-----------|--------------------|
| GPT-5.5 | 0.776     | 3.2                |
| Sonnet  | 0.741     | 4.8                |
| Grok    | 0.745     | 7.2                |
| Gemini  | 0.732     | 8.2                |
| gpt-oss | 0.690     | 9.0                |
| Haiku   | 0.685     | 12.2               |
| Mistral | 0.643     | 14.3               |
| Llama   | 0.642     | 15.9               |
| Nova    | 0.632     | 18.0               |

weaker models — even though prompt diversity adds less to error decorrelation than model-family diversity. The two roles pull apart, and the gap matters downstream: a silver pipeline that locks in one prompt inherits that prompt’s quality profile, an exposure the construction of §6 addresses by randomizing the prompt rather than by treating prompts as interchangeable.

**Finding 4 — General silver–gold divergence (motivation only).** Every labeler scores higher against LLM silver than against gold labels (Figure A.1; per-model mean gaps +0.02 to +0.10, individual labelers up to +0.13). A manual review of disagreement cases suggests part of this is genuine gold error — human labels that are flatly wrong where the model is right — so the divergence is consistent with both shared bias and superior model competence and is reported as motivation, not as evidence of bias.

**Finding 5 — Structured, taxonomy-local disagreement.** Where labelers disagree, the disagreement collapses to a few clusters of semantically adjacent categories, suggesting a portion of the “error” is overlapping category definitions — or too little signal in the call itself to separate them — rather than model failure.

## 6 MITIGATING MODEL–PROMPT BIAS BY COMBINATORIAL SILVER-LABEL CONSTRUCTION

**What this section contributes.** Randomizing the silver source per call equalizes each configuration’s exposure by design, at exactly one inference call — the same as any fixed source (shown below). Its value as a design rule rests on the halo of §5 and on these accounting properties, not

Table 4: De-biasing by randomized construction: a diversity ladder (preliminary, gold-referenced).  $D$  is the accuracy-residualized between-family dispersion of the silver–gold gap — lower is less biased; rung A is a single fixed (model, prompt) silver source, the reference for every  $\Delta D$ . Every diversified construction reduces  $D$ , and the lowest values cross model families, but at  $N = 924$  calls all  $\Delta D$  bootstrap intervals include zero. Rungs A–D are panel silver; the final three rows are per-observation randomized variants, the last being the observed floor.

| Rung | Construction                                | $D$ ( $\downarrow$ ) | $\Delta D$ vs. A (95% CI)  |
|------|---------------------------------------------|----------------------|----------------------------|
| A    | Single fixed source (one model, one prompt) | 0.0172               | —                          |
| B    | Same model, many prompts                    | 0.0161               | −0.0010 (−0.0036, +0.0015) |
| C    | Same family, many models                    | 0.0163               | −0.0009 (−0.0036, +0.0015) |
| D′   | Cross-family, one prompt                    | 0.0156               | −0.0016 (−0.0043, +0.0010) |
| D    | Cross-family and many prompts               | 0.0166               | −0.0005 (−0.0032, +0.0021) |
| —    | Per-observation prompt-diverse (robustness) | 0.0156               | −0.0015 (−0.0042, +0.0010) |
| —    | Per-observation family-diverse (robustness) | 0.0163               | −0.0009 (−0.0036, +0.0017) |
| —    | Per-observation both-diverse (floor)        | 0.0153               | −0.0019 (−0.0045, +0.0006) |

on clearing significance on our gold set; the dispersion evidence itself is directional, and we state it with “suggest” below.

**The idea.** The halo is indexed by the *relationship* between a labeler and the silver’s source, so any *fixed* silver source hands a systematic advantage to its own family: whichever model (or family) authors the answer key is exactly the one whose siblings are flattered. This suggests a construction that removes the fixed relationship. Rather than fix one model and one prompt as the silver source, we draw the source at random *per call* — assigning each call’s silver label from a (model, prompt) configuration sampled from the pool, a bagging construction we call *combinatorial silver-label construction*. The prompt is randomized alongside the model for a reason distinct from decorrelation: §5 shows prompt diversity adds less error independence than model-family diversity, but prompt choice is a first-order driver of label quality (Table 3), so fixing one prompt would bind the whole corpus to that prompt’s quality profile. Randomizing it equalizes prompt exposure just as randomizing the model equalizes family exposure; rendering is held fixed and is not one of the randomized axes. Over a corpus, no single family or prompt is the answer key often enough to enjoy a systematic advantage.

**What it equalizes — and what it does not.** Randomized construction equalizes each family’s *exposure* as the silver source, so the between-family dispersion of bias — how much more one family is flattered than another — shrinks toward a common level. It does **not** remove the shared, correlated errors that two models of a family make together: those survive any majority vote and are invariant to how the silver is built. In other words, randomized construction de-biases in the sense of *removing any single configuration’s systematic family advantage*, not in the sense of eliminating family-shared error.

**Evidence (preliminary, gold-referenced, stated with “suggest”).** Measuring between-family bias by the accuracy-residualized dispersion of the silver–gold gap across families, every diversified construction we tested reduces it below a single fixed silver source, and the least-biased constructions are those that cross model families (per-observation both-model-and-prompt-diverse lowest, versus a single fixed source; Table 4). We report this with “suggest,” not “show”: at  $N = 924$  calls every reduction’s bootstrap confidence interval includes zero, so the effect is directionally consistent but not yet statistically established, and it is a gold-referenced quantity (it uses agreement with human gold) rather than one of the gold-free contrasts that carry our headline claim.

**Deployment at scale.** The construction is meant to run where it matters: producing silver labels for a large unlabeled pool (the roughly 76,000-interaction silver corpus of §2) by randomly assigning a model and a paraphrased prompt to each call. Because the assignment is per call and draws only configurations already in the panel, it uses exactly one model call per label — the same number of calls as any single-source silver, and a fraction of a voting panel’s, which spends  $K$  calls per label to

aggregate  $K$  configurations. The same mixing reduces between-family dispersion in our evaluation — directional only, on the gold-referenced measure detailed above. (Per-token prices differ across models, so the expected dollar cost of a random draw equals the pool’s average per-call price rather than any one model’s; the call count, which dominates at scale, is unchanged.)

## 7 CONCLUSION & BROADER IMPACT

Model agreement is the default currency for trusting LLM-generated labels, but this paper shows that its evidentiary value is source-dependent: a labeler can exhibit excess agreement with labels from a distinct same-family source relative to cross-family sources, a pattern that a gold-referenced measurement cannot cleanly identify but a within-labeler, gold-free contrast can. The evidence points to one operational rule, one construction, and one caution. The rule: **prioritize diversifying model families over prompts when building a panel to vet silver labels**, because family diversity supplies more error decorrelation; prompt choice, meanwhile, remains a first-order driver of label quality, so the two are levers for different ends rather than substitutes. The construction: **randomize the silver source per call so no single configuration is systematically flattered**. The caution: **do not treat same-family consensus as stronger independent confirmation than cross-family agreement**.

We *conjecture*, rather than demonstrate, that similar source dependence may arise wherever LLM consensus certifies labels over a large, fine-grained label space — in medical coding (ICD/SNOMED), content-moderation policy taxonomies, and the growing practice of using LLM juries to score other models: each relies on the same agreement premise, and in each the same-family halo raises the possibility that a panel drawn from one vendor’s model line will certify its own errors. Testing this prediction on an independent, publicly labeled taxonomy (e.g., a consumer-complaint or intent-classification corpus) is the most direct next step, and one that requires no new human annotation: because the halo is gold-free, it can be recomputed wherever several model families label the same items.

We close by marking the frontier this work does not cross. Because our evidence is drawn from a single task and corpus, the halo’s *magnitude* on other label spaces is untested. The ingredients it rests on — a labeler’s self-preference for its own family’s generations, and the correlated errors that same-family models share (Panickssery et al., 2024; Wataoka et al., 2024; Kim et al., 2025) — are documented beyond our setting, but whether the same-family halo itself generalizes across tasks and label spaces remains to be tested. Within our corpus the halo is estimated across the full combinatorial grid of labeler pairs (§5), not a single comparison. Two features of the panel nonetheless bound how far the family claim currently reaches. First, only two vendor lines contribute sibling models (Anthropic and OpenAI), so the within-panel same-family contrasts are concentrated rather than spread across many families; the effect is most directly evidenced for the Anthropic pair, and a roster with more multi-model families would test whether it is equally pronounced elsewhere. Second, same- and cross-family silver sources are not matched on capability, so a residual concern is that proximity in capability — not lineage *per se* — drives part of the halo. Disentangling the two, ideally by contrasting sources matched on strength, is a limitation we flag rather than resolve. We measure bias in the agreement signal; we do not demonstrate that a family-diverse panel yields a more accurate downstream model, and we do not resolve, against the imperfect gold available to us, how much of the general LLM–human divergence is bias versus competence. The randomized construction of §6 rests on the same footing: it is a design rule that is exposure-equalizing and inference-count-neutral by construction, while its empirical dispersion-reduction stays directional at our sample size — suggested, not shown. These two gaps have clear next steps: a downstream training study would test whether de-biased silver labels yield a more accurate model, and an adjudication study that replaces the broken ruler with case-level ground truth would separate bias from competence. The gold-free halo, however, stands without any of them.

## ETHICS STATEMENT

The data are human-subject customer-service call transcripts containing personally identifiable information. All records are de-identified and access-controlled; we report only genericized category names and aggregate metrics, and no raw transcripts or operator identifiers. Our finding carries a fairness dimension: the same-family halo means a single-vendor labeling panel can silently certify

its own errors, most harmfully on rare or sensitive call reasons (e.g., vulnerable-customer intents) where labels are scarce and least checkable; the family-diversity recommendation is in part a mitigation for that risk. No individual-level decisions are made from model outputs in this study.

#### REPRODUCIBILITY STATEMENT

The evaluation protocol (the  $N = 924$  human-gold set, the 48-labeler combinatorial grid, and the within-labeler halo contrast  $H$ ) is specified in §3 and §4; the bias-grid construction with its two leave-out rules, the paired McNemar / Bonferroni / Benjamini–Hochberg testing, and the decorrelation statistics ( $\bar{\rho}$ ,  $M_{\text{eff}}$ ) are given in §4; the randomized silver construction and its dispersion metric in §6. The data are real customer-service call transcripts and cannot be released; we describe all processing steps and provide the category-level statistics needed to interpret the results.

#### AI USE STATEMENT

Large language models are the object of study (the labelers) and the source of the silver labels we audit, as described in §3 and §4. LLM assistance was also used in drafting and editing this manuscript; all technical claims, experiments, and numbers were produced and verified by the authors.

#### REFERENCES

- Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.16737.
- Yijun Bian and Huanhuan Chen. When does diversity help generalization in classification ensembles? *arXiv preprint arXiv:1910.13631*, 2019.
- Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2211.13972.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2312.09390.
- Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.
- Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS)*, pp. 1–15, 2000.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences (PNAS)*, 2023. arXiv:2303.15056.
- Brian Hedden and Manish Raghavan. Algorithmic monoculture and its critics. *arXiv preprint arXiv:2604.06047*, 2026.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NeurIPS Deep Learning Workshop.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2306.02561.

- Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.07962.
- Jon M. Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences (PNAS)*, 2021. arXiv:2101.05853.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.
- Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1994.
- Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.
- Gregory Kang Ruey Lau, Wenyang Hu, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. DIPPER: Diversity in prompts for producing large language model ensembles in reasoning tasks. *arXiv preprint arXiv:2412.15238*, 2024.
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2402.05120.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2303.16634.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.13076.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment (PVLDB)*, 11(3), 2017. arXiv:1711.10160.
- Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1605.07723.
- Chenglei Si, Weijia Shi, Chen Zhao, Luke Zettlemoyer, and Jordan L. Boyd-Graber. Getting MoRE out of mixture of language model reasoning experts. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023. arXiv:2305.14628.
- Alberto Andres Valdes Gonzalez. Cost-aware model selection for text classification: Multi-objective trade-offs between fine-tuned encoders and LLM prompting in production. *arXiv preprint arXiv:2602.06370*, 2026.
- Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b. arXiv:2305.17926.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022. arXiv:2110.07178.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.05685.

## A SUPPLEMENTARY FIGURE

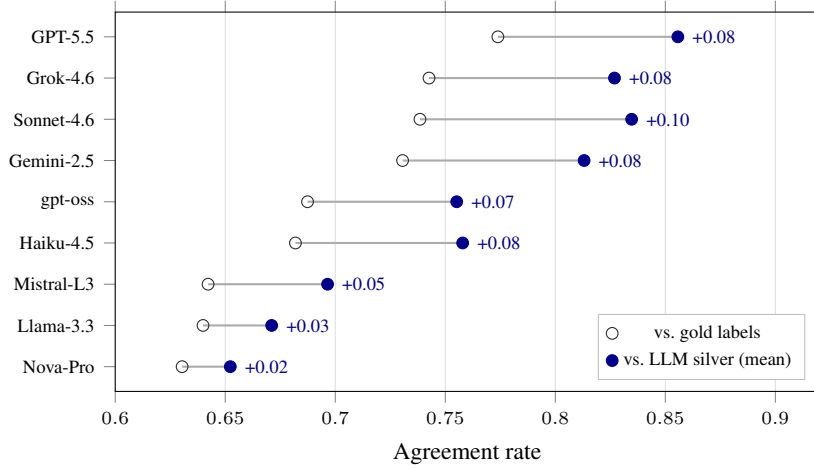


Figure A.1: **The broken ruler.** For each model (aggregated over prompts and silver answer keys), agreement with gold labels (hollow marker) versus mean agreement with LLM silver (filled marker); the segment is the silver–gold gap. Every segment points right — each model agrees more with other models’ labels than with humans (per-model mean gaps +0.02 to +0.10, up to +0.13 individually). Both endpoints are gold-referenced; the load-bearing claims rest on the gold-free halo (§5, Finding 1).  $N = 924$ .

## B EXTENDED RELATED WORK

This work makes a small number of choices that together distinguish it from the neighboring literature, and each meets a separate line of prior work. It measures label-source bias without gold, holding the labeler fixed and varying only the silver source. It treats the models as *labelers*, so a labeler’s agreement with a silver label is read as an implicit endorsement rather than as an elicited quality verdict. It indexes the effect by the model-family relation between labeler and silver source, distinguishing same-family (sibling) from cross-family pairs. It separates family diversity from prompt diversity through the effective-labeler count  $M_{\text{eff}}$ , which tracks decorrelation rather than member count. And its mitigation preserves cost, spending exactly one inference call per example. Below we trace what each line established in concrete terms, what setting it assumed, and which condition changes here. We describe every cited work at its strongest and, where its authors state a limitation in their own words, we cite that statement rather than assert one.

### B.1 AGGREGATING AND ENSEMBLING LLM LABELERS

A first line of work aggregates several LLM outputs to raise answer quality, and supplies the raw fact that ensembles of language models can beat their best member. Li et al. (2024) sample one model many times and majority-vote by cumulative similarity; a Llama2-13B ensemble rises from 0.35 to 0.59 on GSM8K, overtaking a single Llama2-70B at 0.54, with reported gains “ranging

from 12% to 24%.” Their default is homogeneous self-ensembling, and they note in conclusion that multiple calls bring “escalating costs” and that “the sampling phase can be optimized to reduce the cost ... We leave it as future work.” Wang et al. (2023) similarly sample diverse reasoning paths from a single model and marginalize over final answers, reporting absolute gains of “GSM8K (+17.9%), SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%) and ARC-challenge (+3.9%),” while conceding that self-consistency “incurs more computation cost” and applies only where the answer comes from a fixed set. Both establish that agreement among samples tracks accuracy, but the diversity they exploit is decoding or prompt diversity *within one model family*, priced at many calls per example.

Other systems ensemble across heterogeneous models or roles. Jiang et al. (2023) rank eleven open-source models’ candidates with a trained pairwise ranker (DeBERTa-400M) and fuse the top three with a Flan-T5-XL generator, reaching a GPT-Rank of 3.01 against the best single model’s 3.90 on their MixInstruct benchmark; they note the ranker “may need to call the model  $O(n^2)$  times” and that they “cannot afford large-scale human evaluation,” so ChatGPT supplies the oracle rankings. Wang et al. (2024a) layer proposer and aggregator LLMs, reaching 65.1% on AlpacaEval 2.0 against GPT-4 Omni’s 57.5%, and show a multiple-proposer mix beats a single-proposer setup at matched budget (their Table 3); they flag “high Time to First Token” from iterative aggregation. Si et al. (2023) specialise one Codex backbone into four prompted “reasoning experts” and route among them with a random forest that uses an inter-expert *agreement* feature, reaching 57.6 macro-average accuracy against the best single expert’s 49.6; their stated limits are that they “only focused on the Codex model” and on QA. Chen et al. (2023) cascade twelve commercial APIs behind a learned DistilBERT scorer, reporting up to “98%” inference-cost reduction while matching GPT-4 on HEADLINES; they state the cascade “need[s] some labeled examples” drawn “from the same or similar distribution as the test examples.”

These systems demonstrate that combining models improves output quality, and several already use agreement as a signal. The hinge is that each elicits or optimises toward a *correct* answer measured against gold, and each spends a variable and often large number of inference calls to do so. This work instead reads agreement as a gold-free bias indicator, indexes it by the labeler–source family relation, and fixes exactly one call per example. Where Si et al. (2023) and Wang et al. (2023) draw their diversity from prompts on a fixed model, this paper argues that the family axis, not the prompt axis, is what drives decorrelation.

## B.2 DIVERSITY AS DECORRELATION, NOT MEMBER COUNT

The claim that ensemble accuracy comes from decorrelated errors rather than from adding members has a long theoretical lineage, and this line supplies the formal basis for the effective-labeler count. Krogh & Vedelsby (1994) derive the error–ambiguity decomposition  $E = \bar{E} - \bar{A}$ , showing the ensemble error equals mean individual error minus an “ambiguity” (disagreement) term that “can be estimated entirely from unlabeled data.” Breiman (2001) bounds forest generalisation error as  $PE^* \leq \bar{\rho}(1 - s^2)/s^2$  in the mean inter-tree correlation  $\bar{\rho}$  and strength  $s$ , and states plainly that the mechanism behind the accuracy gain “is not obvious.” Dietterich (2000) gives the accuracy-plus-diversity condition and the binomial-voting intuition (for 21 hypotheses at error 0.3 with independent errors, the chance of eleven or more being simultaneously wrong is 0.026), leaving open the interaction between boosting and the base learner. Kuncheva & Whitaker (2003) catalogue ten diversity measures and report that their link to accuracy is weak on real data, concluding that “the problem of measuring this diversity and so using it effectively ... is still to be solved.” Bian & Chen (2019) extend the decomposition to classification ( $\bar{G} = \bar{A} - \bar{D}$ ) and show diversity helps “only ... in a few specific ranges,” leaving the multi-class case to future work.

This body of theory concerns trained classifiers on tabular or vision data under 0/1 or squared loss; “diversity” there means differing errors among independently trained models, not a relation between an LLM labeler and a silver source. This work carries the decorrelation principle, not member count, into LLM silver labeling: the effective-labeler count  $M_{\text{eff}}$  is the direct descendant of ambiguity and inter-model correlation.

Two recent works bring diversity to LLM ensembles specifically and frame the prompt-versus-model contrast this paper sharpens. Lau et al. (2024) build a training-free inference-time ensemble from a *single* base model fed a diversity-maximising subset of reasoning prompts, reporting for DIPPER

that “ $n = 9$  has close to a 10%-pt increase ( $\sim 20\%$  accuracy gain) compared to the single LLM baseline”; the authors state that DIPPER “currently does not take into account . . . model diversity.” That is precisely the axis this paper foregrounds: it spends its diversity budget on model families rather than on prompts, and treats the DIPPER position as the directly contrasted prompt-diversity emphasis.

### B.3 CORRELATED ERRORS AND ALGORITHMIC MONOCULTURE

The closest prior work asks whether different models err together. Kim et al. (2025) measure pairwise error correlation as the agreement rate when both models are wrong across large panels—349 LLMs on 12,032 MMLU items, 71 on HELM, and 20 on a resume-screening task—and report that “pairs of models agree on average about 60% of the time when both models are incorrect,” with a mean error-agreement rate of 0.423 on the HuggingFace panel and 0.6 on HELM. They regress correlation on model attributes and find that “even after conditioning on these factors, pairs of models that are more accurate individually also have more correlated errors,” and they trace downstream harms in LLM-as-judge inflation and matching-market simulations. Crucially, their correlation is measured *against reference answers*: they ask “are different LLMs more correlated with each other than they are with ground truth?” and, for the judge setting, recommend to “calibrate error metrics for each model-judge pair using ground-truth data”; on resumes they “treat the human rating as ‘ground truth’.” They also state their current metrics “treat incorrect answers identically . . . Future work should consider developing a metric,” and they do not claim to fully separate lineage from individual capability.

This is the paper’s central point of contact and its clearest hinge. The setting here is gold-free: agreement is measured among LLM labelers with no reference answer, in a fixed-label silver-labeling workflow rather than on MMLU multiple-choice or LLM-as-judge, and the effect is organised by the labeler–source *family* relation. The two works agree that capability proximity is a plausible driver of shared error, and this paper is explicit that it does not resolve lineage-per-se versus capability proximity—an open question it shares with Kim et al. (2025).

The systemic-cost framing comes from the monoculture literature. Kleinberg & Raghavan (2021) prove in a hiring model that adopting a shared, more accurate ranking can be a strictly dominant strategy yet lower social welfare (roughly “4% less” in a Gaussian three-candidate instance), and leave the multi-firm and larger-candidate-set extensions as open questions. Bommasani et al. (2022) formalise “outcome homogenization”—systemic-failure rate over its expected rate—and find a fixed-data setting “reliably shows more homogeneity than the disjoint setting,” while conceding the sharing–homogenization link is “not fully explained by our hypothesis” and that the notion’s “statistical estimation, mitigation, and connections to monoculture remain poorly understood.” Hedden & Raghavan (2026) weigh three objections to monoculture and note in passing that “algorithms are also likely to make correlated errors . . . trained on overlapping datasets,” while describing the polyculture independence assumption as “implausible” and their analysis as idealised. These are welfare and allocation arguments about shared versus diverse decision-makers; they motivate the concern that correlated behaviour has systemic costs, but they are not gold-free measurements of labeler agreement, and none is indexed by model family.

### B.4 JUDGES THAT ISSUE VERDICTS VERSUS LABELERS WHOSE AGREEMENT ENDORSES

A large literature studies LLMs as *judges* and documents their biases, and this paper draws its labelers-not-judges stance by contrast with it. In these works an evaluator is prompted to rate or choose the higher-quality response, and the resulting verdict is benchmarked against human gold. Zheng et al. (2023) establish that strong judges reach “over 80% agreement, the same level of agreement between humans” (GPT-4 vs. humans 85% against 81% human–human), yet state that their “study cannot determine whether the models exhibit a self-enhancement bias.” Liu et al. (2023) score outputs by chain-of-thought form-filling and reach a Spearman correlation of 0.514 with humans on summarisation, while flagging that G-EVAL “always gives higher scores to GPT-3.5 summaries than human-written summaries” and calling their analysis “a preliminary study on this issue.” Wang et al. (2024b) reveal positional bias—“Vicuna-13B could beat ChatGPT on 66 over 80 tested queries” by reordering—and propose calibration that raises alignment by 9.8%/14.3%. Koo et al. (2023) bench-

mark six cognitive biases across sixteen evaluators and report a human–machine agreement of 44% and egocentric self-preference above 50% for the largest models.

Self-preference in judging is the sub-thread nearest the same-family halo, and its findings sharpen rather than pre-empt the present claim. Panickssery et al. (2024) show that “GPT-4 is 73.5% accurate distinguishing itself from two other LLMs and humans” and report a linear correlation between self-recognition and self-preference after fine-tuning, but the self-preference is read from an explicit rating or pairwise choice, concerns a single model’s own outputs rather than family gradations, and is not tied to any label’s correctness; the authors add that their “experiments can only provide evidence towards the causal hypothesis without fully validating it.” Wataoka et al. (2024) quantify self-preference against human pairwise preferences and, tellingly for this paper, *disconnect* it from authorship: “the factor influencing the LLM evaluators’ judgments is not whether the response is their own but rather the perplexity of the responses.”

Across this line the unit is a judge that issues an elicited verdict on a candidate answer, scored against human gold. The hinge is that this paper’s models are labelers, not judges: they assign a fixed label to a call, no verdict on quality is elicited, and a labeler’s agreement with a silver label is itself the endorsement signal, measured without gold. The self-preference results are complementary—they concern a model preferring its own generations under an elicited rating, whereas the halo here is a family-organised agreement among independent labelers, and one of these works locates the cause in perplexity rather than authorship, leaving the family-indexed, gold-free question open.

## B.5 LEARNING FROM LLM LABELS, AND CONSENSUS OVER NOISY LABELERS

The workflow this paper audits—train or decide on labels produced by models rather than humans—has its own established line, and this supplies both the practical motivation and the one-call mitigation framing. Gilardi et al. (2023) show a single LLM annotator exceeds crowd-workers “by about 25 percentage points on average” at “thirty times cheaper” cost, validating LLM labeling but against human gold throughout. Hinton et al. (2015) transfer a teacher’s temperature-softened outputs to a smaller student, carrying “more than 80% of the improvement,” and West et al. (2022) prompt GPT-3 into a machine-authored knowledge graph, filter it with a learned critic, and lift accuracy from “78.5 to 88.4”; both treat a single source’s outputs as a target to imitate rather than examining agreement across sources. Burns et al. (2024) fine-tune a strong student on a weak supervisor’s labels, recovering “more than 20%” of the performance gap naively and “nearly 80%” with an auxiliary loss, and warn that the student can “imitate” the weak model’s systematic errors and that “none of our methods work consistently in all settings.” Bansal et al. (2025) show compute-matched data from a weaker, cheaper model can beat a stronger one’s, with “relative gains of up to 31.6%,” while treating diversity as within-source solution variety (coverage, diversity, and false-positive rate). These works establish that downstream models can learn from LLM labels and that such labels carry source-specific error, but they do not measure whether the *agreement* among label sources is organised by model family, and their diversity notion is within-source rather than cross-family.

The consensus-modeling thread addresses aggregation of noisy labelers directly, and it is where the paper’s caution bites hardest. Ratner et al. (2016) recover labeling-function accuracies without ground truth, reporting an “average 2.34 point F1” gain, under a model in which functions “label independently, given the true label class”; dependencies must be user-declared. Ratner et al. (2017) model labeling functions as an “independent noisy voter” generative model, achieve “132% average improvements ... over prior heuristic approaches,” and can learn pairwise correlations from co-occurrence statistics—yet Example 3.1 shows correlated functions causing an estimated accuracy of 100%, a catastrophic failure, and the correlations are never indexed by the sources’ model family. The ancestor of this thread, Dawid & Skene (1979), estimates per-observer error rates by EM under an explicit conditional-independence assumption: responses “to successive clinicians ... are independent, given the true response,” with “no patient-by-clinician interaction.” The authors themselves warn that “the restrictive nature of these assumptions should be carefully noted in any application” and point to intervening variables “common to all clinicians” that would break independence.

This is the formal crux. Consensus and weak-supervision aggregation presume that noisy labelers err independently given the truth, or that any dependence is declared or learnable from co-occurrence alone. The concern raised here is that LLM labelers drawn from shared model families violate that premise in a structured, family-indexed way—the same-family halo—so that a consensus built on

the Dawid–Skene assumption can be miscalibrated exactly when the raters are sibling models. The paper does not claim its family-indexed dependence structure is the only such violation, nor that removing it raises downstream accuracy, which our evidence does not test.

## B.6 COST-AWARE LABEL PRODUCTION

Finally, one recent work optimises the economics of producing labels and is complementary to the credibility question asked here. Valdes Gonzalez (2026) frame model selection for fixed-label text classification as a multi-objective trade-off among macro-F1, latency, and cost, and find that “fine-tuned encoder-based models from the BERT family achieve competitive, and often superior, classification performance while operating at one to two orders of magnitude lower cost and latency” (for example, IMDB DistilBERT at \$12.44 versus Claude 4.5 zero-shot at \$1174.95 per one million requests). The paper positions LLMs as upstream “knowledge generators” that supply “weak supervision signals, candidate labels,” but it examines only cost, accuracy, and latency (with governance concerns), and states that objectives such as “probability calibration and abstention logic . . . warrant further investigation.” It never audits the trustworthiness, independence, or family bias of the labels produced. The hinge is one of complementary axes: that work optimises the price of label production, whereas this paper sits upstream of cost and asks whether the labels such pipelines rely on or agree upon can be trusted at all.

## B.7 WHAT IS AND IS NOT NEW

The ingredients this paper uses have precedent, and we name them. That ensembles beat their best member, and that the benefit comes from decorrelated errors rather than member count, is classical (Krogh & Vedelsby, 1994; Breiman, 2001; Dietterich, 2000; Kuncheva & Whitaker, 2003; Bian & Chen, 2019) and has been carried to LLMs (Li et al., 2024; Wang et al., 2023; 2024a; Lau et al., 2024). That different models make correlated errors, and that shared components carry systemic cost, is established (Kim et al., 2025; Kleinberg & Raghavan, 2021; Bommasani et al., 2022; Hedden & Raghavan, 2026). That LLM evaluators exhibit self-preference and other biases is documented (Panickssery et al., 2024; Wataoka et al., 2024; Zheng et al., 2023; Liu et al., 2023; Koo et al., 2023; Wang et al., 2024b). That models can label data, that downstream models learn from those labels, and that consensus over noisy labelers presumes conditional independence is long known (Gilardi et al., 2023; Hinton et al., 2015; West et al., 2022; Burns et al., 2024; Bansal et al., 2025; Ratner et al., 2016; 2017; Dawid & Skene, 1979).

What is new is the combination and the measurement. No prior work measures source-dependent agreement *without gold*, holding the labeler fixed and varying only the silver source; reads a labeler’s agreement as an implicit endorsement rather than eliciting a judge’s verdict; indexes the effect by the labeler–source model-family relation, separating same-family from cross-family pairs; distinguishes family diversity from prompt diversity through an effective-labeler count  $M_{\text{eff}}$  that tracks decorrelation rather than member count; and does so under a cost-preserving budget of one inference call per example. The nearest prior measurement, Kim et al. (2025), quantifies error correlation against reference answers on multiple-choice and judging tasks; this paper detects a family-organised agreement halo in a gold-free silver-labeling workflow. Consistent with the paper’s limitations, we do not claim to separate lineage per se from capability proximity, we do not treat silver–gold divergence as itself evidence of bias, and we do not claim that a de-biased panel raises downstream accuracy—these remain open.

## C ADDITIONAL EXPERIMENTAL DETAIL

**The two roles of a silver label.** The same kind of object plays two roles in the audit. The silver labels under audit are a *labeler’s* own predictions  $L(x)$ , while a silver source’s labels  $\{S(x)\}$  serve only as the answer key those predictions are scored against. For example, Haiku’s label for a call can be the prediction under audit, while Sonnet’s independently produced label for the same call is the silver answer key used only to score Haiku’s agreement.

**What a rendering is.** Of the three labeler axes, the *rendering* is the one least familiar from the labeling literature, so we make it concrete. Every prompt contains a placeholder for the enumerated

list of candidate categories, and the rendering decides how that list is filled in. Under the `full` rendering each entry pairs the category name with its business description — for example, “Name: Payment Arrangement / Extension; Description: customer requests more time to pay or to defer a due date” — so the model is handed the definitions that separate near-twin categories. Under the `none` rendering each entry is the bare name — “Name: Payment Arrangement / Extension” — and the model must rely on the label string alone. Nothing else changes between the two: the instructions and the required output format are held fixed, so rendering isolates the effect of showing versus withholding the definitions. A rendering therefore only bites when a prompt actually consults those definitions, which is why the two name-only prompts (which never refer to a description) take the single `none` setting while the three definition-based prompts take both — the source of the “8 configurations from 5 prompts” count.

**Paired comparability.** Fixing all labelers to the same 924 calls makes any two labelers directly comparable and paired, which the paired statistics of §4 exploit.

**What the analysis deliberately avoids.** Every reported bias quantity is either gold-free (the halo, decorrelation) or explicitly flagged as gold-referenced and hence motivational (the divergence). No claim rests on the raw silver–gold gap.

## D SUPPLEMENTARY RESULTS

**Structured, taxonomy-local disagreement.** Where labelers disagree, the disagreement collapses to a few clusters of semantically adjacent categories (the largest: Customer Prospecting ↔ Inquiries about Services, 293 co-confusions), suggesting a portion of the “error” is overlapping category definitions rather than model failure — a lever orthogonal to panel composition.

**A definition-sharpening lever.** Independently of model choice, because disagreement concentrates in a few semantically adjacent category pairs, sharpening these definitions is a plausible system-level lever that could lift multiple labelers at once — one the practitioners can act on without changing the panel.