

Systematic Failure Analysis of Open-Weight LLM Annotators for Classification Tasks: Taxonomy, Uncertainty Signals, and Deployment Boundaries

Nahid Hossain^{♠*†} Siam Rahman Karip^{♠*} Md Faisal Kabir[♡]

Swakkhar Shatabda[◇] Salekul Islam[♣]

[♠]United International University, Bangladesh [♡]Pennsylvania State University, United States

[◇]BRAC University, Bangladesh [♣]North South University, Bangladesh

Abstract

Large-scale annotated datasets underpin progress in natural language understanding, yet human annotation remains a persistent bottleneck. Large Language Models promise to bridge this gap, but their reliability as annotators is poorly understood. We evaluate 14 open-weight instruction-tuned models and 5 matched base-instruct pairs across 51,189 instances spanning 10 heterogeneous datasets, revealing that annotation failures are structurally predictable rather than random. We identify three recurring failure modes: Safety Hallucination, Intensity Collapse, and Entailment Bias. Safety hallucination can precede alignment rather than being introduced by it: some base models flag up to 100% of benign text as hate speech. Alignment benefits scale inversely with base model capability, yielding a +21.6pp gain for the weakest base model yet a negligible +0.5pp shift for the strongest. KL divergence between base and instruct distributions is $2.0\times$ higher on incorrect predictions, providing a lightweight instance-level signal for alignment-associated errors. Inter-prompt unanimity achieves 72–94% accuracy at 71–90% coverage on clean tasks, yet can actively degrade performance on tasks dominated by shared systematic bias. We translate these findings into a practical deployment framework with pre-screening diagnostics and tiered quality-routing policies, establishing principled boundaries for when LLM annotation can be trusted and when human intervention remains necessary.

1 Introduction

Large-scale annotated datasets underpin progress in natural language understanding, yet acquiring quality annotations remains a persistent bottleneck. Traditional human annotation produces gold-standard labels but faces scalability constraints from costs,

time requirements, and domain expertise needs (Tan et al., 2024). Large Language Models (LLMs) enable democratized annotation at unprecedented scales (Gilardi et al., 2023; Wang et al., 2021), but reliability concerns persist regarding hallucination, instruction-following failures, and unpredictable task-specific performance (Ji et al., 2023). Prior empirical evaluations remain limited in scope, examining isolated tasks, small model sets, or simplistic prompting. While existing work shows promise on sentiment and inference tasks, systematic failure patterns, calibration quality, and practical uncertainty quantification remain underexplored for accessible mid-sized open-weight models. The crucial question for practitioners is not merely “how accurate are LLM annotations” but “when and why do they fail, and can we detect failures without ground truth.”

This paper addresses these limitations through a systematic empirical evaluation of 14 open-weight instruction-tuned models and 5 matched base-instruct pairs across 51,189 instances spanning 10 heterogeneous datasets. Moving beyond descriptive aggregate metrics, we subject base and aligned variants to identical validation formats to isolate behavioral shifts from pre-training artifacts. Our analysis reveals that model failures are highly systematic and predictable, shaped by a structural tension between text-completion habits and preference tuning. We identify that (1) safety hallucination can originate as a pre-training artifact rather than an alignment byproduct, with some base models classifying nearly all benign content as hate-speech before any preference tuning, and alignment amplifying rather than introducing the collapse in 3 of 5 lineages; (2) intensity collapse and entailment bias are architecture-specific rather than universal LLM properties; and (3) alignment benefits scale inversely with foundational capability, showing a large macro gain for the weakest base model but a negligible shift for the strongest, with KL di-

* Equal contribution

† Corresponding author: nahid@uiu.ac.bd

vergence running higher on incorrect predictions across 9 of 10 datasets.

Our primary contributions are:

- We conduct a large-scale evaluation of 14 instruction-tuned models and 5 matched base-instruct pairs across 51,189 instances and 10 heterogeneous datasets, benchmarking mid-sized (7–11B) open-weight LLM annotation, the regime deployable on single-GPU hardware, across diverse task complexities and label cardinalities.
- We introduce a three-mode failure taxonomy (Safety Hallucination, Intensity Collapse, and Entailment Bias) and provide evidence that safety hallucination can originate as a pre-training artifact rather than being introduced solely by alignment. We further observe that alignment benefits tend to scale inversely with base-model capability, and establish base-instruct KL divergence as a lightweight instance-level predictor of alignment-associated errors.
- We propose a practical ground-truth-free deployment framework combining Label Bias Index and Expected Calibration Error pre-screening diagnostics with a tiered annotation-routing policy, reducing human annotation burden across tasks while maintaining quality above the unanimous-agreement accuracy floor.

2 Related Work

The use of LLMs as automated annotators has grown rapidly, prompting a wave of empirical studies that vary widely in scope, task coverage, and methodology (Mohammadi et al., 2025). Tan et al. (2024) categorize LLM annotation into generation, assessment, and utilization; our work falls within the assessment category applied strictly to classification tasks. Comparative studies yield mixed conclusions: Calderon et al. (2025) and Horych et al. (2025) show closed-source models can replace humans across ten languages, while Mohta et al. (2023) find supervised models trained on human labels outperform LLM-labeled counterparts. Gu et al. (2025) show LLMs assist experts more effectively than full automation. More recent work has expanded this evaluation to complex, subjective tasks, such as propaganda detection using multi-model voting setups (Sharma et al., 2026), and highlighted structural prompt-sensitivity and semantic label instability in specific areas like natural language inference (Panahi et al., 2026). Unlike prior single-dataset error analyses (Liu et al.,

2021) (Luo et al., 2022), we treat error patterns as cross-task structural phenomena tied to shared pre-training and alignment mechanisms, yielding a taxonomy of recurring failure modes rather than a per-dataset diagnosis. Complementary to this, Baumann et al. (2025) replicate 37 annotation tasks across 18 models and show that implementation choices such as model selection and prompt phrasing can drive downstream statistical conclusions, with error behaviour differing systematically by model scale. Read against our taxonomy, greater capacity may plausibly mitigate modes that reflect limited fine-grained discrimination, such as Intensity Collapse, while leaving or exacerbating modes we trace partly to pre-training rather than capacity, such as Safety Hallucination. Their finding also motivates our prompt-ensemble design: we hold prompts fixed across all models and report inter-prompt disagreement explicitly rather than selecting a single favourable configuration.

Alignment training improves helpfulness and safety (Ouyang et al., 2022) but introduces systematic biases, with helpfulness optimization exhibiting sycophancy (Perez et al., 2023), and safety training creating overcautious behavior (Röttger et al., 2024). These impact benchmarks but remain unexplored in annotation. We connect alignment and pretraining artifacts to annotation failures, framing Safety Hallucination, Intensity Collapse, and Entailment Bias as predictable consequences of alignment-driven model behavior.

Instruction tuning increases confidence without proportional accuracy gains (Kadavath et al., 2022; Tian et al., 2023). Self-consistency offers an alternative: Wang et al. (2023) show aggregating sampled outputs improves accuracy and correlates with correctness. Recent benchmarks confirm this universal overconfidence across open-weight lineages, prompting the extraction of reasoning-based behavioral features and sample-consistency metrics to map task-dependent calibration gaps (Testoni and Calixto, 2026). Concurrently, optimization frameworks have begun using multi-agent consensus and internal disagreement patterns to navigate epistemic uncertainty and determine model routing pathways (Verma et al., 2026). We build upon these paradigms by demonstrating that a simple, hierarchical multi-prompt vote agreement can serve as a robust, ground-truth-free quality filter for standard classification annotation pipelines.

3 Methodology

3.1 Dataset Selection and Configuration

Datasets were chosen to (i) have verified human gold-standard labels, (ii) be publicly accessible for reproducibility, (iii) span diverse task types and label cardinalities ($|\mathcal{Y}| \in [2, 6]$), and (iv) be widely-adopted benchmarks with established baselines. We evaluate across 10 datasets spanning 8 task categories (Table 2, App. A): Sentiment Classification (SST2, SST5), Grammatical Acceptability (CoLA), Natural Language Inference (SNLI, MNLI), Paraphrase Detection (MRPC), Commonsense Reasoning (SWAG), Toxicity Detection (HateXplain), Question Answering (BoolQ) and Question Classification (TREC). The evaluation corpus comprises 51,189 annotated instances with label cardinality from 2 to 6 classes. We use official test sets when publicly available, validation sets when test labels are unavailable, and 15% random training samples otherwise, maintained consistently across models. The corpus is unbalanced by construction: SWAG contributes 20,006 of the 51,189 instances, or 39.1%. Unless stated otherwise, every cross-dataset mean reported in this paper is an unweighted macro-average over the ten per-dataset values, each weighted equally regardless of instance count, so this imbalance does not drive our aggregates. Any instance-pooled quantity is noted explicitly where it appears and is SWAG-weighted.

3.2 Model Selection

3.2.1 Instruction-Tuned Models

We select 14 instruction-tuned models representing diverse architectural lineages and training methodologies, alongside 5 matched base counterparts to evaluate the specific effects of alignment on annotation behavior. Complete model specifications are provided in Table 3 (App. A). Selection criteria ensure fair comparison and reproducibility: (1) 7–11B parameters for computational feasibility on consumer hardware (RTX 3090 24GB VRAM) while maintaining competitive performance; (2) open-weight models from Hugging Face for transparency and reproducibility; (3) instruction-tuned variants enabling zero-shot task performance; and (4) broad applicability rather than domain specialization.

We prioritize architectural diversity over scale, selecting models from organizations across multiple regions to capture varied instruction-tuning

philosophies. Under FP16, the 7–11B range is the practical single-GPU ceiling; we hold precision fixed across all models (Section 3.5) to avoid confounding accuracy and latency with quantization artifacts. The pool includes current releases (Granite-4.1, Falcon3-10B, Ministral-3-8B-Instruct-2512, Qwen3.5-9B); Gemma-2-9b is retained for its matched base/instruct checkpoints, required by our alignment analysis. We exclude closed-source models because failure-mode analysis requires open-weight transparency, and because API pricing makes runs at this scale prohibitive.

3.2.2 Base-Instruct Pairs

To systematically investigate the specific effects of alignment training on annotation behavior, we supplement our primary evaluation pool with a controlled subset of base models. We evaluate five matched base–instruct pairs spanning dominant open-weight model lineages: Llama-3.1-8B and Llama-3.1-8B-Instruct, Gemma-2-9b and Gemma-2-9b-it, Qwen3.5-9B-Base and Qwen3.5-9B, Mistral-7B-v0.3 and Mistral-7B-Instruct-v0.3, and Granite-4.1-8B-Base and Granite-4.1-8B. Subjecting these matched architectures to identical prompt configurations allows us to isolate native pre-training artifacts from instructional and preference-tuning constraints.

3.3 Instruction-Tuned Annotation Setup

For our instruction tuned models, to mitigate systematic biases from single-prompt configurations and capture the full spectrum of model capabilities, we employ a prompt ensembling strategy with hierarchical aggregation.

3.3.1 Prompt Taxonomy and Ensemble Design

Each prompt follows structure $p = [r, t, l, x]$: role definition, task instruction, label specification, and input text. Prompts were not optimized per model to maintain fairness; all models received identical prompts for each task. Representative prompts for final annotation are provided in Appendices B.1–B.3 (Figures 1–30). Each prompt was selected over two alternative formulations per prompt type based on a preliminary experiment conducted on a held-out subset (5% of the data); representative alternative prompts are provided in Appendices C.1–C.3 (Figures 31–42). We report these alternatives so that the sensitivity of our results to prompt phrasing remains inspectable.

Each model is queried with three strategically

designed prompts that systematically explore different reasoning paradigms:

Prompt 1 (Zero-shot Direct): A minimal, instruction-only prompt that elicits immediate classification without exemplars. This configuration evaluates the model’s capacity for label inference from task description alone: $\hat{y} = P_\theta(y|x, p_{\text{zero}})$.

Prompt 2 (Few-shot Learning): A 2-shot prompting strategy where 2 randomly sampled labeled examples (x_i, y_i) are prepended to the query. To ensure fair cross-model comparison, identical exemplars are used across all evaluated LLMs. This configuration assesses in-context learning capacity: $\hat{y} = P_\theta(y|x, p_{\text{few}}, \{(x_1, y_1), (x_2, y_2)\})$.

Prompt 3 (Chain-of-Thought Reasoning): An explicit reasoning elicitation prompt that instructs the model to perform step-wise analysis before producing the final label. This probes the model’s interpretive depth and structured reasoning capabilities: $\hat{y} = P_\theta(y|x, p_{\text{CoT}})$.

3.3.2 Hierarchical Aggregation with Self-Consistency

To improve prediction reliability and quantify model uncertainty, we employ a 2-stage hierarchical voting strategy built upon self-consistency sampling. This approach leverages both intra-prompt consistency and inter-prompt ensemble diversity to produce robust final predictions.

Stage 1: Intra-Prompt Aggregation: For each prompt variant $p_t \in \{p_{\text{zero}}, p_{\text{few}}, p_{\text{CoT}}\}$ and instance x_i , we generate $K = 3$ independent predictions by sampling from the model’s output distribution: $\{\hat{y}_i^{(t,k)}\}_{k=1}^K \sim P_{\theta_j}(\cdot|x_i, p_t)$. The prompt-specific prediction is determined through majority voting within each prompt type: $\hat{y}_i^{(t)} = \arg \max_{y \in \mathcal{Y}} \sum_{k=1}^K \mathbb{1}[\hat{y}_i^{(t,k)} = y]$, where $\mathbb{1}[\cdot]$ is the indicator function. This intra-prompt aggregation reduces variance induced by sampling stochasticity within each prompting paradigm.

Stage 2: Inter-Prompt Ensemble Aggregation: Given the three prompt-specific predictions $\{\hat{y}_i^{(t)}\}_{t=1}^3$, the final ensemble prediction is obtained through a second majority vote across prompt types: $\hat{y}_i^{\text{final}} = \arg \max_{y \in \mathcal{Y}} \sum_{t=1}^3 \mathbb{1}[\hat{y}_i^{(t)} = y]$. This hierarchical aggregation systematically combines predictions from diverse reasoning modalities, exploiting complementary strengths while mitigating prompt-specific biases. To quantify prediction uncertainty at the ensemble level, we compute the normalized vote entropy across the inter-prompt ag-

gregation: $H(x_i) = -\frac{1}{\log 3} \sum_{y \in \mathcal{Y}} \frac{c_y}{3} \log \frac{c_y}{3}$, where $c_y = \sum_{t=1}^3 \mathbb{1}[\hat{y}_i^{(t)} = y]$ is the vote count for label y across the three prompt types. Here $H(x_i) = 0$ denotes unanimous agreement and $H(x_i) = 1$ maximum disagreement, giving a per-instance measure of prompt sensitivity.

3.3.3 Conditional Log-Probability Extraction

Separate from the free-generation annotation pipeline, we extract conditional log-probabilities to serve as exact calibration and confidence signals. For each evaluation instance, we use teacher forcing with length-normalized log-probability to score the valid label space $\mathcal{Y}$. Specifically, for each candidate label, we append it to the prompt string and compute the sum of token-level log-probabilities at the label positions via a single forward pass. This raw score is divided by the label length (measured in total tokens) to eliminate inherent geometric biases toward shorter label structures. These length-normalized scores are subsequently processed through a softmax normalization across all candidate labels to produce a localized probability distribution over the allowed label space.

3.4 Comparative Base-Instruct Alignment Framework

To isolate the behavioral modifications introduced by alignment paradigms from native pre-training artifacts, we implement a controlled evaluation framework using our matched base-instruct model pairs.

3.4.1 Simple Few-Shot Design

Base models lack instruction-following training and are highly sensitive to prompt variation and unconstrained generation. For stability, we evaluate this subgroup with a uniform simple-completion template. This format structures the context as an ordered sequence of feature-value pairs immediately followed by its ground-truth label token. We construct balanced exemplars to ensure full label coverage, mapping exactly n examples for n distinct classes within the target dataset’s label space. The prompt terminates abruptly at the final target label indicator prefix of the evaluation instance, ensuring a clean continuation environment that matches the text-completion objective of unaligned foundational models. Concrete structural examples of these continuation prompts are detailed in App. C.4. We adopt this two-part design deliberately. Base and instruct variants receive identical inputs, but a single shared prompt cannot

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	HateX	BoolQ	TREC	Mean
Llama	92.2	49.6	57.2	51.7	66.3	76.1	53.8	32.4	77.9	58.0	61.5
Gemma	91.1	54.0	82.6	66.3	79.0	71.1	73.8	37.3	88.1	57.2	70.0
Qwen	85.3	47.5	77.0	79.4	83.5	67.6	76.4	25.8	86.1	46.6	67.5
Mistral	91.5	42.3	81.5	62.7	73.4	76.2	60.8	29.5	82.4	74.2	67.4
DeepSeek	77.8	39.4	53.4	41.6	45.7	59.8	34.5	32.7	69.8	35.0	49.0
EuroLLM	86.1	44.0	77.6	63.4	62.6	72.6	57.1	39.4	81.7	59.6	64.4
EXAONE	83.4	50.5	75.7	77.0	81.8	72.9	71.4	34.0	83.8	56.8	68.7
Ministral	92.6	45.9	78.6	62.3	67.4	64.4	64.8	32.4	84.0	88.0	68.0
Granite	88.9	52.5	82.4	82.7	83.3	68.8	75.0	29.2	85.7	83.8	73.2
Kanana	89.6	46.8	73.2	75.5	62.7	69.2	65.9	29.0	84.0	75.4	67.1
Mengzi	88.9	45.9	73.2	55.9	67.0	78.3	60.0	27.9	81.8	62.8	64.2
GLM-4	91.1	52.6	82.6	79.0	83.5	77.9	66.1	26.4	85.9	82.2	72.7
Falcon	89.0	44.9	79.3	87.0	85.1	72.6	69.2	27.9	85.2	77.8	71.8
SOLAR	82.1	52.5	74.3	74.8	82.3	78.2	75.9	31.9	86.5	90.0	72.8
Mean	87.8	47.7	74.9	68.5	73.1	71.8	64.6	31.1	83.1	67.7	67.0

Table 1: Main evaluation accuracy across evaluated models. All values are produced by the ensembled free-generation pipeline of Section 3.3, with structurally invalid outputs penalised as errors (Section 3.6).

be neutral between them: an instruction-formatted prompt advantages aligned models, while a bare completion template advantages unaligned ones. The cost of this choice is that base and instruct variants are not scored under identical prompt formats, so an unknown portion of the base-instruct difference is confounded with prompt format rather than attributable to alignment. We therefore report all such differences as alignment-associated rather than alignment-caused throughout Section 4.4.

3.4.2 Log-Probability Parity Evaluation

To eliminate formatting failures and label-space deviations, we avoid free-form generation for base models. Both base and instruct models are scored with the length-normalized log-probability teacher-forcing routine of Section 3.3.3, isolating the shifts introduced by alignment under identical inputs and a single forward pass.

3.5 Token Configuration and Modeling Parameters

Input token limits are set at the 99th percentile of tokenizer-specific distributions, as tokenizer design alone shifts effective sequence length across lineages (Hossain and Faisal Kabir, 2026). Generative self-consistency sampling uses $\tau = 0.7$ with extended output limits for Chain-of-Thought; log-probability extraction uses $\tau = 0.0$ for both instruction-tuned and base models. We evaluate all models in FP16 rather than 4- or 8-bit quantization: although quantization fits larger models within the 24GB budget, it perturbs output distributions and calibration, confounding the signals our analysis depends on (KL divergence, ECE, Label Bias In-

dex, log-probabilities). We therefore treat precision as a controlled variable. Full token configuration tables are provided in Tables 4–13 in App. A.

3.6 Response Processing and Evaluation

Generated text responses from the instruction-tuned models are processed through an automated extraction pipeline. We use regular expression matching to capture the first valid label token in the sequence, followed by lowercase normalization to standardize outputs against gold-standard labels. Let $f : \text{String} \rightarrow \mathcal{Y} \cup \{\text{invalid}\}$ denote this extraction and normalization function.

To strictly assess deployment reliability, structural formatting anomalies are not discarded. Self-consistency majority votes that fall outside the permitted label space or fail to match a valid token are treated as incorrect predictions. Formally, standard classification accuracy is computed over all N instances within the dataset: $\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[f(\text{response}_i) = y_i]$. This formulation directly penalizes format failures as annotation errors rather than excluding them from the calculation. For multi-class metrics where structural modeling is required, macro- F_1 scores are computed exclusively over the subset of valid, interpretable outputs where $f(\text{response}_i) \neq \text{invalid}$.

4 Experimental Results and Analysis

4.1 Performance Across Label Space Complexity

Ensembling across 14 instruction-tuned architectures and 10 datasets yields a macro-accuracy mean of 67.0%. Performance declines non-linearly

with label cardinality. Binary benchmarks set the highest baseline but remain fragmented by task difficulty: shallow pattern-matching tasks (SST-2, BoolQ) lead, whereas relational grammar and semantic-matching tasks (CoLA, MRPC) lag. Table 1 reports the accuracy of all instruction-tuned models, while Table 14 (App. A) reports the macro F1 score.

Transitioning from binary to three-way classification spaces induces a sharp performance reduction, with the three-class average falling to 57.6%. Natural language inference tasks manage moderate performance, with MNLI leading at a 73.1% mean accuracy (71.6%) and SNLI following at 68.5% (66.5%). However, domain-specific classification architectures suffer severe drops; the HateXplain toxicity benchmark collapses to a 31.1% mean accuracy across models. This is accompanied by a mean macro- F_1 of 22.6%. The two figures are different metrics rather than two estimates of one quantity: with a near-degenerate prediction distribution, recall on the normal and offensive classes is very low, so macro- F_1 sits below accuracy by construction. Both reflect the difficulty of processing localized context rather than basic semantic connections.

Importantly, our analysis are inconsistent with the assumption that annotation quality scales downward monotonically with label size. Higher-order multi-class datasets with explicit structural framing challenge traditional constraints. 4-class SWAG outperforms 3-class NLI, and 6-class TREC (67.7%) beats 5-class SST-5 (47.7%) by 20.0pp; clear category definitions matter more than label count. This suggests that clear category definitions allow models to easily eliminate impossible options, making task clarity more important than the total number of choices. Pairwise McNemar significance tests across all model pairs are reported in Table 15 (App. A).

Downstream performance varies sharply across distinct open-weight model lineages. Granite-4.1-8b sets the empirical ceiling at 73.2% accuracy (71.7%), followed closely by SOLAR-10.7B (72.8%) and GLM-4-9b (72.7%). Deepseek-llm-7b-chat anchors the fleet floor at 49.0% accuracy (43.4%), revealing a 24.2pp gap below the ceiling. Granite’s leading performance is consistent with recent architectural and data curation optimizations. 95% confidence intervals for all models are reported in Tables 16, 17, 18 and 19 (App. A).

4.2 Failure Mode Taxonomy

Each of the three modes requires a specific label structure in order to surface: binary safety triggers (HateXplain), ordinal intensity (SST-5), and relational inference (SNLI, MNLI). Our remaining datasets lack these structures, so the concentration of evidence for each mode reflects dataset design rather than a gap in coverage. We accordingly frame each mode by its demonstrated scope on the tasks that expose it, rather than as a universal property of large language models, and we note where a mode’s measured rate is contingent on our particular label verbalization.

4.2.1 Safety Hallucination

Evaluating HateXplain reveals extreme vulnerability to safety over-activation, yielding a 95.00% mean false positive rate (FPR) on benign text. Every model exceeds an 85% FPR while maintaining a low 4.82% mean FNR on true hate text (Table 20, App. A). Crucially, base models exhibit near-identical or extensive over-flagging patterns prior to any alignment: Mistral base registers the highest rate at 100.0%, Llama base at 97.9%, and Gemma base at 79.3%, while Granite base sits at the opposite extreme with only 0.3% (Table 21, App. A). Alignment training merely worsens or triggers this vocabulary collapse (e.g., Gemma instruct reaches 99.8% hatespeech classifications, a +20.5pp shift). This challenges the common assumption that over-cautiousness is introduced only by safety alignment, providing evidence that safety hallucination can also originate as a pre-training artifact. This collapse is not an artifact of counting invalid outputs as errors. The mean HateXplain invalid rate across the 14 models is 2.5%, ranging from 0.0% (Ministral, Kanana, Mengzi) to 7.7% (DeepSeek), recoverable from the N_{invalid} columns of Tables 26–39. This cannot account for a 31.1% mean accuracy or a 95.00% mean false positive rate: recomputing over valid outputs only shifts the mean to 32.0%, with no model moving more than 2.8pp. The collapse reflects models confidently emitting valid but incorrect labels, not format failures.

4.2.2 Intensity Collapse

When tasked with annotating fine-grained sentiment categories in the SST-5 dataset, models exhibit a systematic bias, labeled as "Intensity Collapse", where extreme emotional labels are compressed into moderate options. SST-5 contains

279 "Very Negative" and 399 "Very Positive" test instances; we report dampening as a percentage over these fixed denominators so that rates computed over the smaller extreme-label subsets can be judged for reliability. As shown in Figure 53, on "Very Negative" text, EuroLLM dampens intensity in 90% of instances, followed by DeepSeek (85%) and Kanana (84%). Symmetrically, EuroLLM dampens "Very Positive" text in 91% of cases, with Kanana and DeepSeek at 86% and 85%. Intensity collapse is highly model-dependent rather than a universal LLM property: Mistral resist compression for "Very Positive" only and Falcon resist compression for both extreme labels, restricting sentiment dampening to under 15% of samples. No model reaches 100% dampening, confirming this functions as a soft statistical bias rather than a hard constraint. As with all our results, these dampening rates are measured under a fixed five-way label verbalization and may shift under alternative phrasings of the extreme categories.

4.2.3 Entailment Bias

On NLI datasets, models systematically misclassify neutral context relationships as positive entailments. This agreeableness bias is highly architecture-specific, spanning a massive 78pp variance across lineages. DeepSeek and Llama collapse severely (well over half of neutral pairs flagged as entailment), whereas highly optimized architectures such as SOLAR and Qwen stay below 6% (Figure 54, App. D). Because the entailment/neutral/contradiction verbalization can itself influence NLI predictions, these rates reflect the model under our prompt ensemble rather than the model in isolation.

4.3 Uncertainty-Based Quality Filtering

We evaluate whether model consensus across our hierarchical prompt-ensemble structure acts as a reliable, ground-truth-free quality control signal.

Enforcing full inter-prompt unanimity (Agreement Level 3) systematically amplifies accuracy compared to non-unanimous majorities. On binary sentiment (SST-2), absolute unanimity scales classification accuracy to 93.9% while retaining 88.9% data coverage; simplicity majorities (Agreement Level 2) drop sharply to 64.6% accuracy (5.3% coverage). Symmetrically, binary BoolQ reaches 86.3% accuracy under unanimous consensus at 89.9% coverage, but collapses to 56.1% accuracy under partial disagreement. Full self-consistency

error matrices at Agreement Level 3 are in Tables 22 – 24 (App. A). This precision signal remains robust in higher-order multi-class spaces: across TREC, SWAG, MNLI, and SNLI, unanimity sustains 72–82% accuracy at 70–86% coverage, whereas non-unanimous instances fall to 41–52% (Table 25, App. A). Figure 55 (App. D) summarizes the accuracy–coverage tradeoff across all agreement levels and datasets.

However, this precision filtering mechanism completely breaks down on datasets experiencing deep, systematic group biases. On HateXplain, forcing absolute unanimity actually degrades accuracy (34.5% → 31.4%) at 74.7% coverage. This indicates that when models share a collective, structural failure mode, such as safety hallucination, the architectures will confidently and unanimously agree on the identical incorrect token assignment. The relationship between ensemble entropy and prediction accuracy across all tasks is visualized in Figure 56 (App. D). Inter-prompt consensus effectively filters out random textual uncertainty but is entirely incapable of detecting errors born from shared pre-training artifacts. Per-model self-consistency agreement breakdowns are provided in Tables 26 – 39 (App. A)

4.4 Base vs. Instruct Alignment Analysis

4.4.1 Inverse Alignment Benefit

Comparative evaluation of matched base–instruct pairs isolates preference-tuning shifts. Accuracy gains scale inversely with foundational base model capabilities across all five lineages. Complete base vs. instruct accuracy comparisons are in Table 40 (App. A). The weakest base model, Gemma (38.3%), registers the largest alignment benefit (+21.6pp, to 59.9%), driven by large gains on CoLA, TREC, SWAG, and MNLI. Mistral (46.6%) gains +13.0pp, including the single largest task improvement in the study (MRPC), offset only by a small HateXplain regression. Medium-capability Llama (57.1%) gains a modest but inconsistent +6.3pp, with large BoolQ/TREC gains offset by a severe CoLA regression. For the strongest base models, Granite (70.5%) and Qwen (72.1%), benefits flatten almost entirely (+1.0pp and +0.5pp net), as per-task gains and regressions cancel out. Alignment primarily forces formatting template compliance rather than injecting analytical reasoning; practitioners should prioritize raw foundational base strength over alignment density. A heatmap of alignment effects on annotation accu-

racy across all model–dataset combinations is provided in Figure 57. Detailed failure attribution rates are in Table 41 (App. A) and Figure 58 (App. D).

4.4.2 KL Divergence as an Instance-Level Predictor of Alignment Errors

Localized probability space transitions reveal the statistical origin of alignment regressions. Across the 10-dataset matrix, the Kullback-Leibler divergence ratio, $KL(P_{\text{instruct}} \parallel P_{\text{base}})$, is systematically $2.0\times$ higher on incorrect than correct predictions (range $1.3\text{--}4.5\times$ across datasets), with HateXplain the sole exception ($1.0\times$), caused by total label collapse where high divergence occurs globally regardless of accuracy (Figure 59). This provides instance-level evidence consistent with the hypothesis that when alignment overcorrects and heavily distorts the label space away from pre-training beliefs, those specific instances become errors. Label probability shift plots for each model family are provided in Figures 60, 61, and 62 (App. D)

4.4.3 Calibration Updates and the Toxicity Collapse Boundary

Instruction tuning improves Expected Calibration Error (ECE) across 9 out of 10 tasks. On CoLA, Gemma’s ECE falls from 0.395 to 0.118 (-0.277 shift); on MNLI, Llama’s error drops from 0.247 to 0.054 (-0.193 shift); on BoolQ, Qwen’s ECE sharpens from 0.155 to 0.090 (-0.065 shift). However, calibration completely breaks down on HateXplain, inducing extreme overconfidence on error cases. Post-alignment, toxicity ECE doubles for Llama (0.152 to 0.396, $+0.244$ shift) and expands for Gemma (0.441 to 0.620, $+0.179$ worse) and Qwen (0.172 to 0.197). Because aligned models generate near-total false positives with extreme confidence, log-prob metrics are least reliable exactly where practitioners need them most. Full margin and overconfidence gap breakdowns are in Tables 42–44 (App. A). Full ECE results across both base vs instruct and instruct-only setup are in Tables 45–46 (App. A) and Figure 65 (App. D).

4.5 Analytical Trade-offs: SC vs. Logprob Pipelines

Production engineering demands evaluating the latency-accuracy trade-offs between generative Self-Consistency ensembling versus direct conditional log-probability extraction under teacher forcing. In predictive alignment, SC acts as a more robust accuracy indicator, outperforming log-probabilities across 7 out of 10 tasks to secure an

aggregate mean win rate of 57.9% (Table 47, App. A). On CoLA, SC correlates with ground-truth correctness at 0.312 versus 0.180 for log-probabilities, a trend mirrored on SST-2 (0.555 vs. 0.329) and HateXplain (0.052 vs. -0.022). Logprobs only break this pattern on complex multiple-choice reasoning (SWAG) and refined sentiment classification (SST5). However, SC imposes significant computational latency overhead. Running $K = 3$ generations across 3 distinct prompt paradigms requires a 9-pass decoding loop per sample, pushing total fleet execution time to a $1.72\times$ latency expansion over a single teacher-forced forward pass, with the gap widening on context-dense tasks (Table 48 and 49, App. A). SC majorities should be deployed only where absolute accuracy maximization can absorb high compute token cycles, while log-probability pipelines provide a lean, high-speed alternative for real-time indexing workloads. Input-side reduction offers a complementary lever, since linguistically informed context minimization lowers per-instance cost without a proportional accuracy loss (Hossain and Kabir, 2025). Detailed logprob confidence correctness breakdowns by dataset group are in Tables 50 – 53 (App. A).

5 Practical Deployment Guidelines

Our analysis reveals that LLM annotation failures are structured and task-predictable. We distill our findings into guidelines for practitioners designing ground-truth-free pipelines (see Figure 66, App. D).

5.1 Pre-Deployment Task Screening

The most consequential decision in an LLM annotation pipeline is whether to deploy at all. We recommend computing two diagnostics on a small unlabeled probe set (~ 200 instances) via a single logprob forward pass before committing to full annotation.

Label Bias Index (LBI) measures skew in the predicted label distribution. Clean tasks cluster at $LBI \leq 0.14$ (SST-2: 0.005, BoolQ: 0.013, SWAG: 0.038), while structurally compromised tasks cluster at $LBI \geq 0.30$ (HateXplain: 0.34), yielding a natural screening threshold of $LBI > 0.30$. Full LBI results across all models are provided in Table 54 (App. A). Expected Calibration Error (ECE) separates well-calibrated tasks (SST-2: 0.092, BoolQ: 0.122) from catastrophically miscalibrated ones (HateXplain: 0.564). We propose $ECE > 0.45$ as a corroborating criterion.

If $\text{LBI} > 0.30$ or $\text{ECE} > 0.45$, the task should be routed to human annotation entirely. The correct response on a compromised task is *rejection*, not degraded filtering: on HateXplain, enforcing unanimous inter-prompt agreement, the strongest available quality signal, actively degrades accuracy from 34.5% to 31.4% at 74.7% coverage. No confidence filter can correct a failure mode that is shared simultaneously across all models and prompt types. While our most severe failure case is observed on a single toxicity benchmark, the LBI and ECE diagnostics that flag it are task-agnostic and computed without reference to any specific failure mode.

5.2 Quality Filtering on Clean Tasks

For tasks passing screening, our hierarchical self-consistency framework provides a reliable ground-truth-free quality signal, with inter-prompt unanimity (Agreement Level 3) consistently separating high- from low-quality predictions (Table 25, App. A). We recommend a three-tier routing policy keyed to the level of agreement. Instances on which all three prompts agree (Level 3) are auto-accepted, a tier that yields 72–94% accuracy at 71–90% coverage across clean tasks. Instances on which only two of three prompts agree (Level 2) are passed to logprob arbitration: those whose confidence exceeds the task-specific P_{75} threshold are accepted and the remainder escalated. Finally, instances producing a three-way split (Level 1) are sent to human review; this mass is small, comprising only 0.1–4.9% of instances. Together, this structure reduces human annotation burden to 5–15% of instances on binary tasks and 20–30% on multi-class tasks, while maintaining quality above the Level 3 accuracy floor. All accepted instances are emitted as a unified dataset, each tagged with its agreement level and confidence margin for downstream quality stratification.

6 Conclusion

This work delivers a systematic failure analysis of LLM annotators, showing that annotation failures are structurally predictable and cluster into three recurring failure modes tied to pre-training artifacts and alignment dynamics, with safety hallucination preceding alignment rather than being introduced by it. We show that alignment benefits scale inversely with base model capability, that KL divergence serves as an instance-level predictor of alignment-associated errors, and that inter-prompt consensus provides a reliable ground-truth-free

quality signal on well-behaved tasks while collapsing under shared systematic bias. These findings are distilled into a practical deployment framework with pre-screening diagnostics and tiered quality-routing policies, providing practitioners with principled criteria for determining when annotation by mid-sized open-weight models can be trusted and when human intervention remains necessary. We frame these criteria as a diagnosis of the accessible 7–11B regime rather than a claim that necessarily extends to frontier-scale annotators.

Limitations

First, our core signals (KL divergence, Expected Calibration Error, Label Bias Index, and conditional log-probabilities) are sensitive to output-distribution perturbations, so we hold precision fixed at FP16 and avoid quantization; running models beyond 11B unquantized requires multi-GPU resources outside our budget, which bounds the study at 7–11B. Our findings should therefore be read as a characterisation of accessible mid-sized open-weight annotators and may not extend to frontier-scale models. Second, we exclude closed-source models, precluding comparison with the strongest available annotators. Third, our causal attributions rest on five matched base-instruct pairs under a single log-probability parity protocol; while this isolates alignment-associated shifts, five lineages cannot fully separate alignment effects from residual factors. Fourth, penalizing structurally invalid outputs to measure deployment reliability makes format-failure rates an upper bound constrained decoding can largely suppress. Fifth, we cover only fixed-label-space classification tasks; this enables cross-task comparison, but findings may not generalize to open-ended generative annotation. Sixth, the corpus is unbalanced by construction: SWAG contributes 20,006 of 51,189 instances (39.1%); cross-dataset means are unweighted macro-averages over ten datasets, while instance-pooled quantities are SWAG-weighted and flagged as such. Seventh, base and instruct variants are scored under the simple-completion template of Section 3.4.1, not those of Section 3.3, so the base-instruct difference is confounded with prompt format to an unknown extent; we call these effects alignment-associated, not alignment-caused. Eighth, dampening and entailment rates (Sections 4.2.2, 4.2.3) use a single label verbalization and may shift under alternative phrasings of extreme sentiment and inference categories.

References

- Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. [Large language model hacking: Quantifying the hidden risks of using llms for text annotation](#). *Preprint*, arXiv:2509.08825.
- Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. [The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Ma"el Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. [Large language models are effective human annotation assistants, but not good independent annotators](#). *Preprint*, arXiv:2503.06778.
- Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. [The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nahid Hossain and Md. Faisal Kabir. 2026. [Efficient monolingual pretraining in low-resource settings through morphology-aware tokenization, principled corpus denoising, and benchmark-driven evaluation](#). *IEEE Access*, 14:91979–92003.
- Nahid Hossain and Md Faisal Kabir. 2025. [Context minimization for resource-constrained text classification: Optimizing performance-efficiency trade-offs through linguistic features](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15412–15426, Suzhou, China. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Jia, and 1 others. 2022. [Language models \(mostly\) know what they know](#). *arXiv preprint arXiv:2207.05221*.
- Zhe Liu, Yufan Guo, and Jalal Mahmud. 2021. [When and why a model fails? a human-in-the-loop error detection framework for sentiment analysis](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 170–177, Online. Association for Computational Linguistics.
- Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. [Simple but challenging: Natural language inference models fail on simple sentences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3449–3462, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro, Massimo Poesio, Ayoub Bagheri, and Anastasia Giachanou. 2025. [Assessing the reliability of LLMs annotations in the context of demographic bias and model explanation](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 92–104, Vienna, Austria. Association for Computational Linguistics.
- Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. [Are large language models good annotators?](#) In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 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.
- Solmaz Panahi, John Kelleher, and Vasudevan Nandumpozhimana. 2026. [When LLMs annotate: Reliability challenges in low-resource NLI](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 178–188, Rabat, Morocco. Association for Computational Linguistics.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [XSTest: A test suite for identifying exaggerated](#)

safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Vivek Sharma, Shweta Jain, Mohammad Shokri, Sarah Ita Levitan, and Elena Filatova. 2026. **Council of LLMs: Evaluating capability of large language models to annotate propaganda.** In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. **Large language models for data annotation and synthesis: A survey.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Alberto Testoni and Iacer Calixto. 2026. **Mind the gap: Benchmarking LLM uncertainty and calibration with specialty-aware clinical QA and reasoning-based behavioural features.** In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2364–2382, Rabat, Morocco. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. **Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Akshay Verma, Swapnil Gupta, Deepak Gupta, Praateek Sircar, and Siddharth Pillai. 2026. **SELENE: Selective and evidence-weighted LLM debating for efficient and reliable reasoning.** In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 95–104, Rabat, Morocco. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. **Want to reduce labeling cost? GPT-3 can help.** In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. **Self-consistency improves chain of thought reasoning in language models.** In *International Conference on Learning Representations*.

A App.: Dataset, Model, and Result Tables

Task	Dataset	Labels	Samples
Sentiment Analysis	SST-2	2	872
	SST-5	5	2,210
Grammatical Acceptability	CoLA	2	1,043
Natural Language Inference	SNLI	3	9,824
	MNLI	3	9,815
Paraphrase Detection	MRPC	2	1,725
Commonsense Reasoning	SWAG	4	20,006
Toxicity Detection	HateXplain	3	1,924
Question Answering	BoolQ	2	3,270
Question Classification	TREC	6	500

Table 2: Overview of evaluation datasets spanning eight task categories across multiple domains.

Model	Organization	Parameters
Llama-3.1-8B	Meta	8B
Gemma-2-9b	Google	9B
Qwen3.5-9B-Base	Alibaba	9B
granite-4.1-8b-base	IBM	8B
Mistral-7B-v0.3	Mistral AI	7B
Llama-3.1-8B-Instruct	Meta	8B
Gemma-2-9b-it	Google	9B
Qwen3.5-9B	Alibaba	9B
Mistral-7B-Instruct-v0.3	Mistral AI	7B
deepseek-llm-7b-chat	DeepSeek	7B
EXAONE-3.5-7.8B-Instruct	LG AI Research	7.8B
Ministral-3-8B-Instruct-2512	Mistral AI	8B
granite-4.1-8b	IBM	8B
kanana-1.5-8b-instruct-2505	Kakao Corp	8B
Mengzi3-8B-Chat	Langboat	8B
EuroLLM-9B-Instruct	Utter Project	9B
glm-4-9b-chat-hf	Zhipu AI	9B
Falcon3-10B-Instruct	TII UAE	10B
SOLAR-10.7B-Instruct-v1.0	Upstage	10.7B

Table 3: Overview of evaluated models

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	65.57	10.18	90	99.57	10.18	124	103.57	10.18	128
Gemma	67.03	9.94	91	102.03	9.94	126	108.03	9.94	132
Qwen	66.55	10.17	91	100.55	10.17	125	104.55	10.17	129
Mistral	72.94	11.37	100	112.94	11.37	140	116.94	11.37	144
DeepSeek	68.81	10.39	94	106.81	10.39	132	109.81	10.39	135
EuroLLM	72.35	11.04	98	112.35	11.04	138	115.35	11.04	141
Mengzi	65.57	10.18	90	109.57	10.18	134	103.57	10.18	128
EXAONE	69.90	10.44	95	106.90	10.44	132	111.90	10.44	137
Granite	65.58	10.18	90	99.58	10.18	124	103.58	10.18	128
GLM-4	65.57	10.17	90	99.57	10.17	124	103.57	10.17	128
Falcon	71.44	10.90	98	108.44	10.90	135	113.44	10.90	140
SOLAR	72.94	11.37	100	112.94	11.37	140	116.94	11.37	144
Ministral	67.18	10.52	93	113.18	10.52	139	108.18	10.52	134
Kanana	65.57	10.18	90	109.57	10.18	134	103.57	10.18	128

Table 4: Token Length Distribution: **SST2**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	72.49	10.76	101	185.49	10.76	214	121.49	10.76	150
Gemma	74.01	10.55	101	192.01	10.55	219	126.01	10.55	153
Qwen	74.46	10.77	102	193.46	10.77	221	123.46	10.77	151
Mistral	78.77	11.94	110	209.77	11.94	241	134.77	11.94	166
DeepSeek	75.73	10.99	104	199.73	10.99	228	127.73	10.99	156
EuroLLM	79.23	11.61	110	216.23	11.61	247	135.23	11.61	166
Mengzi	72.49	10.76	101	195.49	10.76	224	121.49	10.76	150
EXAONE	76.82	11.09	106	206.82	11.09	236	129.82	11.09	159
Granite	72.50	10.77	101	185.50	10.77	214	121.50	10.77	150
GLM-4	72.49	10.76	101	186.49	10.76	215	121.49	10.76	150
Falcon	78.39	11.53	108	210.39	11.53	240	131.39	11.53	161
SOLAR	78.77	11.94	110	209.77	11.94	241	134.77	11.94	166
Ministral	73.09	11.06	102	197.09	11.06	226	125.09	11.06	154
Kanana	72.49	10.76	101	195.49	10.76	224	121.49	10.76	150

Table 5: Token Length Distribution: **SST5**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	57.57	4.56	74	90.57	4.56	107	100.57	4.56	117
Gemma	58.54	4.78	76	94.54	4.78	112	101.54	4.78	119
Qwen	59.55	4.60	76	94.55	4.60	111	102.55	4.60	119
Mistral	69.26	5.00	88	111.26	5.00	130	119.26	5.00	138
DeepSeek	58.69	4.83	77	96.69	4.83	115	101.69	4.83	120
EuroLLM	69.13	4.94	87	115.13	4.94	133	117.13	4.94	135
Mengzi	57.57	4.56	74	100.57	4.56	117	100.57	4.56	117
EXAONE	61.70	4.80	80	101.70	4.80	120	105.70	4.80	124
Granite	57.57	4.56	74	90.57	4.56	107	100.57	4.56	117
GLM-4	57.56	4.55	74	90.56	4.55	107	100.56	4.55	117
Falcon	59.53	4.77	77	97.53	4.77	115	103.53	4.77	121
SOLAR	69.26	5.00	88	111.26	5.00	130	119.26	5.00	138
Ministral	58.67	4.6	75	103.67	4.6	120	101.67	4.6	118
Kanana	57.57	4.56	74	100.57	4.56	117	100.57	4.56	117

Table 6: Token Length Distribution: **CoLA**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	78.62	8.02	103	180.62	8.02	205	123.62	8.02	148
Gemma	78.02	7.97	102	175.02	7.97	199	125.02	7.97	149
Qwen	81.35	8.04	105	189.35	8.04	213	126.35	8.04	150
Mistral	90.93	8.69	117	217.93	8.69	244	140.93	8.69	167
DeepSeek	81.42	8.09	106	188.42	8.09	213	128.42	8.09	153
EuroLLM	87.97	8.70	115	213.97	8.70	241	139.97	8.70	167
Mengzi	78.62	8.02	103	190.62	8.02	215	123.62	8.02	148
EXAONE	82.50	8.17	107	192.50	8.17	217	130.50	8.17	155
Granite	78.62	8.02	103	180.62	8.02	205	123.62	8.02	148
GLM-4	78.62	8.02	103	180.62	8.02	205	123.62	8.02	148
Falcon	81.97	7.97	106	187.97	7.97	212	129.97	7.97	154
SOLAR	90.93	8.69	117	217.93	8.69	244	140.93	8.69	167
Ministral	79.04	8.19	104	191.04	8.19	216	126.04	8.19	151
Kanana	78.62	8.02	103	190.62	8.02	215	123.62	8.02	148

Table 7: Token Length Distribution: **SNLI**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	90.08	18.87	147	236.08	18.87	293	135.08	18.87	192
Gemma	90.08	19.24	147	230.08	19.24	287	137.08	19.24	194
Qwen	92.94	19.23	149	242.94	19.23	299	137.94	19.23	194
Mistral	104.41	21.12	168	275.41	21.12	339	154.41	21.12	218
DeepSeek	93.84	19.65	153	244.84	19.65	304	140.84	19.65	200
EuroLLM	100.47	20.42	161	267.47	20.42	328	152.47	20.42	213
Mengzi	90.08	18.87	147	246.08	18.87	303	135.08	18.87	192
EXAONE	94.89	19.68	153	248.89	19.68	307	142.89	19.68	201
Granite	90.09	18.88	147	236.09	18.88	293	135.09	18.88	192
GLM-4	90.11	18.89	147	236.11	18.89	293	135.11	18.89	192
Falcon	94.18	19.33	152	243.18	19.33	301	142.18	19.33	200
SOLAR	104.41	21.12	168	275.41	21.12	339	154.41	21.12	218
Ministral	90.97	19.41	148	248.97	19.41	306	137.97	19.41	195
Kanana	90.08	18.87	147	246.08	18.87	303	135.08	18.87	192

Table 8: Token Length Distribution: **MNLI**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	109.61	13.37	139	244.61	13.37	274	150.61	13.37	180
Gemma	111.84	15.00	149	255.84	15.00	293	154.84	15.00	192
Qwen	114.35	14.92	151	261.35	14.92	298	155.35	14.92	192
Mistral	133.12	16.41	173	296.12	16.41	336	180.12	16.41	220
DeepSeek	114.19	15.38	151	261.19	15.38	298	157.19	15.38	194
EuroLLM	130.43	15.82	168	293.43	15.82	331	176.43	15.82	214
Mengzi	109.61	13.37	139	254.61	13.37	284	150.61	13.37	180
EXAONE	120.00	15.43	158	278.00	15.43	316	164.00	15.43	202
Granite	109.62	13.37	139	244.62	13.37	274	150.62	13.37	180
GLM-4	109.74	13.48	139	245.74	13.48	275	150.74	13.48	180
Falcon	119.56	16.66	161	275.56	16.66	317	163.56	16.66	205
SOLAR	133.12	16.41	173	296.12	16.41	336	180.12	16.41	220
Ministral	115.22	15.14	152	272.22	15.14	309	158.22	15.14	195
Kanana	109.61	13.37	139	254.61	13.37	284	150.61	13.37	180

Table 9: Token Length Distribution: **MRPC**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	118.90	15.52	165	406.90	15.52	453	165.90	15.52	212
Gemma	122.93	15.32	168	425.93	15.32	471	170.93	15.32	216
Qwen	123.86	15.50	169	427.86	15.50	473	170.86	15.50	216
Mistral	135.09	16.45	184	465.09	16.45	514	189.09	16.45	238
DeepSeek	124.98	15.51	171	433.98	15.51	480	172.98	15.51	219
EuroLLM	133.61	16.36	181	470.61	16.36	518	185.61	16.36	233
Mengzi	118.9	15.52	165	416.9	15.52	463	165.9	15.52	212
EXAONE	125.00	15.54	171	433.00	15.54	479	174.00	15.54	220
Granite	118.90	15.52	165	406.90	15.52	453	165.90	15.52	212
GLM-4	118.89	15.52	165	406.89	15.52	453	165.89	15.52	212
Falcon	124.83	15.35	170	431.83	15.35	477	173.83	15.35	219
SOLAR	135.09	16.45	184	465.09	16.45	514	189.09	16.45	238
Ministral	120.93	15.79	167	423.93	15.79	470	169.93	15.79	216
Kanana	118.9	15.52	165	416.9	15.52	463	165.9	15.52	212

Table 10: Token Length Distribution: **SWAG**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	73.50	15.82	109	201.50	15.82	237	115.50	15.82	151
Gemma	73.26	15.38	109	195.26	15.38	231	117.26	15.38	153
Qwen	74.39	15.72	111	201.39	15.72	238	116.39	15.72	153
Mistral	81.90	17.24	123	228.90	17.24	270	128.90	17.24	170
DeepSeek	75.98	16.01	113	209.98	16.01	247	119.98	16.01	157
EuroLLM	82.39	16.78	120	221.39	16.78	259	128.39	16.78	166
Mengzi	73.5	15.82	109	211.5	15.82	247	115.5	15.82	151
EXAONE	76.80	16.03	114	213.80	16.03	251	121.80	16.03	159
Granite	73.52	15.83	109	201.52	15.83	237	115.52	15.83	151
GLM-4	73.48	15.80	109	201.48	15.80	237	115.48	15.80	151
Falcon	77.25	15.67	115	202.25	15.67	240	122.25	15.67	160
SOLAR	81.90	17.24	123	228.90	17.24	270	128.90	17.24	170
Ministral	77.38	16.22	114	218.38	16.22	255	122.38	16.22	159
Kanana	73.5	15.82	109	211.5	15.82	247	115.5	15.82	151

Table 11: Token Length Distribution: **HateXplain**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	176.95	71.32	413	562.95	71.32	799	224.95	71.32	461
Gemma	179.51	73.07	418	545.51	73.07	784	228.51	73.07	467
Qwen	182.95	73.97	422	568.95	73.97	808	230.95	73.97	470
Mistral	199.79	81.44	469	640.79	81.44	910	250.79	81.44	520
DeepSeek	186.05	75.41	426	586.05	75.41	826	235.05	75.41	475
EuroLLM	195.26	78.28	449	605.26	78.28	859	248.26	78.28	502
Mengzi	176.95	71.32	413	572.95	71.32	809	224.95	71.32	461
EXAONE	185.73	75.34	429	593.73	75.34	837	235.73	75.34	479
Granite	177.08	71.42	413	564.08	71.42	800	225.08	71.42	461
GLM-4	177.20	71.52	413	561.20	71.52	797	225.20	71.52	461
Falcon	184.46	74.29	422	571.46	74.29	809	234.46	74.29	472
SOLAR	199.79	81.44	469	640.79	81.44	910	250.79	81.44	520
Ministral	182.35	74.95	428	584.35	74.95	830	232.35	74.94	478
Kanana	176.95	71.32	413	572.95	71.32	809	224.95	71.32	461

Table 12: Token Length Distribution: **BoolQ**

Tokenizer	Zero-Shot Direct			Few-Shot ICL			Chain-of-Thought		
	Mean	Std	P_{99}	Mean	Std	P_{99}	Mean	Std	P_{99}
Llama	61.21	2.88	70	159.21	2.88	168	105.21	2.88	114
Gemma	63.05	3.15	72	166.05	3.15	175	109.05	3.15	118
Qwen	63.24	2.99	72	167.24	2.99	176	107.24	2.99	116
Mistral	70.08	3.15	79	193.08	3.15	202	117.08	3.15	126
DeepSeek	66.32	3.09	75	175.32	3.09	184	112.32	3.09	121
EuroLLM	71.52	3.23	81	201.52	3.23	211	119.52	3.23	129
Mengzi	61.21	2.88	70	169.21	2.88	178	105.21	2.88	114
EXAONE	67.22	3.15	76	180.22	3.15	189	114.22	3.15	123
Granite	61.21	2.88	70	159.21	2.88	168	105.21	2.88	114
GLM-4	61.21	2.88	70	159.21	2.88	168	105.21	2.88	114
Falcon	67.16	3.32	79	181.16	3.32	193	114.16	3.32	126
SOLAR	70.08	3.15	79	193.08	3.15	202	117.08	3.15	126
Ministral	63.28	3.06	72	174.28	3.06	183	109.28	3.06	118
Kanana	61.21	2.88	70	169.21	2.88	178	105.21	2.88	114

Table 13: Token Length Distribution: **TREC**

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	HateX	BoolQ	TREC	Mean
Llama	92.7	46.0	57.3	48.4	64.5	71.5	53.6	23.3	78.2	51.6	58.7
Gemma	93.7	51.1	82.0	66.0	80.1	72.3	74.8	32.2	88.6	55.7	69.6
Qwen	94.2	43.4	79.8	84.6	85.2	70.1	78.5	22.2	86.8	78.6	72.3
Mistral	93.4	38.4	79.0	61.9	72.2	69.4	61.0	22.9	82.8	64.5	64.5
DeepSeek	82.7	26.8	53.7	35.3	40.4	39.8	31.4	24.5	69.0	30.2	43.4
EuroLLM	90.7	27.1	77.0	56.7	55.6	70.4	57.4	19.5	81.9	43.2	58.0
EXAONE	89.3	44.5	75.6	76.5	81.6	70.3	71.5	25.1	83.6	88.6	70.7
Ministral	93.3	40.6	78.0	59.6	65.5	64.6	64.8	22.0	84.2	87.7	66.0
Granite	92.4	46.5	82.4	80.4	82.5	67.5	74.9	22.0	85.9	82.2	71.7
Kanana	92.3	40.8	73.0	74.5	59.3	68.0	65.5	18.2	84.3	74.2	65.0
Mengzi	92.1	39.1	73.2	49.3	65.1	75.1	59.5	16.4	81.7	55.6	60.7
GLM-4	92.9	48.7	82.4	78.9	83.4	75.8	65.8	21.0	86.0	80.8	71.6
Falcon	92.7	38.6	79.0	86.9	84.7	70.9	69.0	21.3	85.3	74.9	70.3
SOLAR	91.5	46.8	73.9	72.5	81.7	76.3	75.6	25.7	86.8	84.4	71.5
Mean	91.7	41.3	74.7	66.5	71.6	68.7	64.5	22.6	83.2	68.0	65.1

Table 14: Main Evaluation F_1 -Score Results across Evaluated Models. The Mean column is an unweighted macro-average over the ten datasets, not an instance-weighted pooled score.

	Deep	EXAO	Euro	Falc	GLM-	Gemm	Gran	Kana	Llam	Meng	Mini	Mist	Qwen	SOLA
DeepSeek		9	10	10	10	10	10	10	8	10	9	9	10	9
EXAONE	9		7	8	9	8	9	8	7	9	7	10	8	5
EuroLLM	10	7		7	10	7	10	9	9	6	7	7	8	9
Falcon	10	8	7		7	8	7	5	10	7	7	6	6	10
GLM-4	10	9	10	7		7	4	8	8	8	9	7	6	7
Gemma	10	8	7	8	7		6	8	8	9	9	8	10	9
Granite	10	9	10	7	4	6		7	10	8	10	8	7	8
Kanana	10	8	9	5	8	8	7		10	6	8	7	8	8
Llama	8	7	9	10	8	8	10	10		9	7	8	9	9
Mengzi	10	9	6	7	8	9	8	6	9		8	7	9	8
Ministral	9	7	7	7	9	9	10	8	7	8		7	8	8
Mistral	9	10	7	6	7	8	8	7	8	7	7		10	10
Qwen	10	8	8	6	6	10	7	8	9	9	8	10		7
SOLAR	9	5	9	10	7	9	8	8	9	8	8	10	7	

Table 15: Pairwise McNemar Significance Matrix: Count of Datasets (out of 10) with Statistically Significant Performance Differences (Mirrored Matrix)

Model	SST2	SST5	CoLA	SNLI	MNLI
Llama	90.4–93.9	47.6–51.5	54.6–59.9	51.0–52.5	65.5–67.1
Gemma	89.1–92.9	52.1–55.8	80.2–84.8	65.5–67.1	78.2–79.8
Qwen	82.9–87.6	45.6–49.3	74.5–79.5	78.6–80.2	82.7–84.2
Mistral	89.6–93.4	40.5–44.1	79.1–83.8	61.9–63.6	72.6–74.2
DeepSeek	75.1–80.4	37.8–40.9	50.4–56.5	41.0–42.3	45.0–46.4
EuroLLM	83.8–88.4	42.8–45.2	75.1–80.1	62.8–64.0	62.0–63.2
EXAONE	81.0–85.8	48.7–52.4	73.4–78.0	76.2–77.8	81.1–82.5
Ministral	90.8–94.3	43.9–47.9	76.4–80.7	61.4–63.1	66.5–68.2
Granite	86.7–90.9	50.5–54.4	80.1–84.6	82.0–83.4	82.6–84.0
Kanana	87.6–91.5	45.2–48.5	70.6–75.7	74.8–76.2	62.1–63.4
Mengzi	86.7–90.8	44.1–47.8	70.5–75.8	55.1–56.7	66.2–67.8
GLM-4	89.2–92.9	50.7–54.6	80.4–84.8	78.3–79.8	82.7–84.2
Falcon	86.9–91.1	43.1–46.7	76.8–81.7	86.3–87.6	84.4–85.8
SOLAR	79.6–84.5	50.4–54.6	72.2–76.4	74.0–75.6	81.6–83.1

Table 16: Model 95% Confidence Intervals (Accuracy %) — Part A

Model	MRPC	SWAG	HateXplain	BoolQ	TREC
Llama	74.3–77.9	53.2–54.5	30.9–34.0	76.4–79.3	54.4–61.8
Gemma	69.0–73.2	73.1–74.4	35.4–39.1	87.0–89.2	54.0–60.4
Qwen	65.4–69.9	75.8–77.0	24.3–27.3	85.0–87.3	43.2–50.0
Mistral	74.6–77.9	60.1–61.4	27.8–31.2	81.1–83.7	71.4–76.8
DeepSeek	58.6–60.9	33.9–35.1	31.1–34.2	68.3–71.4	32.0–38.2
EuroLLM	70.5–74.6	56.4–57.8	38.8–39.9	80.4–83.1	55.6–63.8
EXAONE	70.7–75.0	70.8–72.0	32.3–35.9	82.5–85.1	53.0–60.6
Ministral	62.3–66.4	64.1–65.5	31.1–33.7	82.8–85.3	85.4–90.4
Granite	66.8–70.7	74.4–75.6	27.3–31.1	84.5–86.9	80.8–86.8
Kanana	67.1–71.2	65.3–66.6	27.3–30.7	82.8–85.3	72.4–78.4
Mengzi	76.5–80.1	59.3–60.7	27.2–28.6	80.4–83.1	59.2–66.4
GLM-4	75.9–79.8	65.5–66.8	24.6–28.2	84.7–87.1	79.4–84.8
Falcon	70.5–74.6	68.6–69.8	26.2–29.6	84.0–86.4	74.6–81.0
SOLAR	76.4–80.0	75.3–76.5	30.0–33.7	85.4–87.7	87.4–92.4

Table 17: Model 95% Confidence Intervals (Accuracy %) — Part B

Model	SST2	SST5	CoLA	SNLI	MNLI
Llama	90.9–94.5	43.9–48.0	54.6–60.0	47.5–49.3	63.6–65.4
Gemma	92.1–95.3	48.9–53.3	79.4–84.6	65.0–66.8	79.3–80.8
Qwen	92.5–95.7	41.2–45.6	76.9–82.5	83.8–85.3	84.5–85.9
Mistral	91.7–95.0	36.6–40.1	76.4–81.7	61.0–62.8	71.3–73.0
DeepSeek	80.1–85.2	25.5–28.0	50.5–56.7	34.4–36.2	39.4–41.3
EuroLLM	88.8–92.6	25.7–28.5	74.4–79.6	56.0–57.5	54.9–56.4
EXAONE	93.0–96.1	46.0–50.6	74.2–80.4	85.1–86.5	82.0–83.5
Ministral	90.8–94.3	44.4–48.3	68.7–75.0	61.8–63.7	66.4–68.2
Granite	93.3–96.3	48.9–53.2	77.8–83.0	87.6–88.9	86.0–87.4
Kanana	87.5–91.5	34.1–38.0	70.1–74.9	72.4–74.1	54.8–56.4
Mengzi	90.0–93.7	40.6–44.7	68.9–74.2	52.1–54.0	64.0–65.7
GLM-4	89.4–93.1	47.4–51.7	76.4–82.1	78.6–80.2	82.8–84.2
Falcon	93.3–96.3	43.0–47.0	77.2–82.5	89.6–90.8	85.5–86.9
SOLAR	93.6–96.7	52.1–56.3	63.5–70.2	76.0–77.6	82.9–84.3

Table 18: Model 95% Bootstrap Confidence Intervals (F1-Score %) — Part A

Model	MRPC	SWAG	HateXplain	BoolQ	TREC
Llama	69.2–73.8	53.0–54.3	21.9–24.8	76.8–79.6	47.9–55.2
Gemma	70.2–74.3	74.1–75.3	30.7–33.7	87.5–89.7	51.7–59.8
Qwen	68.0–72.2	77.9–79.1	20.7–23.7	85.6–87.9	72.9–84.3
Mistral	67.0–71.9	60.3–61.7	21.4–24.4	81.5–84.1	57.5–69.8
DeepSeek	37.7–40.8	30.8–32.0	22.8–26.3	67.3–70.6	26.5–33.9
EuroLLM	68.1–72.7	56.7–58.1	19.1–20.0	80.6–83.2	57.5–66.0
EXAONE	71.9–76.2	71.7–73.0	29.2–32.4	83.4–85.8	83.6–92.5
Ministral	62.2–66.4	64.1–65.4	18.2–20.5	82.9–85.3	79.4–88.4
Granite	72.0–77.0	74.7–75.9	23.9–26.9	85.0–87.4	84.8–91.0
Kanana	66.8–70.8	65.2–66.5	22.4–25.4	82.5–85.1	63.9–71.9
Mengzi	71.7–76.2	59.2–60.6	16.0–17.8	80.4–83.0	56.3–66.7
GLM-4	73.5–77.8	65.3–66.6	21.6–24.8	84.9–87.3	71.8–80.9
Falcon	69.9–74.2	69.3–70.6	20.7–23.6	83.8–86.3	70.7–78.3
SOLAR	73.1–77.6	75.8–77.0	26.2–29.2	86.2–88.5	89.0–95.2

Table 19: Model 95% Bootstrap Confidence Intervals (F1-Score %) — Part B

Model	FPR (%)	FNR (%)	Safety
			Collapse (%)
Llama	96.97	6.39	16.94
Gemma	94.78	0.91	55.51
Qwen	89.56	7.12	66.74
Mistral	96.97	7.85	23.54
DeepSeek	85.35	8.76	11.69
EuroLLM	96.63	0.18	0.88
EXAONE	88.55	1.82	57.59
Ministral	98.15	3.47	6.13
Granite	97.31	4.38	51.25
Kanana	99.66	2.37	65.96
Mengzi	98.48	2.74	91.27
GLM-4	97.31	5.84	59.04
Falcon	93.77	6.39	24.90
SOLAR	95.96	8.21	40.18
Mean	95.00	4.82	40.90

Table 20: Safety Hallucination on HateXplain averaged across prompt types. FPR = False Positive Rate on normal content; FNR = False Negative Rate on hate content; Safety Collapse = rate of offensive content misclassified as hate.

Model	Type	Hatespeech (%)	Offensive (%)	Normal (%)
Llama	base	97.9	0.0	2.1
	instruct	91.3	0.1	8.6
Gemma	base	79.3	4.5	16.2
	instruct	99.8	0.1	0.1
Mistral	base	100.0	0.0	0.0
	instruct	83.8	4.0	12.2
Qwen	base	28.6	63.6	7.8
	instruct	41.9	38.9	19.2
Granite	base	0.3	78.3	21.4
	instruct	7.1	71.1	21.8

Table 21: Distribution of safety classification categories across base and instruction-tuned variants on HateXplain. Values are marginal distributions of predicted labels over the full test set and are therefore not directly comparable with the class-conditional rates of Table 20: a high conditional false positive rate on the gold-normal subset is arithmetically compatible with a moderate marginal share of normal predictions, because the marginal also collects normal predictions made on offensive and hate instances.

Model	Dataset	True Positive	Consistent Error	Inconsistent Correct	True Negative	Consistent Error Rate (%)
DeepSeek	BoolQ	1734	540	550	446	16.51
	CoLA	401	289	156	197	27.71
	HateXplain	317	562	312	733	29.21
	MNLI	3747	3828	736	1504	39.00
	MRPC	831	331	200	363	19.19
	SNLI	3607	4371	485	1361	44.49
	SST2	626	115	52	79	13.19
	SST5	593	781	277	559	35.34
	SWAG	4003	5902	2893	7208	29.50
	TREC	113	113	62	212	22.60
EXAONE	BoolQ	2650	429	91	100	13.12
	CoLA	763	146	27	107	14.00
	HateXplain	488	902	167	367	46.88
	MNLI	7670	1247	357	541	12.71
	MRPC	1198	333	59	135	19.30
	SNLI	7120	891	444	1369	9.07
	SST2	707	36	20	109	4.13
	SST5	1014	804	103	289	36.38
	SWAG	13274	4150	1004	1578	20.74
	TREC	206	12	78	204	2.40
EuroLLM	BoolQ	2353	352	320	245	10.76
	CoLA	674	124	135	110	11.89
	HateXplain	616	929	141	238	48.28
	MNLI	5115	1912	1033	1755	19.48
	MRPC	980	249	272	224	14.43
	SNLI	4999	1645	1227	1953	16.74
	SST2	589	28	162	93	3.21
	SST5	736	796	237	441	36.02
	SWAG	7754	3235	3665	5352	16.17
	TREC	167	31	131	171	6.20
Falcon	BoolQ	2767	424	20	59	12.97
	CoLA	807	163	20	53	15.63
	HateXplain	477	1162	60	225	60.40
	MNLI	8167	1129	188	331	11.50
	MRPC	1222	393	30	80	22.78
	SNLI	8428	833	114	449	8.48
	SST2	771	39	5	57	4.47
	SST5	815	889	177	329	40.23
	SWAG	13444	5308	403	851	26.53
	TREC	375	74	14	37	14.80
GLM-4	BoolQ	2576	294	234	166	8.99
	CoLA	803	127	59	54	12.18
	HateXplain	331	1011	177	405	52.55
	MNLI	7638	1093	554	530	11.14
	MRPC	1093	195	251	186	11.30
	SNLI	6960	1400	806	658	14.25
	SST2	766	56	28	22	6.42
	SST5	808	624	355	423	28.24
	SWAG	10151	3023	3079	3753	15.11
	TREC	372	42	39	47	8.40

Table 22: Self-Consistency Error Matrix (Threshold = 3): Evaluation Across Model Families and Datasets

Model	Dataset	True Positive	Consistent Error	Inconsistent Correct	True Negative	Consistent Error Rate (%)
Gemma	BoolQ	2845	327	37	61	10.00
	CoLA	844	135	17	47	12.94
	HateXplain	592	955	126	251	49.64
	MNLI	7469	1633	289	424	16.64
	MRPC	1191	417	36	81	24.17
	SNLI	6205	2772	306	541	28.22
	SST2	771	44	23	34	5.05
	SST5	1081	820	112	197	37.10
	SWAG	14109	4134	646	1117	20.66
	TREC	274	100	12	114	20.00
Granite	BoolQ	2772	383	30	85	11.71
	CoLA	827	155	32	29	14.86
	HateXplain	465	1198	97	164	62.27
	MNLI	7288	908	892	727	9.25
	MRPC	1007	211	179	328	12.23
	SNLI	7238	631	888	1067	6.42
	SST2	760	33	15	64	3.78
	SST5	1078	902	82	148	40.81
	SWAG	14510	4124	500	872	20.61
	TREC	388	42	31	39	8.40
Kanana	BoolQ	2622	402	126	120	12.29
	CoLA	733	249	30	31	23.87
	HateXplain	314	900	244	466	46.78
	MNLI	5704	2871	455	785	29.25
	MRPC	1028	407	165	125	23.59
	SNLI	6505	1606	914	799	16.35
	SST2	764	81	17	10	9.29
	SST5	818	850	217	325	38.46
	SWAG	11675	4359	1512	2460	21.79
	TREC	318	68	59	55	13.60
Llama	BoolQ	2029	379	517	345	11.59
	CoLA	378	254	219	192	24.35
	HateXplain	333	634	291	666	32.95
	MNLI	5395	2184	1109	1127	22.25
	MRPC	1085	263	227	150	15.25
	SNLI	4294	3707	790	1033	37.73
	SST2	727	27	77	41	3.10
	SST5	598	515	499	598	23.30
	SWAG	5221	1679	5550	7556	8.39
	TREC	178	80	112	130	16.00
Mengzi	BoolQ	2588	503	86	93	15.38
	CoLA	751	261	12	19	25.02
	HateXplain	497	1294	40	93	67.26
	MNLI	6192	2783	385	455	28.35
	MRPC	1279	305	72	69	17.68
	SNLI	5153	3942	341	388	40.13
	SST2	731	51	44	46	5.85
	SST5	824	920	191	275	41.63
	SWAG	10348	5344	1657	2657	26.71
	TREC	281	112	33	74	22.40
Ministral	BoolQ	2590	375	157	148	11.47
	CoLA	662	101	158	122	9.68
	HateXplain	487	969	137	331	50.36
	MNLI	5356	1950	1257	1252	19.87
	MRPC	815	374	295	241	21.68
	SNLI	4850	2508	1266	1200	25.53
	SST2	766	41	41	24	4.70
	SST5	671	664	344	531	30.05
	SWAG	10667	3725	2296	3318	18.62
	TREC	403	34	37	26	6.80

Table 23: Self-Consistency Error Matrix (Threshold = 3) Part 2

Model	Dataset	True Positive	Consistent Error	Inconsistent Correct	True Negative	Consistent Error Rate (%)
Mistral	BoolQ	2653	511	42	64	15.63
	CoLA	847	168	3	25	16.11
	HateXplain	499	1177	68	180	61.17
	MNLI	7000	2323	204	288	23.67
	MRPC	1306	376	9	34	21.80
	SNLI	5881	3263	280	400	33.21
	SST2	791	52	7	22	5.96
	SST5	864	1110	70	166	50.23
	SWAG	11067	5977	1086	1876	29.88
	TREC	362	112	9	17	22.40
Qwen	BoolQ	2612	278	204	176	8.50
	CoLA	697	98	106	142	9.40
	HateXplain	302	908	194	520	47.19
	MNLI	7291	752	901	871	7.66
	MRPC	949	330	218	228	19.13
	SNLI	7163	965	641	1055	9.82
	SST2	715	31	29	97	3.56
	SST5	780	700	269	461	31.67
	SWAG	12889	1777	2403	2937	8.88
	TREC	183	17	50	250	3.40
SOLAR	BoolQ	2799	364	31	76	11.13
	CoLA	694	141	81	127	13.52
	HateXplain	588	1220	25	91	63.41
	MNLI	7971	1527	110	207	15.56
	MRPC	1337	338	12	38	19.59
	SNLI	7213	2188	137	286	22.27
	SST2	706	33	10	123	3.78
	SST5	1105	929	56	120	42.04
	SWAG	14956	4296	226	528	21.47
	TREC	448	31	2	19	6.20

Table 24: Self-Consistency Error Matrix (Threshold = 3): Part 3

Dataset	Agreement Level	Accuracy (%)	Coverage (%)
BoolQ	1	77.5	0.1
	2	56.1	9.0
	3	86.3	89.9
CoLA	1	65.1	0.5
	2	51.1	12.9
	3	79.7	84.2
HateXplain	1	32.6	1.7
	2	34.5	21.2
	3	31.4	74.7
MNLI	1	43.2	0.9
	2	50.7	11.8
	3	77.4	86.0
MRPC	1	54.2	1.6
	2	53.9	13.6
	3	77.1	82.2
SNLI	1	45.4	1.0
	2	52.1	11.6
	3	73.6	84.6
SST2	1	78.4	1.0
	2	64.6	5.3
	3	93.9	88.9
SST5	1	34.8	1.6
	2	42.5	21.4
	3	51.0	74.6
SWAG	1	31.4	3.5
	2	41.6	20.5
	3	72.2	75.4
TREC	1	47.0	4.9
	2	50.7	15.6
	3	81.6	70.5

Table 25: Self-Consistency Performance: Accuracy and Coverage Breakdown by Agreement Level

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	97.5	845	22	5	94.4	27.3	0.6717
SST5	79.0	1740	462	8	53.4	36.2	0.1730
CoLA	85.6	891	150	2	59.4	45.3	0.1404
SNLI	93.8	9167	603	54	51.8	56.0	-0.0428
MNLI	92.8	9026	704	85	68.2	49.2	0.1908
MRPC	93.5	1607	112	6	78.5	45.5	0.3293
SWAG	69.6	13906	6071	29	67.2	23.4	0.4383
HateXplain	80.0	1531	382	11	34.4	25.4	0.0903
BoolQ	90.0	2914	323	33	82.3	45.5	0.3682
TREC	80.7	402	96	2	67.7	18.8	0.4891

Table 26: Self-Consistency and Logprob Consistency Metrics for Llama

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	99.3	841	6	25	94.0	50.0	0.4405
SST5	69.6	1497	653	60	59.8	45.5	0.1437
CoLA	94.5	958	56	29	85.7	71.4	0.1427
SNLI	88.7	8454	1076	294	69.3	60.9	0.0840
MNLI	87.2	8442	1238	135	83.5	57.5	0.2595
MRPC	82.1	1381	302	42	70.4	84.4	-0.1405
SWAG	97.5	19251	486	269	75.6	39.3	0.3635
HateXplain	72.2	1338	514	72	34.2	50.8	-0.1662
BoolQ	99.4	3217	18	35	89.2	61.1	0.2813
TREC	95.1	388	20	92	72.4	25.0	0.4742

Table 27: Self-Consistency and Logprob Consistency Metrics for Gemma

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	98.0	774	16	82	95.5	31.2	0.6423
SST5	10.7	223	1861	126	60.5	49.1	0.1142
CoLA	80.5	778	189	76	86.6	68.2	0.1838
SNLI	33.2	3065	6171	588	93.0	80.3	0.1276
MNLI	34.0	3269	6356	190	93.4	80.8	0.1263
MRPC	72.4	1193	455	77	65.4	85.0	-0.1967
SWAG	90.8	17692	1788	526	83.2	32.2	0.5103
HateXplain	71.1	1284	523	117	29.3	22.9	0.0634
BoolQ	96.4	3116	115	39	88.4	54.8	0.3357
TREC	74.0	199	70	231	89.4	78.6	0.1088

Table 28: Self-Consistency and Logprob Consistency Metrics for Qwen

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	99.9	853	1	18	93.6	0.0	0.9355
SST5	93.8	2050	135	25	43.0	39.3	0.0372
CoLA	97.4	1007	27	9	83.3	40.7	0.4258
SNLI	97.2	9450	276	98	63.9	44.9	0.1896
MNLI	97.8	9503	215	97	74.4	63.3	0.1112
MRPC	99.6	1701	7	17	77.2	28.6	0.4862
SWAG	95.1	18980	973	53	62.2	35.4	0.2686
HateXplain	94.4	1798	107	19	29.9	28.0	0.0183
BoolQ	99.4	3215	19	36	83.4	63.2	0.2029
TREC	96.8	480	16	4	76.0	37.5	0.3854

Table 29: Self-Consistency and Logprob Consistency Metrics for Mistral

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	97.3	794	22	56	83.6	63.6	0.1999
SST5	19.1	400	1697	113	44.5	40.8	0.0372
CoLA	85.5	850	144	49	55.2	61.1	-0.0593
SNLI	87.3	8266	1200	358	47.2	15.7	0.3156
MNLI	79.0	7577	2016	222	54.2	18.6	0.3555
MRPC	95.1	1484	77	164	68.1	27.3	0.4079
SWAG	80.4	15870	3881	255	37.4	24.8	0.1254
HateXplain	61.1	1085	691	148	34.5	36.9	-0.0243
BoolQ	91.8	2949	262	59	73.5	44.7	0.2883
TREC	68.5	329	151	20	40.7	27.2	0.1358

Table 30: Self-Consistency and Logprob Consistency Metrics for DeepSeek

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	94.0	778	50	44	94.2	36.0	0.5822
SST5	2.8	60	2100	50	36.7	45.3	-0.0862
CoLA	38.0	388	633	22	63.4	88.9	-0.2554
SNLI	35.7	3435	6198	191	83.7	54.1	0.2959
MNLI	34.7	3357	6324	134	82.8	53.3	0.2951
MRPC	33.0	560	1135	30	61.3	80.1	-0.1884
SWAG	82.2	16311	3543	152	64.6	25.1	0.3947
HateXplain	0.9	17	1865	42	11.8	40.5	-0.2872
BoolQ	91.1	2947	287	36	86.5	43.6	0.4291
TREC	22.2	101	355	44	77.2	62.0	0.1526

Table 31: Self-Consistency and Logprob Consistency Metrics for EuroLLM

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	99.1	762	7	103	95.1	28.6	0.6657
SST5	28.6	594	1479	137	62.1	50.6	0.1155
CoLA	96.3	919	35	89	84.2	45.7	0.3851
SNLI	43.5	3815	4961	1048	82.6	89.0	-0.0641
MNLI	42.1	4040	5557	218	80.1	86.2	-0.0612
MRPC	39.5	651	998	76	62.2	85.4	-0.2316
SWAG	94.4	18621	1112	273	74.3	39.8	0.3446
HateXplain	62.6	1114	666	144	31.1	46.4	-0.1534
BoolQ	98.6	3182	46	42	85.6	34.8	0.5086
TREC	71.3	224	90	186	92.4	85.6	0.0686

Table 32: Self-Consistency and Logprob Consistency Metrics for EXAONE

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	97.6	851	21	0	93.1	71.4	0.2164
SST5	39.0	862	1348	0	50.5	43.0	0.0744
CoLA	67.0	699	344	0	87.8	59.9	0.2796
SNLI	18.6	1822	8002	0	84.0	57.3	0.2673
MNLI	19.3	1890	7925	0	84.5	63.3	0.2120
MRPC	64.2	1107	618	0	49.2	91.4	-0.4219
SWAG	87.9	17578	2428	0	68.9	34.8	0.3418
HateXplain	6.1	118	1806	0	15.2	33.6	-0.1830
BoolQ	95.7	3112	140	18	84.9	75.7	0.0915
TREC	90.4	452	48	0	91.4	56.2	0.3512

Table 33: Self-Consistency and Logprob Consistency Metrics for Ministral

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	99.1	810	7	55	95.1	71.4	0.2363
SST5	41.1	886	1268	56	55.6	52.6	0.0304
CoLA	95.8	987	43	13	84.8	51.2	0.3364
SNLI	89.1	8188	1006	630	90.3	72.6	0.1776
MNLI	86.4	8115	1273	427	88.9	76.0	0.1283
MRPC	88.9	1312	164	249	82.7	61.6	0.2111
SWAG	94.1	18769	1173	64	77.3	43.0	0.3423
HateXplain	79.5	1500	387	37	26.2	43.7	-0.1747
BoolQ	99.2	3206	27	37	87.1	40.7	0.4631
TREC	81.2	388	90	22	88.9	82.2	0.0670

Table 34: Self-Consistency and Logprob Consistency Metrics for Granite

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	98.7	861	11	0	89.6	90.9	-0.0136
SST5	5.4	119	2091	0	47.9	46.8	0.0113
CoLA	80.9	844	199	0	68.5	93.0	-0.2448
SNLI	66.0	6485	3339	0	82.6	61.7	0.2094
MNLI	55.2	5413	4402	0	78.9	42.9	0.3605
MRPC	92.4	1594	131	0	71.0	47.3	0.2363
SWAG	89.2	17850	2156	0	69.0	40.2	0.2881
HateXplain	66.0	1269	655	0	26.1	34.7	-0.0857
BoolQ	91.9	2992	263	15	85.6	70.7	0.1491
TREC	71.7	358	141	1	78.2	68.8	0.0942

Table 35: Self-Consistency and Logprob Consistency Metrics for Kanana

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	98.9	834	9	29	92.2	66.7	0.2554
SST5	31.1	688	1522	0	42.2	47.6	-0.0548
CoLA	98.2	1024	19	0	72.8	94.7	-0.2198
SNLI	69.2	6794	3030	0	57.8	51.8	0.0594
MNLI	77.5	7609	2206	0	70.8	54.0	0.1680
MRPC	85.0	1467	258	0	81.4	60.8	0.2054
SWAG	91.3	18259	1747	0	62.6	33.3	0.2931
HateXplain	91.9	1769	155	0	29.1	14.2	0.1492
BoolQ	98.1	3195	61	14	82.6	59.0	0.2355
TREC	79.2	384	101	15	71.1	40.6	0.3050

Table 36: Self-Consistency and Logprob Consistency Metrics for Mengzi

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	94.1	819	51	2	92.6	70.6	0.2196
SST5	27.7	611	1595	4	55.6	51.6	0.0405
CoLA	83.4	865	172	6	88.9	54.1	0.3483
SNLI	37.0	3620	6171	33	87.2	74.7	0.1243
MNLI	43.0	4200	5575	40	86.2	82.0	0.0427
MRPC	32.8	562	1150	13	69.2	83.0	-0.1383
SWAG	80.5	16090	3900	16	76.5	23.6	0.5285
HateXplain	63.2	1213	707	4	24.4	30.0	-0.0558
BoolQ	91.7	2977	269	24	87.8	73.2	0.1454
TREC	76.4	382	118	0	88.0	63.6	0.2440

Table 37: Self-Consistency and Logprob Consistency Metrics for GLM-4

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	99.9	817	1	54	94.9	100.0	-0.0514
SST5	66.1	1396	716	98	42.3	56.2	-0.1388
CoLA	95.9	963	41	39	83.0	68.3	0.1468
SNLI	80.4	7613	1859	352	92.5	80.8	0.1161
MNLI	72.9	7047	2621	147	87.8	82.6	0.0524
MRPC	39.9	674	1015	36	59.4	83.9	-0.2459
SWAG	97.2	19220	559	227	71.0	37.0	0.3394
HateXplain	26.6	496	1369	59	19.2	32.3	-0.1313
BoolQ	99.5	3219	15	36	86.3	60.0	0.2630
TREC	80.8	384	91	25	86.5	62.6	0.2382

Table 38: Self-Consistency and Logprob Consistency Metrics for Falcon

Dataset	Agreement (%)	N_{agree}	N_{disagree}	N_{invalid}	Acc (Agree)	Acc (Diag)	Acc Gap
SST2	100.0	752	0	120	95.2	—	—
SST5	79.7	1719	439	52	54.3	51.9	0.0234
CoLA	31.3	318	697	28	86.2	71.9	0.1428
SNLI	74.4	7170	2466	188	85.2	50.2	0.3499
MNLI	81.0	7858	1839	118	90.0	55.0	0.3493
MRPC	98.0	1661	34	30	80.2	47.1	0.3319
SWAG	97.4	19353	527	126	77.6	33.0	0.4453
HateXplain	56.7	1074	820	30	19.4	49.4	-0.3002
BoolQ	99.0	3189	32	49	88.4	37.5	0.5087
TREC	26.9	130	354	16	95.4	92.1	0.0329

Table 39: Self-Consistency and Logprob Consistency Metrics for SOLAR

Model	Type	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	HateXplain	BoolQ	TREC	Mean
Gemma	base	51.6	26.9	34.2	36.0	35.7	63.2	24.8	28.6	59.4	22.2	38.3
	instruct	59.3	44.4	69.5	61.9	63.1	73.7	56.2	28.5	87.5	55.2	59.9
Llama	base	88.0	51.5	79.7	37.7	62.0	73.7	48.6	28.3	56.8	44.6	57.1
	instruct	93.8	50.9	61.5	50.7	64.6	75.5	63.6	27.6	82.4	63.2	63.4
Mistral	base	93.2	43.8	31.1	35.6	58.5	37.0	37.0	28.5	57.4	42.8	46.6
	instruct	93.2	53.4	62.5	58.9	65.5	75.8	50.4	25.0	59.5	51.8	59.6
Qwen	base	94.6	53.4	79.9	85.9	86.3	75.0	68.1	31.3	72.4	74.2	72.1
	instruct	95.0	50.9	71.1	87.6	87.1	59.0	79.7	23.0	86.7	85.8	72.6
Granite	base	94.8	50.6	82.7	87.4	84.5	64.4	75.1	26.3	67.5	72.0	70.5
	instruct	94.7	57.4	83.5	79.6	88.3	73.2	76.7	22.4	71.6	67.6	71.5

Table 40: Accuracy Comparison of Base vs. Instruct Models Across Datasets

Model Group	Dataset	<i>N</i>	Both Correct %	Alignment Induced [95% CI]	Alignment Corrected [95% CI]	Net Effect
Llama	SST2	872	86.81	1.2 [0.00, 0.02]	7.0 [0.05, 0.09]	5.8
	SST5	2210	36.65	14.9 [0.13, 0.16]	14.2 [0.13, 0.16]	-0.7
	CoLA	1043	49.09	30.6 [0.28, 0.33]	12.4 [0.10, 0.14]	-18.2
	SNLI	9824	37.44	0.2 [0.00, 0.00]	13.2 [0.13, 0.14]	13.0
	MNLI	9815	60.59	1.4 [0.01, 0.02]	4.0 [0.04, 0.04]	2.6
	MRPC	1725	69.51	4.2 [0.03, 0.05]	6.0 [0.05, 0.07]	1.8
	SWAG	20006	37.71	10.9 [0.10, 0.11]	25.8 [0.25, 0.26]	14.9
	HateXplain	1924	27.44	0.9 [0.00, 0.01]	0.1 [0.00, 0.00]	-0.8
	BoolQ	3270	48.62	8.1 [0.07, 0.09]	33.8 [0.32, 0.35]	25.7
	TREC	500	39.40	5.2 [0.03, 0.07]	23.8 [0.20, 0.28]	18.6
Gemma	SST2	872	11.58	40.0 [0.37, 0.43]	47.7 [0.44, 0.51]	7.7
	SST5	2210	18.51	8.4 [0.07, 0.10]	25.9 [0.24, 0.28]	17.5
	CoLA	1043	5.85	28.4 [0.26, 0.31]	63.7 [0.61, 0.67]	35.3
	SNLI	9824	23.47	12.5 [0.12, 0.13]	38.4 [0.37, 0.39]	25.9
	MNLI	9815	22.78	12.9 [0.12, 0.14]	40.4 [0.39, 0.41]	27.5
	MRPC	1725	60.06	3.2 [0.02, 0.04]	13.6 [0.12, 0.15]	10.4
	SWAG	20006	13.09	11.7 [0.11, 0.12]	43.1 [0.42, 0.44]	31.4
	HateXplain	1924	22.92	5.7 [0.05, 0.07]	5.6 [0.05, 0.07]	-0.1
	BoolQ	3270	53.49	5.9 [0.05, 0.07]	34.0 [0.32, 0.36]	28.1
	TREC	500	3.60	18.6 [0.15, 0.22]	51.6 [0.47, 0.56]	33.0
Mistral	SST2	872	91.74	1.5 [0.01, 0.02]	1.5 [0.01, 0.02]	0.0
	SST5	2210	30.72	13.0 [0.12, 0.14]	22.7 [0.21, 0.24]	9.7
	CoLA	1043	29.63	1.4 [0.01, 0.02]	32.9 [0.30, 0.36]	31.5
	SNLI	9824	35.40	0.1 [0.00, 0.00]	23.5 [0.23, 0.24]	23.3
	MNLI	9815	57.77	0.8 [0.01, 0.01]	7.8 [0.07, 0.08]	7.0
	MRPC	1725	30.49	6.6 [0.05, 0.08]	45.3 [0.43, 0.48]	38.8
	SWAG	20006	33.15	3.8 [0.03, 0.05]	17.2 [0.16, 0.19]	13.4
	HateXplain	1924	23.80	4.7 [0.04, 0.06]	1.2 [0.01, 0.02]	-3.5
	BoolQ	3270	53.24	4.2 [0.04, 0.05]	6.3 [0.05, 0.07]	2.1
	TREC	500	41.20	1.6 [0.01, 0.03]	10.6 [0.08, 0.13]	9.0
Qwen	SST2	872	93.12	1.5 [0.01, 0.02]	1.8 [0.01, 0.03]	0.3
	SST5	2210	39.50	13.8 [0.12, 0.15]	11.4 [0.10, 0.13]	-2.4
	CoLA	1043	66.44	13.4 [0.11, 0.16]	4.7 [0.03, 0.06]	-8.7
	SNLI	9824	84.43	1.4 [0.01, 0.02]	3.2 [0.03, 0.04]	1.8
	MNLI	9815	83.58	2.7 [0.02, 0.03]	3.5 [0.03, 0.04]	0.8
	MRPC	1725	55.48	19.5 [0.18, 0.21]	3.5 [0.03, 0.04]	-16.0
	SWAG	20006	64.22	3.9 [0.04, 0.04]	15.5 [0.15, 0.16]	11.6
	HateXplain	1924	16.32	15.0 [0.13, 0.17]	6.7 [0.06, 0.08]	-8.3
	BoolQ	3270	65.84	6.5 [0.06, 0.07]	20.9 [0.19, 0.22]	14.4
	TREC	500	73.40	0.8 [0.00, 0.02]	12.4 [0.10, 0.15]	11.6
Granite	SST2	872	93.81	1.0 [0.00, 0.02]	0.9 [0.00, 0.02]	-0.1
	SST5	2210	41.54	9.1 [0.08, 0.10]	15.9 [0.14, 0.17]	6.8
	CoLA	1043	75.46	7.2 [0.06, 0.09]	8.1 [0.06, 0.10]	0.9
	SNLI	9824	76.70	10.7 [0.10, 0.11]	2.9 [0.03, 0.03]	-7.7
	MNLI	9815	80.35	4.1 [0.04, 0.05]	7.9 [0.07, 0.08]	3.8
	MRPC	1725	58.49	5.9 [0.05, 0.07]	14.7 [0.13, 0.16]	8.8
	SWAG	20006	68.15	7.0 [0.06, 0.08]	8.6 [0.07, 0.10]	1.6
	HateXplain	1924	17.88	8.4 [0.07, 0.10]	4.5 [0.04, 0.05]	-3.9
	BoolQ	3270	62.91	4.6 [0.04, 0.05]	8.7 [0.08, 0.10]	4.1
	TREC	500	60.20	11.8 [0.09, 0.15]	7.4 [0.05, 0.10]	-4.4

Table 41: Failure Attribution and Alignment Effects Across Model Families and Datasets

Model Group	Dataset	Type	Mean Margin	Margin Correct	Margin Incorrect	Overconf. Gap	<i>N</i>
Llama	SST2	base	0.744	0.786	0.433	-0.353	872
		instruct	0.883	0.903	0.582	-0.321	872
	SST5	base	0.171	0.199	0.140	-0.059	2210
		instruct	0.284	0.329	0.238	-0.091	2210
	CoLA	base	0.284	0.310	0.180	-0.131	1043
		instruct	0.360	0.384	0.320	-0.064	1043
	SNLI	base	0.737	0.847	0.670	-0.177	9824
		instruct	0.721	0.760	0.681	-0.078	9824
	MNLI	base	0.373	0.422	0.293	-0.129	9815
		instruct	0.595	0.670	0.457	-0.214	9815
	MRPC	base	0.208	0.228	0.150	-0.078	1725
		instruct	0.330	0.361	0.235	-0.126	1725
	SWAG	base	0.160	0.214	0.109	-0.105	20006
		instruct	0.568	0.658	0.409	-0.249	20006
	HateXplain	base	0.394	0.392	0.395	0.003	1924
		instruct	0.664	0.675	0.660	-0.014	1924
	BoolQ	base	0.496	0.535	0.444	-0.091	3270
		instruct	0.753	0.797	0.546	-0.251	3270
	TREC	base	0.390	0.621	0.205	-0.417	500
		instruct	0.380	0.488	0.193	-0.295	500
Gemma	SST2	base	0.811	0.825	0.796	-0.029	872
		instruct	0.610	0.733	0.432	-0.301	872
	SST5	base	0.242	0.257	0.237	-0.021	2210
		instruct	0.260	0.302	0.226	-0.077	2210
	CoLA	base	0.678	0.669	0.683	0.014	1043
		instruct	0.578	0.598	0.535	-0.063	1043
	SNLI	base	0.724	0.737	0.716	-0.021	9824
		instruct	0.808	0.866	0.713	-0.153	9824
	MNLI	base	0.795	0.796	0.794	-0.002	9815
		instruct	0.500	0.576	0.372	-0.204	9815
	MRPC	base	0.433	0.444	0.415	-0.029	1725
		instruct	0.470	0.504	0.373	-0.131	1725
	SWAG	base	0.420	0.420	0.420	0.000	20006
		instruct	0.499	0.585	0.389	-0.196	20006
	HateXplain	base	0.669	0.654	0.675	0.021	1924
		instruct	0.904	0.912	0.901	-0.011	1924
	BoolQ	base	0.709	0.731	0.676	-0.055	3270
		instruct	0.880	0.910	0.667	-0.244	3270
	TREC	base	0.570	0.596	0.563	-0.034	500
		instruct	0.782	0.821	0.733	-0.088	500

Table 42: Margin Comparison and Overconfidence Gaps (Part 1: Llama, Gemma)

Model Group	Dataset	Type	Mean Margin	Margin Correct	Margin Incorrect	Overconf. Gap	<i>N</i>
Mistral	SST2	base	0.693	0.713	0.421	-0.292	872
		instruct	0.928	0.943	0.718	-0.225	872
	SST5	base	0.116	0.139	0.098	-0.040	2210
		instruct	0.258	0.288	0.224	-0.065	2210
	CoLA	base	0.315	0.456	0.251	-0.205	1043
		instruct	0.429	0.456	0.384	-0.072	1043
	SNLI	base	0.475	0.617	0.397	-0.220	9824
		instruct	0.695	0.717	0.662	-0.055	9824
	MNLI	base	0.195	0.223	0.156	-0.067	9815
		instruct	0.639	0.740	0.448	-0.292	9815
	MRPC	base	0.048	0.064	0.039	-0.025	1725
		instruct	0.230	0.251	0.165	-0.085	1725
	SWAG	base	0.144	0.185	0.121	-0.064	20006
		instruct	0.617	0.719	0.513	-0.206	20006
	HateXplain	base	0.487	0.464	0.496	0.032	1924
		instruct	0.429	0.387	0.443	0.056	1924
	BoolQ	base	0.576	0.600	0.543	-0.057	3270
		instruct	0.603	0.635	0.555	-0.080	3270
	TREC	base	0.268	0.421	0.154	-0.267	500
		instruct	0.416	0.531	0.292	-0.239	500
Qwen	SST2	base	0.908	0.925	0.597	-0.328	872
		instruct	0.819	0.840	0.440	-0.400	872
	SST5	base	0.161	0.191	0.127	-0.063	2210
		instruct	0.270	0.293	0.246	-0.047	2210
	CoLA	base	0.429	0.461	0.302	-0.159	1043
		instruct	0.494	0.525	0.419	-0.107	1043
	SNLI	base	0.678	0.723	0.404	-0.319	9824
		instruct	0.784	0.824	0.501	-0.323	9824
	MNLI	base	0.560	0.605	0.272	-0.334	9815
		instruct	0.632	0.676	0.335	-0.341	9815
	MRPC	base	0.378	0.410	0.283	-0.128	1725
		instruct	0.431	0.471	0.374	-0.097	1725
	SWAG	base	0.428	0.529	0.213	-0.316	20006
		instruct	0.641	0.719	0.337	-0.382	20006
	HateXplain	base	0.170	0.176	0.167	-0.009	1924
		instruct	0.226	0.172	0.242	0.070	1924
	BoolQ	base	0.724	0.764	0.619	-0.144	3270
		instruct	0.797	0.833	0.563	-0.270	3270
	TREC	base	0.607	0.700	0.338	-0.362	500
		instruct	0.750	0.803	0.428	-0.375	500

Table 43: Margin Comparison and Overconfidence Gaps (Part 2: Mistral, Qwen)

Model Group	Dataset	Type	Mean Margin	Margin Correct	Margin Incorrect	Overconf. Gap	<i>N</i>
Granite	SST2	base	0.563	0.578	0.297	-0.281	872
		instruct	0.978	0.987	0.812	-0.175	872
	SST5	base	0.244	0.264	0.224	-0.039	2210
		instruct	0.404	0.439	0.357	-0.082	2210
	CoLA	base	0.641	0.695	0.383	-0.311	1043
		instruct	0.847	0.887	0.645	-0.242	1043
	SNLI	base	0.728	0.766	0.462	-0.304	9824
		instruct	0.904	0.929	0.805	-0.123	9824
	MNLI	base	0.601	0.646	0.356	-0.290	9815
		instruct	0.898	0.927	0.683	-0.244	9815
	MRPC	base	0.464	0.514	0.374	-0.140	1725
		instruct	0.729	0.762	0.640	-0.122	1725
	SWAG	base	0.550	0.629	0.309	-0.320	20006
		instruct	0.827	0.888	0.630	-0.258	20006
	HateXplain	base	0.370	0.333	0.383	0.050	1924
		instruct	0.496	0.469	0.504	0.035	1924
	BoolQ	base	0.550	0.560	0.531	-0.029	3270
		instruct	0.781	0.818	0.688	-0.130	3270
	TREC	base	0.482	0.577	0.236	-0.342	500
		instruct	0.809	0.875	0.672	-0.203	500

Table 44: Margin Comparison and Overconfidence Gaps (Part 3: Granite)

Model Family	Type	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	Hate Xplain	BoolQ	TREC	Mean
Gemma	base	0.334	0.158	0.395	0.415	0.485	0.214	0.250	0.441	0.242	0.371	0.330
	instruct	0.111	0.185	0.118	0.241	0.135	0.267	0.087	0.620	0.063	0.285	0.211
Llama	base	0.136	0.345	0.513	0.399	0.247	0.530	0.327	0.152	0.221	0.101	0.297
	instruct	0.055	0.245	0.283	0.331	0.054	0.425	0.096	0.396	0.116	0.253	0.225
Qwen	base	0.039	0.373	0.372	0.181	0.303	0.371	0.253	0.172	0.155	0.135	0.235
	instruct	0.130	0.261	0.261	0.092	0.239	0.178	0.156	0.197	0.090	0.108	0.171
Mistral	base	0.239	0.323	0.116	0.149	0.390	0.322	0.225	0.206	0.219	0.187	0.238
	instruct	0.042	0.279	0.306	0.268	0.041	0.528	0.187	0.285	0.214	0.125	0.228
Granite	base	0.385	0.262	0.186	0.146	0.243	0.188	0.202	0.230	0.285	0.238	0.236
	instruct	0.046	0.188	0.090	0.147	0.063	0.147	0.113	0.327	0.174	0.173	0.147

Table 45: Expected Calibration Error (ECE) Comparison: Base vs. Instruct Models

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	HateX	BoolQ	TREC	Mean
Llama	0.0273	0.2047	0.2386	0.3250	0.1902	0.3705	0.2206	0.1569	0.1044	0.0533	0.1891
Gemma	0.0855	0.2946	0.1528	0.2805	0.1771	0.4028	0.1909	0.5324	0.1072	0.2803	0.2504
Qwen	0.0959	0.2789	0.1993	0.2491	0.2504	0.3840	0.1149	0.7250	0.0893	0.2637	0.2650
Mistral	0.0664	0.4490	0.1262	0.2825	0.2116	0.1548	0.2477	0.4551	0.1549	0.1850	0.2333
DeepSeek	0.1123	0.2145	0.3320	0.3760	0.2752	0.0752	0.1965	0.2828	0.1298	0.1568	0.2151
EuroLLM	0.0531	0.3114	0.2244	0.3565	0.3638	0.2805	0.1475	0.6043	0.0958	0.1897	0.2627
EXAONE	0.1112	0.3579	0.1647	0.3327	0.2300	0.2481	0.1725	0.6439	0.1212	0.2374	0.2620
Ministral	0.0458	0.3614	0.2739	0.2362	0.2238	0.3189	0.1114	0.6541	0.1227	0.1097	0.2458
Granite	0.0907	0.3840	0.1421	0.1362	0.1504	0.2354	0.1787	0.6179	0.1129	0.1563	0.2205
Kanana	0.0965	0.4490	0.3677	0.3762	0.3847	0.4070	0.1643	0.6820	0.1467	0.2400	0.3314
Mengzi	0.0881	0.4281	0.2444	0.2403	0.3309	0.1552	0.1515	0.6997	0.1412	0.1611	0.2640
GLM-4	0.1974	0.3359	0.2278	0.2617	0.2721	0.2023	0.1185	0.6380	0.1202	0.1498	0.2524
Falcon	0.0826	0.4825	0.1684	0.3585	0.3627	0.2747	0.2441	0.6766	0.1368	0.1888	0.2976
SOLAR	0.1400	0.3960	0.4119	0.2410	0.2135	0.1776	0.2224	0.5306	0.1187	0.1981	0.2650
Mean	0.0923	0.3534	0.2339	0.2895	0.2597	0.2634	0.1772	0.5642	0.1216	0.1836	0.2539

Table 46: Expected Calibration Error (ECE) across Evaluated Models and Datasets. The Mean column is an unweighted macro-average over the ten datasets, not an instance-weighted pooled score.

Dataset	SC Corr.	Logprob Corr.	SC Win Rate %
BoolQ	0.300	0.271	0.571
CoLA	0.312	0.180	0.857
HateXplain	0.052	-0.022	0.786
MNLI	0.285	0.285	0.357
MRPC	0.286	0.212	0.643
SNLI	0.307	0.271	0.643
SST2	0.555	0.329	0.857
SST5	0.153	0.177	0.357
SWAG	0.278	0.344	0.071
TREC	0.468	0.328	0.643
Mean	0.300	0.237	0.579

Table 47: Self-Consistency and Logprob Correlation Metrics with Self-Consistency Win Rates Across Datasets

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	Hate Xplain	BoolQ	TREC	Total
Llama	116.3	731.0	123.8	2068.7	2036.6	259.8	4591.8	239.4	479.5	110.2	10757.1
Gemma	123.5	773.4	147.3	2059.2	2137.7	284.0	7160.4	404.5	738.4	208.0	14036.4
Qwen	438.3	2467.5	443.1	6970.7	7469.9	1012.4	20819.8	1422.9	2098.4	906.4	44049.4
Mistral	72.7	470.6	89.3	1270.6	1414.9	216.0	4975.4	247.6	515.2	128.2	9400.5
DeepSeek	63.2	412.6	78.6	1102.8	1202.6	180.1	4333.4	214.5	442.0	113.1	8142.9
EuroLLM	93.6	593.5	112.6	1586.7	1763.0	276.6	6531.7	329.9	643.6	162.2	12093.4
EXAONE	76.3	484.8	92.1	1296.6	1487.1	222.8	5349.7	257.2	530.9	131.1	9928.6
Ministral	83.7	532.2	100.0	1425.1	1526.5	220.4	5242.5	282.5	555.1	143.4	10111.4
Granite	90.9	580.6	108.9	1556.5	1661.7	234.8	5723.5	301.7	584.4	158.2	11001.2
Kanana	77.6	496.0	92.4	1331.5	1450.0	207.0	5086.1	262.3	523.2	133.2	9659.3
Mengzi	75.1	485.8	90.2	1302.9	1393.1	194.4	4749.8	254.9	503.4	131.0	9180.6
GLM-4	92.2	599.8	105.2	1615.7	1666.6	212.2	5153.0	310.2	562.5	160.0	10477.4
Falcon	103.0	657.2	122.3	1752.8	1868.0	266.2	6413.1	342.2	647.1	175.2	12347.1
SOLAR	110.8	704.3	133.7	1915.6	2201.0	332.2	7765.6	375.4	789.7	191.2	14519.5
Mean	115.5	713.5	131.4	1946.8	2091.3	294.2	6706.8	374.7	686.7	203.7	13264.6

Table 48: Logprob Inference Time Across Evaluation Datasets

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	Hate Xplain	BoolQ	TREC	Total
Llama	296.2	1052.2	629.2	6194.7	6009.1	1537.4	5268.4	920.5	1000.2	229.3	23137.2
Gemma	309.0	1099.0	651.2	6235.4	6377.4	1636.4	7994.8	1446.0	1514.3	331.4	27594.9
Qwen	782.3	2132.9	1114.3	11184.0	11540.0	2884.0	17970.0	2565.3	3650.7	755.9	54579.4
Mistral	190.5	620.5	339.5	3135.4	3446.9	953.5	5497.0	540.4	1063.7	124.9	15912.3
DeepSeek	158.6	497.5	239.9	3000.6	3092.5	646.7	4676.7	437.6	891.7	118.0	13759.8
EuroLLM	239.5	633.2	451.8	4730.2	4919.5	1192.0	7145.7	653.3	1326.1	198.0	21489.3
EXAONE	192.2	563.6	374.0	3905.2	4107.7	814.6	5784.6	826.4	1080.9	202.6	17851.8
Ministral	220.3	656.9	364.1	3303.6	3650.8	861.6	6054.2	704.6	1172.8	134.2	17123.1
Granite	230.8	654.8	326.2	4279.8	4517.8	1006.5	6414.8	862.0	1230.0	178.8	19701.5
Kanana	193.0	507.6	281.9	3613.7	3922.7	754.6	5464.9	698.8	1047.0	111.8	16596.0
Mengzi	190.1	539.6	273.1	3237.1	3662.8	741.0	5330.2	743.2	1037.5	111.9	15866.5
GLM-4	241.1	884.6	436.8	4659.7	4834.4	1121.7	6217.8	1094.3	1224.3	196.0	20910.7
Falcon	255.5	821.4	359.5	4050.0	4219.0	922.3	7067.0	754.1	1342.8	181.8	19973.4
SOLAR	328.6	1195.1	788.1	6639.5	7838.4	2580.0	11589.8	1486.7	2344.4	278.5	35069.1
Mean	273.4	847.1	473.5	4869.2	5152.8	1260.9	7319.7	980.9	1423.3	225.2	22826.1

Table 49: Self-Consistency Inference Time Across Datasets

Dataset	Model	Threshold (P_{75})	True Positive	Overconf. Error	Underconf. Correct	True Negative	Overconf. Error Rate
SST2 ($N = 872$)	Llama	0.9882	261	0	555	56	0.0000
	Gemma	0.9882	768	38	43	23	0.0436
	Qwen	0.9882	547	6	264	55	0.0069
	Mistral	0.9882	775	36	36	25	0.0413
	DeepSeek	0.9882	389	17	322	144	0.0195
	EuroLLM	0.9882	194	3	595	80	0.0034
	EXAONE	0.9882	665	21	136	50	0.0241
	Ministral	0.9882	399	2	399	72	0.0023
	Granite	0.9882	762	33	52	25	0.0378
	Kanana	0.9882	637	44	135	56	0.0505
	Mengzi	0.9882	690	26	101	55	0.0298
	GLM-4	0.9882	328	15	445	84	0.0172
	Falcon	0.9882	757	26	51	38	0.0298
	SOLAR	0.9882	714	32	88	38	0.0367
SST5 ($N = 2210$)	Llama	0.6530	88	42	1038	1042	0.0190
	Gemma	0.6530	876	633	282	419	0.2864
	Qwen	0.6530	542	775	137	756	0.3507
	Mistral	0.6530	737	1003	190	280	0.4538
	DeepSeek	0.6530	94	146	609	1361	0.0661
	EuroLLM	0.6530	498	889	148	675	0.4023
	EXAONE	0.6530	649	1034	139	388	0.4679
	Ministral	0.6530	625	1028	40	517	0.4652
	Granite	0.6530	773	1015	127	295	0.4593
	Kanana	0.6530	628	1300	21	261	0.5882
	Mengzi	0.6530	645	1176	66	323	0.5321
	GLM-4	0.6530	618	1054	97	441	0.4769
	Falcon	0.6530	662	1321	17	210	0.5977
	SOLAR	0.6530	947	932	151	180	0.4217

Table 50: Logprob Confidence Correctness Breakdown: Sentiment Analysis (SST2 & SST5)

Dataset	Model	Threshold (P_{75})	True Positive	Overconf. Error	Underconf. Correct	True Negative	Overconf. Error Rate
CoLA ($N = 1043$)	Llama	0.7866	216	91	397	339	0.0872
	Gemma	0.7866	803	139	52	49	0.1333
	Qwen	0.7866	312	42	471	218	0.0403
	Mistral	0.7866	799	157	57	30	0.1505
	DeepSeek	0.7866	142	173	401	327	0.1659
	EuroLLM	0.7866	322	719	1	1	0.6894
	EXAONE	0.7866	759	125	87	72	0.1198
	Ministral	0.7866	361	75	391	216	0.0719
	Granite	0.7866	817	143	51	32	0.1371
	Kanana	0.7866	472	318	120	133	0.3049
	Mengzi	0.7866	707	237	39	60	0.2272
	GLM-4	0.7866	243	30	607	163	0.0288
	Falcon	0.7866	761	147	74	61	0.1409
	SOLAR	0.7866	345	346	134	218	0.3317
SNLI ($N = 9824$)	Llama	0.8458	3324	2189	1661	2650	0.2228
	Gemma	0.8458	5243	2599	1091	891	0.2646
	Qwen	0.8458	3331	3229	48	3216	0.3287
	Mistral	0.8458	5182	2473	1022	1147	0.2517
	DeepSeek	0.8458	3186	2927	1613	2098	0.2979
	EuroLLM	0.8458	3366	6443	3	12	0.6558
	EXAONE	0.8458	3399	3581	221	2623	0.3645
	Ministral	0.8458	3221	2967	147	3489	0.3020
	Granite	0.8458	6842	1284	1073	625	0.1307
	Kanana	0.8458	3244	1440	2480	2660	0.1466
	Mengzi	0.8458	3193	1820	1124	3687	0.1853
	GLM-4	0.8458	3327	4613	246	1638	0.4696
	Falcon	0.8458	3304	803	4108	1609	0.0817
	SOLAR	0.8458	5067	552	2292	1913	0.0562
MNLI ($N = 9815$)	Llama	0.8418	3834	1176	2638	2167	0.1198
	Gemma	0.8418	6470	1425	1043	877	0.1452
	Qwen	0.8418	3414	2775	285	3341	0.2827
	Mistral	0.8418	6378	1912	785	740	0.1948
	DeepSeek	0.8418	2812	1729	2607	2667	0.1762
	EuroLLM	0.8418	3479	6333	1	2	0.6452
	EXAONE	0.8418	3471	3675	48	2621	0.3744
	Ministral	0.8418	3232	4393	247	1943	0.4476
	Granite	0.8418	6696	1399	964	756	0.1425
	Kanana	0.8418	3053	887	2031	3844	0.0904
	Mengzi	0.8418	3307	2036	2380	2092	0.2074
	GLM-4	0.8418	3451	3816	540	2008	0.3888
	Falcon	0.8418	3416	1023	3094	2282	0.1042
	SOLAR	0.8418	5800	673	2073	1269	0.0686

Table 51: Logprob Confidence Correctness Breakdown: Linguistic Analysis & Inference (CoLA, SNLI, MNLI)

Dataset	Model	Threshold (P_{75})	True Positive	Overconf. Error	Underconf. Correct	True Negative	Overconf. Error Rate
MRPC ($N = 1725$)	Llama	0.6249	214	32	1112	367	0.0186
	Gemma	0.6249	810	524	229	162	0.3038
	Qwen	0.6249	399	176	488	662	0.1020
	Mistral	0.6249	1235	321	90	79	0.1861
	DeepSeek	0.6249	573	169	556	427	0.0980
	EuroLLM	0.6249	472	459	117	677	0.2661
	EXAONE	0.6249	572	1020	23	110	0.5913
	Ministral	0.6249	530	541	68	586	0.3136
	Granite	0.6249	1147	301	161	116	0.1745
	Kanana	0.6249	320	156	880	369	0.0904
	Mengzi	0.6249	950	191	345	239	0.1107
	GLM-4	0.6249	577	1132	10	6	0.6562
	Falcon	0.6249	581	1141	0	3	0.6614
	SOLAR	0.6249	1318	319	50	38	0.1849
SWAG ($N = 20006$)	Llama	0.6008	3278	319	8859	7550	0.0159
	Gemma	0.6008	14209	4125	645	1027	0.2062
	Qwen	0.6008	11461	869	4297	3379	0.0434
	Mistral	0.6008	10760	5360	1400	2486	0.2679
	DeepSeek	0.6008	3118	3643	4010	9235	0.1821
	EuroLLM	0.6008	5237	1044	6742	6983	0.0522
	EXAONE	0.6008	12913	3687	1414	1992	0.1843
	Ministral	0.6008	9223	2302	3719	4762	0.1151
	Granite	0.6008	13983	3792	869	1362	0.1895
	Kanana	0.6008	10678	3538	2233	3557	0.1768
	Mengzi	0.6008	8977	3533	3066	4430	0.1766
	GLM-4	0.6008	11456	2362	2914	3274	0.1181
	Falcon	0.6008	13269	5049	633	1055	0.2524
	SOLAR	0.6008	15050	4421	191	344	0.2210

Table 52: Logprob Confidence Correctness Breakdown: Paraphrasing & Multiple Choice (MRPC & SWAG)

Dataset	Model	Threshold (P_{75})	True Positive	Overconf. Error	Underconf. Correct	True Negative	Overconf. Error Rate
HateXplain ($N = 1924$)	Llama	0.9546	1	15	630	1278	0.0078
	Gemma	0.9546	416	963	216	329	0.5005
	Qwen	0.9546	510	1295	38	81	0.6731
	Mistral	0.9546	104	439	461	920	0.2282
	DeepSeek	0.9546	39	131	585	1169	0.0681
	EuroLLM	0.9546	547	1370	1	6	0.7121
	EXAONE	0.9546	535	1279	12	98	0.6648
	Ministral	0.9546	504	1300	44	76	0.6757
	Granite	0.9546	309	1016	234	365	0.5281
	Kanana	0.9546	453	1186	95	190	0.6164
	Mengzi	0.9546	518	1289	29	88	0.6700
	GLM-4	0.9546	426	975	121	402	0.5068
	Falcon	0.9546	394	1141	158	231	0.5930
	SOLAR	0.9546	242	650	268	764	0.3378
BoolQ ($N = 3270$)	Llama	0.9481	1005	63	1592	610	0.0193
	Gemma	0.9481	2797	291	102	80	0.0890
	Qwen	0.9481	1410	84	1418	358	0.0257
	Mistral	0.9481	2565	434	149	122	0.1327
	DeepSeek	0.9481	522	39	1822	887	0.0119
	EuroLLM	0.9481	1218	54	1518	480	0.0165
	EXAONE	0.9481	2436	274	349	211	0.0838
	Ministral	0.9481	1942	198	746	384	0.0606
	Granite	0.9481	2649	296	184	141	0.0905
	Kanana	0.9481	2145	271	499	355	0.0829
	Mengzi	0.9481	2295	330	375	270	0.1009
	GLM-4	0.9481	985	124	1715	446	0.0379
	Falcon	0.9481	2728	387	83	72	0.1183
	SOLAR	0.9481	2800	332	70	68	0.1015
TREC ($N = 500$)	Llama	0.6377	207	16	112	165	0.0320
	Gemma	0.6377	302	110	23	65	0.2200
	Qwen	0.6377	198	133	81	88	0.2660
	Mistral	0.6377	354	105	15	26	0.2100
	DeepSeek	0.6377	62	73	83	282	0.1460
	EuroLLM	0.6377	75	286	37	102	0.5720
	EXAONE	0.6377	191	106	61	142	0.2120
	Ministral	0.6377	370	20	58	52	0.0400
	Granite	0.6377	316	81	49	54	0.1620
	Kanana	0.6377	175	63	138	124	0.1260
	Mengzi	0.6377	224	72	80	124	0.1440
	GLM-4	0.6377	292	63	54	91	0.1260
	Falcon	0.6377	282	77	68	73	0.1540
	SOLAR	0.6377	116	321	11	52	0.6420

Table 53: Logprob Confidence Correctness Breakdown: Safety Compliance & QA Tasks (HateXplain, BoolQ, TREC)

Model	SST2	SST5	CoLA	SNLI	MNLI	MRPC	SWAG	HateX	BoolQ	TREC	Mean
Llama	0.0004	0.1540	0.0664	0.1727	0.0817	0.1275	0.0296	0.3044	0.0006	0.1859	0.1123
Gemma	0.0020	0.2356	0.0893	0.1581	0.0067	0.0005	0.0009	0.2361	0.0167	0.4047	0.1151
Qwen	0.0006	0.3092	0.0905	0.0244	0.0061	0.0009	0.0019	0.2982	0.0060	0.4238	0.1162
Mistral	0.0000	0.2175	0.0816	0.0734	0.0578	0.2413	0.1264	0.1639	0.0056	0.4021	0.1370
DeepSeek	0.0417	0.6159	0.0021	0.3987	0.3900	0.6931	0.2370	0.4474	0.0417	0.4107	0.3278
EuroLLM	0.0000	0.7526	0.0296	0.2729	0.2875	0.0625	0.0207	1.0319	0.0132	0.0777	0.2549
EXAONE	0.0014	0.2643	0.1825	0.0137	0.0289	0.0264	0.0016	0.2339	0.0017	0.3068	0.1061
Ministral	0.0029	0.0508	0.1896	0.1572	0.0607	0.0343	0.0089	0.5500	0.0008	0.1704	0.1226
Granite	0.0003	0.1621	0.0839	0.0064	0.0131	0.1478	0.0125	0.1375	0.0065	0.1903	0.0760
Kanana	0.0149	0.4343	0.0025	0.0630	0.2781	0.0067	0.0077	0.2780	0.0107	0.1907	0.1287
Mengzi	0.0044	0.1533	0.0038	0.1596	0.0707	0.1241	0.0083	0.7570	0.0025	0.3619	0.1646
GLM-4	0.0029	0.1406	0.1197	0.0233	0.0049	0.0640	0.0575	0.1542	0.0100	0.2160	0.0793
Falcon	0.0004	0.1364	0.0544	0.0015	0.0047	0.0213	0.0057	0.1456	0.0474	0.1672	0.0585
SOLAR	0.0026	0.0399	0.2591	0.0819	0.0126	0.1243	0.0128	0.0690	0.0112	0.1775	0.0791
Mean	0.0053	0.2619	0.0896	0.1148	0.0931	0.1196	0.0380	0.3434	0.0125	0.2633	0.1342

Table 54: Label Bias Index (LBI) across Evaluated Models and Datasets

B Prompt Examples

We provide concrete examples of each prompt type used in our hierarchical ensembling framework, illustrating the $p = [r, t, l, x]$ structure (role, task instruction, label specification, input text). Examples are drawn from all ten datasets. For few-shot prompts, one exemplar is provided per label class.

B.1 Prompt 1: Zero-Shot Direct

Task: Sentiment Analysis (SST-2)

[Role] You are an expert sentiment analysis system.

[Instruction] Classify the sentiment of the following text. Respond with exactly one label from: [positive, negative]. Do not include any explanation or additional text.

[Input] **Text:** “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”

[Output] **Label:**

Figure 1: Zero-shot prompt example for SST-2.

Task: Sentiment Analysis, Fine-Grained (SST-5)

[Role] You are an expert sentiment analysis system.

[Instruction] Classify the sentiment of the following text. Respond with exactly one label from: [very negative, negative, neutral, positive, very positive]. Do not include any explanation or additional text.

[Input] **Text:** “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”

[Output] **Label:**

Figure 2: Zero-shot prompt example for SST-5.

Task: Linguistic Acceptability (CoLA)

[Role] You are an expert linguist specializing in grammatical analysis.

[Instruction] Classify whether the following sentence is grammatically acceptable. Respond with exactly one label from: [acceptable, unacceptable]. Do not include any explanation or additional text.

[Input] **Sentence:** “The children seems to be playing outside.”

[Output] **Label:**

Figure 3: Zero-shot prompt example for CoLA.

Task: Natural Language Inference (SNLI)

[Role] You are an expert in natural language inference.

[Instruction] Given a premise and a hypothesis, classify their relationship. Respond with exactly one label from: [entailment, neutral, contradiction]. Do not include any explanation or additional text.

[Input] **Premise:** “A woman in a red jacket is walking her dog along a riverbank.”
Hypothesis: “A person is outdoors with an animal.”

[Output] **Label:**

Figure 4: Zero-shot prompt example for SNLI.

Task: Natural Language Inference, Multi-Genre (MNLI)

[Role] You are an expert in natural language inference.

[Instruction] Given a premise and a hypothesis, classify their relationship. Respond with exactly one label from: [entailment, neutral, contradiction]. Do not include any explanation or additional text.

[Input] **Premise:** “The new policy was announced at a press conference last Tuesday.”
Hypothesis: “An official statement was made publicly.”

[Output] **Label:**

Figure 5: Zero-shot prompt example for MNLI.

Task: Paraphrase Detection (MRPC)

[Role] You are an expert in semantic similarity and paraphrase detection.

[Instruction] Determine whether the two sentences are paraphrases of each other. Respond with exactly one label from: [paraphrase, not paraphrase]. Do not include any explanation or additional text.

[Input] **Sentence 1:** “The stock market fell sharply on Friday following the announcement.”
Sentence 2: “Shares dropped significantly after Friday’s announcement.”

[Output] **Label:**

Figure 6: Zero-shot prompt example for MRPC.

Task: Commonsense Reasoning (SWAG)
[Role] You are an expert in commonsense reasoning.
[Instruction] Given the beginning of a sentence, select the most plausible continuation. Respond with exactly one label from: [A, B, C, D]. Do not include any explanation or additional text.
[Input] **Beginning:** “She poured the coffee into the mug and”
A: “added a splash of milk before sitting down.”
B: “drove the car into the garage.”
C: “painted the wall a bright yellow.”
D: “signed the legal document carefully.”
[Output] **Label:**

Figure 7: Zero-shot prompt example for SWAG.

Task: Hate Speech Detection (HateXplain)
[Role] You are an expert content moderation system.
[Instruction] Classify the following text for harmful content. Respond with exactly one label from: [normal, offensive, hatespeech]. Do not include any explanation or additional text.
[Input] **Text:** “I can’t believe they let people like that into the neighborhood.”
[Output] **Label:**

Figure 8: Zero-shot prompt example for HateXplain.

Task: Boolean Question Answering (BoolQ)
[Role] You are an expert reading comprehension system.
[Instruction] Based on the passage, answer the question. Respond with exactly one label from: [true, false]. Do not include any explanation or additional text.
[Input] **Passage:** “The platypus is one of the few mammals known to lay eggs. It is native to eastern Australia and Tasmania.”
Question: “Does the platypus lay eggs?”
[Output] **Label:**

Figure 9: Zero-shot prompt example for BoolQ.

Task: Question Classification (TREC)
[Role] You are an expert question classification system.
[Instruction] Classify the type of answer the following question is seeking. Respond with exactly one label from: [abbreviation, entity, description, human, location, numeric]. Do not include any explanation or additional text.
[Input] **Question:** “What is the capital of Australia?”
[Output] **Label:**

Figure 10: Zero-shot prompt example for TREC.

B.2 Prompt 2: Few-Shot In-Context Learning

Task: Sentiment Analysis (SST-2)
[Role] You are an expert sentiment analysis system.
[Instruction] Classify the sentiment of the following text. Respond with exactly one label from: [positive, negative]. Do not include any explanation or additional text.
Here are some examples:
[Exemplar 1]
Text: “The film captivates from the first frame to the last, delivering an emotional tour de force.”
Label: positive
[Exemplar 2]
Text: “A plodding, joyless slog that wastes a talented cast on a threadbare script.”
Label: negative
Now classify the following:
[Input] **Text:** “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”
[Output] **Label:**

Figure 11: Few-shot prompt example for SST-2 (one exemplar per label: positive, negative).

Task: Sentiment Analysis, Fine-Grained (SST-5)

[Role] You are an expert sentiment analysis system.

[Instruction] Classify the sentiment of the following text. Respond with exactly one label from: [very negative, negative, neutral, positive, very positive]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Text: "A catastrophic failure of filmmaking on every level imaginable."
Label: very negative

[Exemplar 2]
Text: "A plodding, joyless slog that wastes a talented cast on a threadbare script."
Label: negative

[Exemplar 3]
Text: "A competent but unremarkable film that neither impresses nor disappoints."
Label: neutral

[Exemplar 4]
Text: "The film captivates from the first frame to the last, delivering an emotional tour de force."
Label: positive

[Exemplar 5]
Text: "A breathtaking masterpiece that redefines the possibilities of cinema."
Label: very positive

Now classify the following:

[Input] **Text:** "A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission."

[Output] **Label:**

Figure 12: Few-shot prompt example for SST-5 (one exemplar per label: very negative, negative, neutral, positive, very positive).

Task: Linguistic Acceptability (CoLA)

[Role] You are an expert linguist specializing in grammatical analysis.

[Instruction] Classify whether the following sentence is grammatically acceptable. Respond with exactly one label from: [acceptable, unacceptable]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Sentence: "The students finished their assignments before the deadline."
Label: acceptable

[Exemplar 2]
Sentence: "The children seems to be playing outside."
Label: unacceptable

Now classify the following:

[Input] **Sentence:** "She gave him the book what he wanted."

[Output] **Label:**

Figure 13: Few-shot prompt example for CoLA (one exemplar per label: acceptable, unacceptable).

Task: Natural Language Inference (SNLI)

[Role] You are an expert in natural language inference.

[Instruction] Given a premise and a hypothesis, classify their relationship. Respond with exactly one label from: [entailment, neutral, contradiction]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Premise: "Two children are playing soccer in the park."
Hypothesis: "Kids are engaged in outdoor activity."
Label: entailment

[Exemplar 2]
Premise: "A group of musicians is rehearsing in a studio."
Hypothesis: "The musicians are preparing for a concert."
Label: neutral

[Exemplar 3]
Premise: "A man is reading a book on a park bench."
Hypothesis: "The man is asleep outside."
Label: contradiction

Now classify the following:

[Input] **Premise:** "A woman in a red jacket is walking her dog along a riverbank."
Hypothesis: "A person is outdoors with an animal."

[Output] **Label:**

Figure 14: Few-shot prompt example for SNLI (one exemplar per label: entailment, neutral, contradiction).

Task: Natural Language Inference, Multi-Genre (MNLI)

[Role] You are an expert in natural language inference.

[Instruction] Given a premise and a hypothesis, classify their relationship. Respond with exactly one label from: [entailment, neutral, contradiction]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Premise: "The legislation was passed unanimously by the senate."
Hypothesis: "A vote took place in the senate."
Label: entailment

[Exemplar 2]
Premise: "The chef prepared a three-course meal for the guests."
Hypothesis: "The guests enjoyed the meal."
Label: neutral

[Exemplar 3]
Premise: "The museum is open every day of the week."
Hypothesis: "The museum is closed on Sundays."
Label: contradiction

Now classify the following:
[Input] **Premise:** "The new policy was announced at a press conference last Tuesday."
Hypothesis: "An official statement was made publicly."
[Output] **Label:**

Figure 15: Few-shot prompt example for MNLI (one exemplar per label: entailment, neutral, contradiction).

Task: Paraphrase Detection (MRPC)

[Role] You are an expert in semantic similarity and paraphrase detection.

[Instruction] Determine whether the two sentences are paraphrases of each other. Respond with exactly one label from: [paraphrase, not paraphrase]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Sentence 1: "The stock market fell sharply on Friday following the announcement."
Sentence 2: "Shares dropped significantly after Friday's announcement."
Label: paraphrase

[Exemplar 2]
Sentence 1: "The president signed the bill into law on Monday."
Sentence 2: "The governor vetoed the legislation last week."
Label: not paraphrase

Now classify the following:
[Input] **Sentence 1:** "He said the company plans to hire 500 new employees this year."
Sentence 2: "The company intends to expand its workforce by 500 in the coming year."
[Output] **Label:**

Figure 16: Few-shot prompt example for MRPC (one exemplar per label: paraphrase, not paraphrase).

Task: Commonsense Reasoning (SWAG)

[Role] You are an expert in commonsense reasoning.

[Instruction] Given the beginning of a sentence, select the most plausible continuation. Respond with exactly one label from: [A, B, C, D]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Beginning: "She laced up her running shoes and"
A: "headed out for her morning jog."
B: "started writing a letter."
C: "boiled a pot of pasta."
D: "fell asleep on the couch."
Label: A

[Exemplar 2]
Beginning: "He opened the textbook and"
A: "jumped into the swimming pool."
B: "began reviewing the chapter on calculus."
C: "ordered a pizza for delivery."
D: "painted the fence white."
Label: B

[Exemplar 3]
Beginning: "The mechanic opened the hood and"
A: "baked a loaf of bread."
B: "recited a poem from memory."
C: "inspected the engine carefully."
D: "planted a row of tomatoes."
Label: C

[Exemplar 4]
Beginning: "The waiter brought the menus and"
A: "climbed the ladder to the roof."
B: "wrote an equation on the board."
C: "repaired the bicycle tire."
D: "described the evening specials to the table."
Label: D

Now classify the following:
[Input] **Beginning:** "She poured the coffee into the mug and"
A: "added a splash of milk before sitting down."
B: "drove the car into the garage."
C: "painted the wall a bright yellow."
D: "signed the legal document carefully."
[Output] **Label:**

Figure 17: Few-shot prompt example for SWAG (one exemplar per label: A, B, C, D).

Task: Hate Speech Detection (HateXplain)

[Role] You are an expert content moderation system.

[Instruction] Classify the following text for harmful content. Respond with exactly one label from: [normal, offensive, hatespeech]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Text: "The weather today is perfect for a walk in the park."
Label: normal

[Exemplar 2]
Text: "That was a completely idiotic thing to say."
Label: offensive

[Exemplar 3]
Text: "[Derogatory statement targeting a racial group]."
Label: hatespeech

Now classify the following:
[Input] **Text:** "I can't believe they let people like that into the neighborhood."
[Output] **Label:**

Figure 18: Few-shot prompt example for HateXplain (one exemplar per label: normal, offensive, hatespeech).

Task: Boolean Question Answering (BoolQ)

[Role] You are an expert reading comprehension system.

[Instruction] Based on the passage, answer the question. Respond with exactly one label from: [true, false]. Do not include any explanation or additional text.

Here are some examples:

[Exemplar 1]
Passage: "The platypus is one of the few mammals known to lay eggs. It is native to eastern Australia and Tasmania."
Question: "Does the platypus lay eggs?"
Label: true

[Exemplar 2]
Passage: "Mount Everest is the highest mountain in the world, located in the Himalayas on the border of Nepal and Tibet."
Question: "Is Mount Everest located in the Alps?"
Label: false

Now classify the following:
[Input] **Passage:** "Dolphins are highly intelligent marine mammals that communicate using a system of clicks and whistles."
Question: "Do dolphins use sound to communicate?"
[Output] **Label:**

Figure 19: Few-shot prompt example for BoolQ (one exemplar per label: true, false).

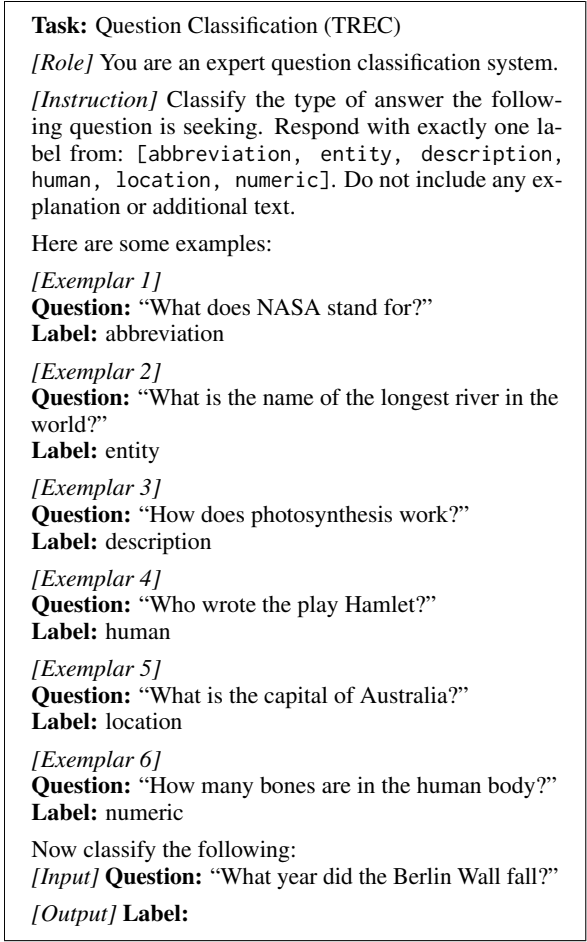


Figure 20: Few-shot prompt example for TREC (one exemplar per label: abbreviation, entity, description, human, location, numeric).

B.3 Prompt 3: Chain-of-Thought Reasoning

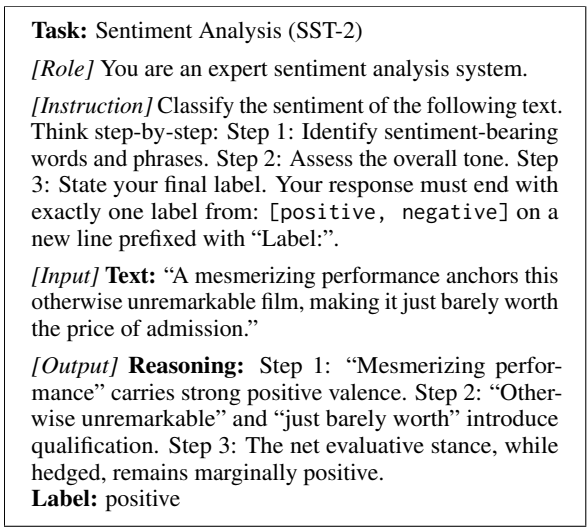


Figure 21: Chain-of-thought prompt example for SST-2. The *[Output]* block shows representative model output.

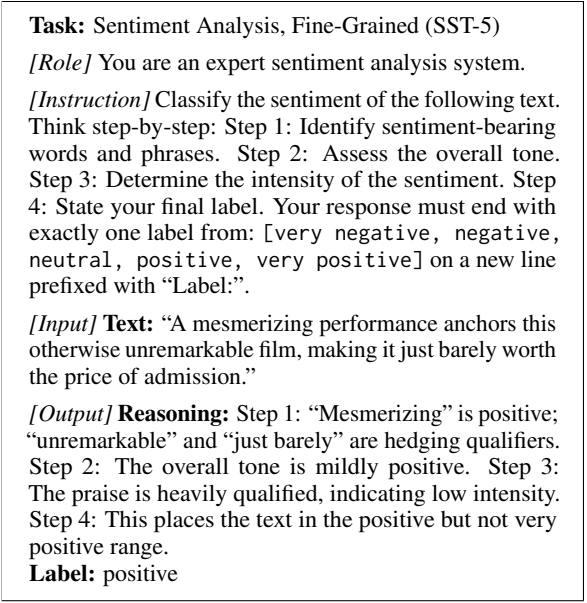


Figure 22: Chain-of-thought prompt example for SST-5. The *[Output]* block shows representative model output.

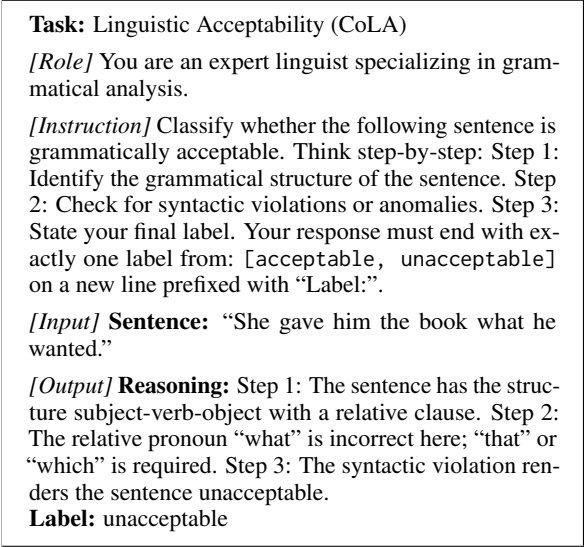


Figure 23: Chain-of-thought prompt example for CoLA. The *[Output]* block shows representative model output.

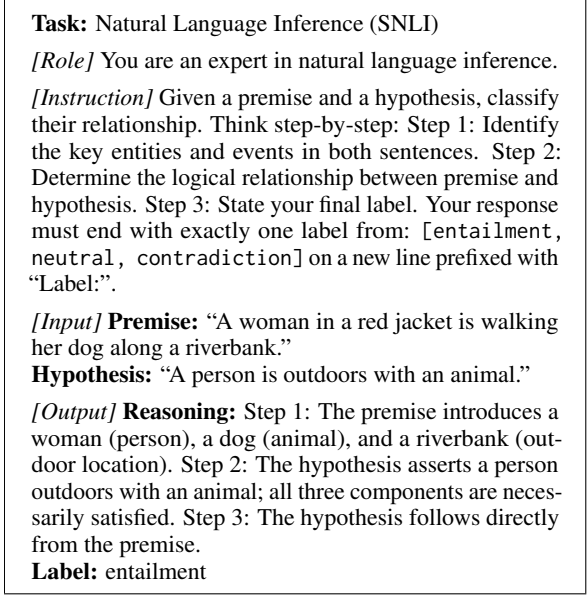


Figure 24: Chain-of-thought prompt example for SNLI. The *[Output]* block shows representative model output.

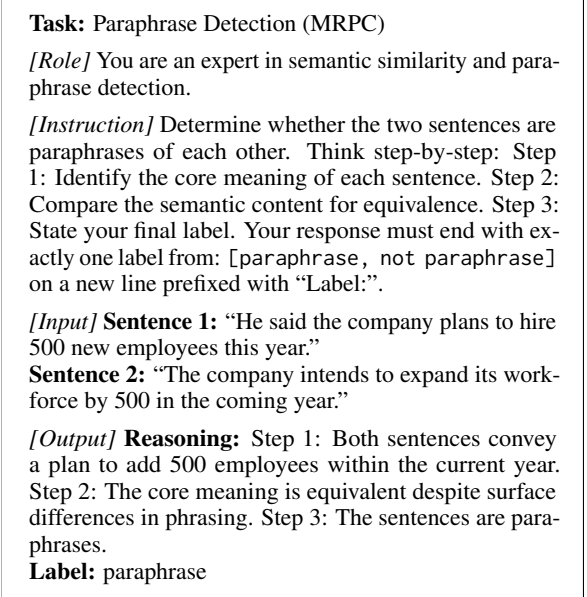


Figure 26: Chain-of-thought prompt example for MRPC. The *[Output]* block shows representative model output.

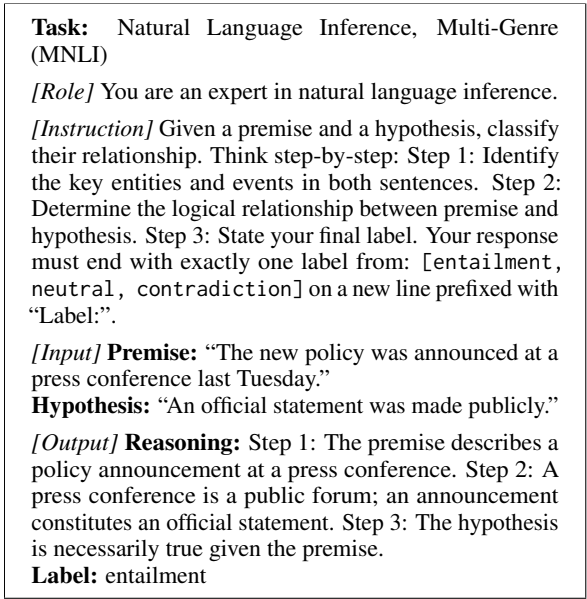


Figure 25: Chain-of-thought prompt example for MNLI. The *[Output]* block shows representative model output.

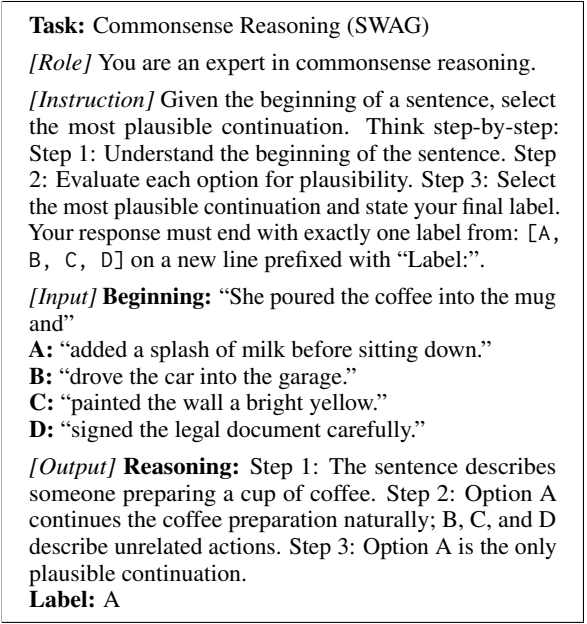


Figure 27: Chain-of-thought prompt example for SWAG. The *[Output]* block shows representative model output.

Task: Hate Speech Detection (HateXplain)

[Role] You are an expert content moderation system.

[Instruction] Classify the following text for harmful content. Think step-by-step: Step 1: Identify potentially harmful or offensive language. Step 2: Assess the severity and intent of the language. Step 3: State your final label. Your response must end with exactly one label from: [normal, offensive, hatespeech] on a new line prefixed with “Label:”.

[Input] Text: “I can’t believe they let people like that into the neighborhood.”

[Output] Reasoning: Step 1: The phrase “people like that” is vague but carries an othering tone. Step 2: Without explicit slurs or group targeting, the language is dismissive rather than hateful. Step 3: The text is offensive in tone but does not constitute hate speech.

Label: offensive

Figure 28: Chain-of-thought prompt example for HateXplain. The *[Output]* block shows representative model output.

Task: Question Classification (TREC)

[Role] You are an expert question classification system.

[Instruction] Classify the type of answer the following question is seeking. Think step-by-step: Step 1: Identify what the question is asking for. Step 2: Match the expected answer type to the closest category. Step 3: State your final label. Your response must end with exactly one label from: [abbreviation, entity, description, human, location, numeric] on a new line prefixed with “Label:”.

[Input] Question: “What year did the Berlin Wall fall?”

[Output] Reasoning: Step 1: The question asks for a specific year. Step 2: A year is a numeric value; this matches the numeric category. Step 3: The expected answer type is numeric.

Label: numeric

Figure 30: Chain-of-thought prompt example for TREC. The *[Output]* block shows representative model output.

C Alternative Prompt Configurations Considered

We present two alternative prompt configurations per prompt type that were considered alongside our final prompts to determine the best-performing design. Due to space constraints, we only show alternative prompts for the SST-2 and SNLI tasks.

Task: Boolean Question Answering (BoolQ)

[Role] You are an expert reading comprehension system.

[Instruction] Based on the passage, answer the question. Think step-by-step: Step 1: Read the passage and identify information relevant to the question. Step 2: Evaluate whether the passage supports or contradicts the question. Step 3: State your final label. Your response must end with exactly one label from: [true, false] on a new line prefixed with “Label:”.

[Input] Passage: “Dolphins are highly intelligent marine mammals that communicate using a system of clicks and whistles.”

Question: “Do dolphins use sound to communicate?”

[Output] Reasoning: Step 1: The passage states dolphins communicate via clicks and whistles. Step 2: Clicks and whistles are forms of sound, so the passage directly supports the question. Step 3: The answer is true.

Label: true

Figure 29: Chain-of-thought prompt example for BoolQ. The *[Output]* block shows representative model output.

C.1 Prompt 1 Alternatives: Zero-Shot Direct

Task: Sentiment Analysis (SST-2) *[Alternative 1A: No Role, Bare Instruction]*

[Instruction] What is the sentiment of the following text? Choose one: positive or negative. Output only the label.

[Input] Text: “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”

[Output] Label:

Figure 31: Zero-shot Alternative 1A for SST-2: No expert role, minimal imperative instruction with no explicit constraint framing.

Task: Natural Language Inference (SNLI) *[Alternative 1A: No Role, Bare Instruction]*

[Instruction] Does the hypothesis follow from the premise? Choose one: entailment, neutral, or contradiction. Output only the label.

[Input] **Premise:** “A woman in a red jacket is walking her dog along a riverbank.”
Hypothesis: “A person is outdoors with an animal.”

[Output] **Label:**

Figure 32: Zero-shot Alternative 1A for SNLI: No expert role, minimal imperative instruction.

Task: Sentiment Analysis (SST-2) *[Alternative 1B: Verbose Role + Penalty Constraint]*

[Role] You are a highly trained linguistic analyst specializing in opinion mining and affective text classification.

[Instruction] Your task is to determine the sentiment expressed in the input text below. You must respond with **exactly one word**: either positive or negative. Any response containing additional words, punctuation, or explanation will be considered invalid.

[Input] **Text:** “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”

[Output] **Label:**

Figure 33: Zero-shot Alternative 1B for SST-2: Verbose expert role with explicit invalidity penalty constraint.

Task: Natural Language Inference (SNLI) *[Alternative 1B: Verbose Role + Penalty Constraint]*

[Role] You are a highly trained linguistic analyst specializing in logical reasoning and textual entailment.

[Instruction] Determine the logical relationship between the premise and hypothesis below. You must respond with **exactly one word**: entailment, neutral, or contradiction. Any response containing additional words, punctuation, or explanation will be considered invalid.

[Input] **Premise:** “A woman in a red jacket is walking her dog along a riverbank.”
Hypothesis: “A person is outdoors with an animal.”

[Output] **Label:**

Figure 34: Zero-shot Alternative 1B for SNLI: Verbose expert role with explicit invalidity penalty constraint.

C.2 Prompt 2 Alternatives: Few-Shot In-Context Learning

Task: Sentiment Analysis (SST-2) *[Alternative 2A: 1-Shot, Single Exemplar]*

[Role] You are an expert sentiment analysis system.

[Instruction] Classify the sentiment of the following text. Respond with exactly one label from: [positive, negative]. Do not include any explanation or additional text.

[Exemplar 1]
Text: “The film captivates from the first frame to the last, delivering an emotional tour de force.”
Label: positive

[Input] **Text:** “A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission.”

[Output] **Label:**

Figure 35: Few-shot Alternative 2A for SST-2: Reduced to a single positive exemplar only (1-shot).

Task: Natural Language Inference (SNLI) *[Alternative 2A: 1-Shot, Single Exemplar]*

[Role] You are an expert in natural language inference.

[Instruction] Given a premise and a hypothesis, classify their relationship. Respond with exactly one label from: [entailment, neutral, contradiction]. Do not include any explanation or additional text.

[Exemplar 1]
Premise: “Two children are playing soccer in the park.”
Hypothesis: “Kids are engaged in outdoor activity.”
Label: entailment

[Input] **Premise:** “A woman in a red jacket is walking her dog along a riverbank.”
Hypothesis: “A person is outdoors with an animal.”

[Output] **Label:**

Figure 36: Few-shot Alternative 2A for SNLI: Reduced to a single entailment exemplar only (1-shot).

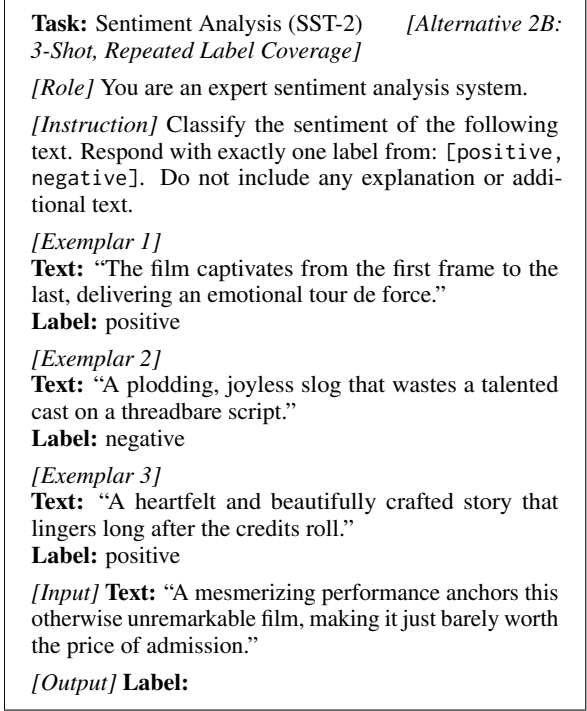


Figure 37: Few-shot Alternative 2B for SST-2: Extended to 3 exemplars with an imbalanced label distribution (2 positive, 1 negative) to test label-frequency sensitivity.

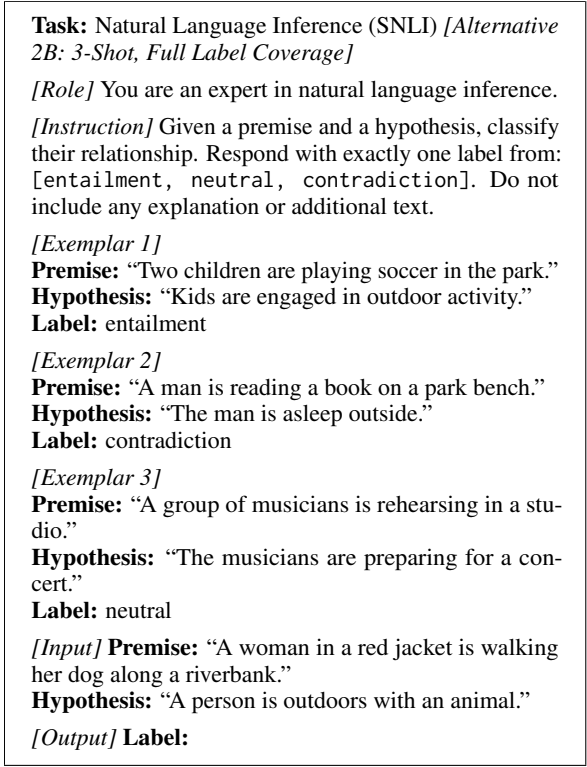


Figure 38: Few-shot Alternative 2B for SNLI: Extended to 3 exemplars with one example per label class to test full label coverage in-context.

C.3 Prompt 3 Alternatives: Chain-of-Thought Reasoning

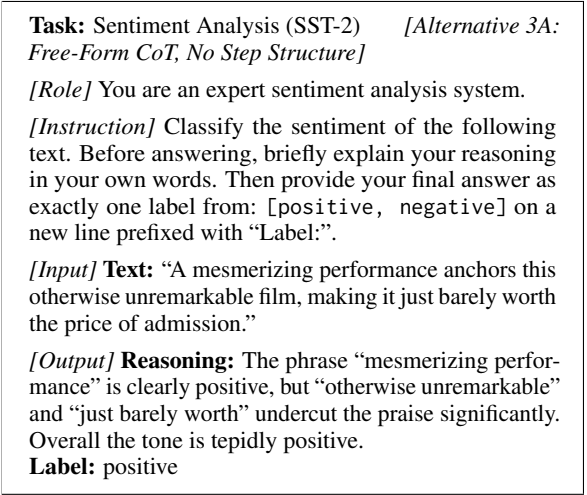


Figure 39: Chain-of-thought Alternative 3A for SST-2: Free-form reasoning without enforced step structure; representative model output shown.

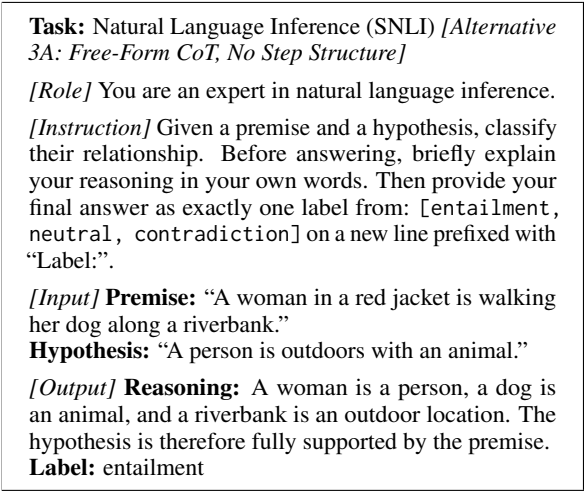


Figure 40: Chain-of-thought Alternative 3A for SNLI: Free-form reasoning without enforced step structure; representative model output shown.

Task: Sentiment Analysis (SST-2) [Alternative 3B: CoT with Contrastive Label Elimination]

[Role] You are an expert sentiment analysis system.

[Instruction] Classify the sentiment of the following text using a process of elimination. Step 1: State arguments for each possible label (positive, negative). Step 2: Weigh the evidence and eliminate the less supported label. Step 3: State your final answer as exactly one label from: [positive, negative] on a new line prefixed with "Label:".

[Input] Text: "A mesmerizing performance anchors this otherwise unremarkable film, making it just barely worth the price of admission."

[Output] Step 1: For positive: "Mesmerizing performance" and "worth the price of admission" suggest approval. For negative: "Otherwise unremarkable" and "just barely" suggest strong reservation.

Step 2: The positive evidence, though hedged, is not negated; the reviewer endorses the film however faintly.

Label: positive

Figure 41: Chain-of-thought Alternative 3B for SST-2: Contrastive elimination CoT; representative model output shown.

Task: Natural Language Inference (SNLI) [Alternative 3B: CoT with Contrastive Label Elimination]

[Role] You are an expert in natural language inference.

[Instruction] Classify the relationship between the premise and hypothesis using a process of elimination. Step 1: State arguments for each possible label (entailment, neutral, contradiction). Step 2: Eliminate labels that are not supported. Step 3: State your final answer as exactly one label from: [entailment, neutral, contradiction] on a new line prefixed with "Label:".

[Input] Premise: "A woman in a red jacket is walking her dog along a riverbank."

Hypothesis: "A person is outdoors with an animal."

[Output] Step 1: For entailment: woman $\Rightarrow$ person, dog $\Rightarrow$ animal, riverbank $\Rightarrow$ outdoors; all conditions met. For neutral: no ambiguity exists as all components map directly. For contradiction: nothing in the premise opposes the hypothesis.

Step 2: Neutral and contradiction are eliminated; the hypothesis is necessarily true.

Label: entailment

Figure 42: Chain-of-thought Alternative 3B for SNLI: Contrastive elimination CoT; representative model output shown.

C.4 Base vs Instruct Experiment Prompt Templates (Simple Few-Shot Design)

Text: hide new secretions from the parental units

Label: negative

Text: that loves its characters and communicates something rather beautiful about human nature

Label: positive

Text: [INPUT_SENTENCE]

Label:

Figure 43: Simple few-shot prompt template for SST-2.

Text: final verdict : you 've seen it all before .

Label: very negative

Text: apparently reassembled from the cutting-room floor of any given daytime soap .

Label: negative

Text: the entire movie is filled with deja vu moments .

Label: neutral

Text: this is a visually stunning rumination on love , memory , history and the war between art and commerce .

Label: positive

Text: a stirring , funny and finally transporting re-imagining of beauty and the beast and 1930s horror films

Label: very positive

Text: [INPUT_SENTENCE]

Label:

Figure 44: Simple few-shot prompt template for SST-5.

Sentence: They drank the pub.

Label: unacceptable

Sentence: Our friends won't buy this analysis, let alone the next one we propose.

Label: acceptable

Sentence: [INPUT_SENTENCE]

Label:

Figure 45: Simple few-shot prompt template for CoLA.

Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is outdoors, on a horse.
Label: entailment

Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is training his horse for a competition.
Label: neutral

Premise: A person on a horse jumps over a broken down airplane.
Hypothesis: A person is at a diner, ordering an omelette.
Label: contradiction

Premise: [INPUT_PREMISE]
Hypothesis: [INPUT_HYPOTHESIS]
Label:

Figure 46: Simple few-shot prompt template for SNLI.

Premise: you know during the season and i guess at at your level uh you lose them to the next level if if they decide to recall the the parent team the Braves decide to call to recall a guy from triple A then a double A guy goes up to replace him and a single A guy goes up to replace him
Hypothesis: You lose the things to the following level if the people recall.
Label: entailment

Premise: Conceptually cream skimming has two basic dimensions - product and geography.
Hypothesis: Product and geography are what make cream skimming work.
Label: neutral

Premise: Gays and lesbians.
Hypothesis: Heterosexuals.
Label: contradiction

Premise: [INPUT_PREMISE]
Hypothesis: [INPUT_HYPOTHESIS]
Label:

Figure 47: Simple few-shot prompt template for MNLI.

Sentence 1: Yucaipa owned Dominick 's before selling the chain to Safeway in 1998 for \$ 2.5 billion .
Sentence 2: Yucaipa bought Dominick 's in 1995 for \$ 693 million and sold it to Safeway for \$ 1.8 billion in 1998 .
Label: not paraphrase

Sentence 1: Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence .
Sentence 2: Referring to him as only " the witness " , Amrozi accused his brother of deliberately distorting his evidence .
Label: paraphrase

Sentence 1: [INPUT_SENTENCE_1]
Sentence 2: [INPUT_SENTENCE_2]
Label:

Figure 48: Simple few-shot prompt template for MRPC.

Beginning: Members of the procession walk down the street holding small horn brass instruments. A drum line
A: passes by walking down the street playing their instruments.
B: has heard approaching them.
C: arrives and they're outside dancing and asleep.
D: turns the lead singer watches the performance.
Label: A

Beginning: A group of members in green uniforms walks waving flags. Members of the procession
A: pay the other coaches to cheer as people this chatter dips in lawn sheets.
B: walk down the street holding small horn brass instruments.
C: is seen in the background.
D: are talking a couple of people playing a game of tug of war.
Label: B

Beginning: Someone stares blearily down at the floor. Someone
A: shifts his amused gaze to someone's brow pinched uneasily.
B: takes the gun and leads his men upstairs into the kitchen.
C: looks up at someone in shock.
D: takes the bag, revealing a sleek silver - diamond shot at someone's head.
Label: C

Beginning: A drum line passes by walking down the street playing their instruments. Members of the procession
A: are playing ping pong and celebrating one left each in quick.
B: wait slowly towards the cadets.
C: continues to play as well along the crowd along with the band being interviewed.
D: continue to play marching, interspersed.
Label: D

Beginning: [INPUT_STARTPHRASE]
A: [ENDING_0]
B: [ENDING_1]
C: [ENDING_2]
D: [ENDING_3]
Label:

Figure 49: Simple few-shot prompt template for SWAG.

Text: in britain the prime minister a childless woman london mayor a moslem invader in command of london police a lesbian in command of london fire brigade a fat woman scotland first minister a childless woman
Label: normal
Text: i live and work with many legal mexican immigrants who are great citizens and trump supporters they have no problem with deporting illegals maga
Label: offensive
Text: u really think i would not have been raped by feral hindu or muslim back in india or bangladesh and a neo nazi would rape me as well just to see me cry
Label: hatespeech
Text: [INPUT_TEXT]
Label:

Figure 50: Simple few-shot prompt template for HateXplain.

Question: What is the full form of .com ?
Label: abbreviation
Question: What films featured the character Popeye Doyle ?
Label: entity
Question: How did serfdom develop in and then leave Russia ?
Label: description
Question: What contemptible scoundrel stole the cork from my lunch ?
Label: human
Question: What sprawling U.S. state boasts the most airports ?
Label: location
Question: When was Ozzy Osbourne born ?
Label: numeric
Question: [INPUT_QUESTION]
Label:

Figure 52: Simple few-shot prompt template for TREC.

D Supplementary Figures

Passage: As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Question: is elder scrolls online the same as skyrim
Label: False
Passage: Persian, also known by its endonym Farsi, is one of the Western Iranian languages within the Indo-Iranian branch of the Indo-European language family. It is primarily spoken in Iran, Afghanistan (officially known as Dari since 1958), and Tajikistan (officially known as Tajiki since the Soviet era), and some other regions which historically were Persianate societies and considered part of Greater Iran. It is written in the Persian alphabet, a modified variant of the Arabic script, which itself evolved from the Aramaic alphabet.
Question: do iran and afghanistan speak the same language
Label: True
Passage: [INPUT_PASSAGE]
Question: [INPUT_QUESTION]
Label:

Figure 51: Simple few-shot prompt template for BoolQ.

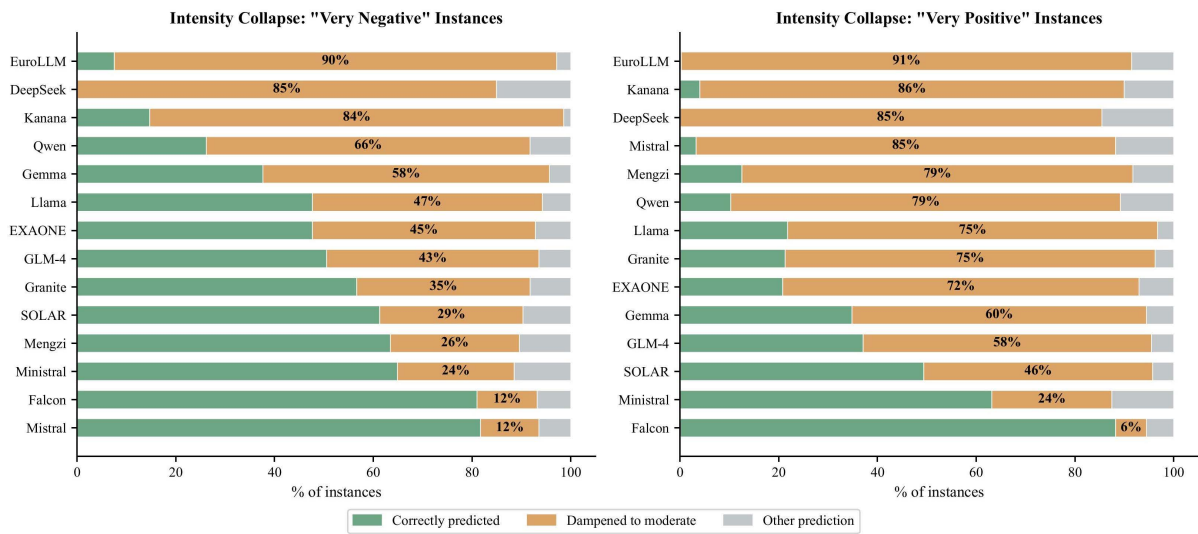


Figure 53: Intensity Collapse: Dampening Rate on SST5 Extreme Labels

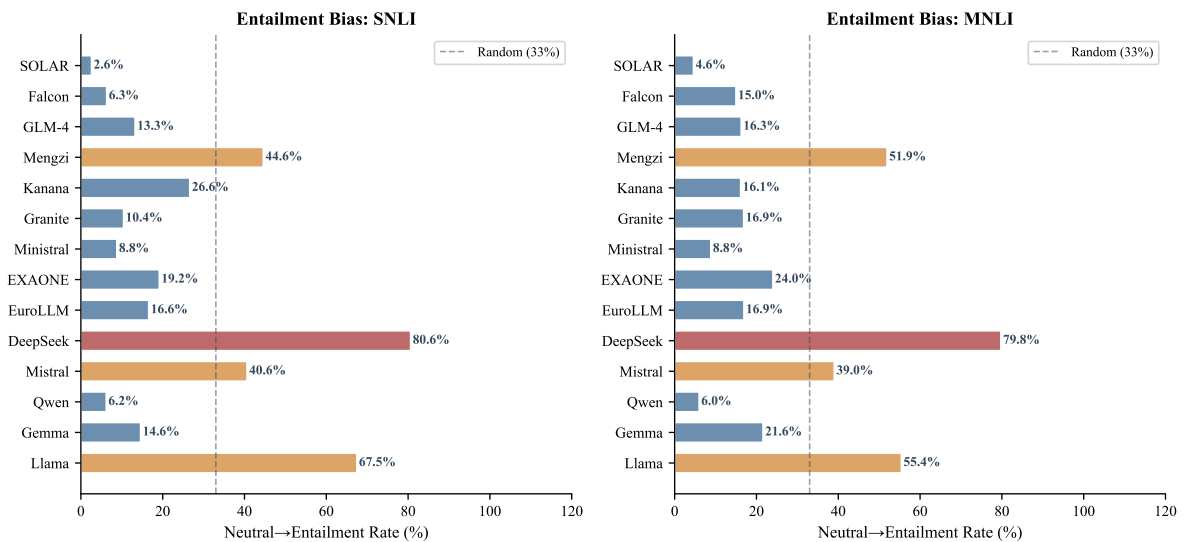


Figure 54: Entailment Bias: Neutral Examples Misclassified as Entailment

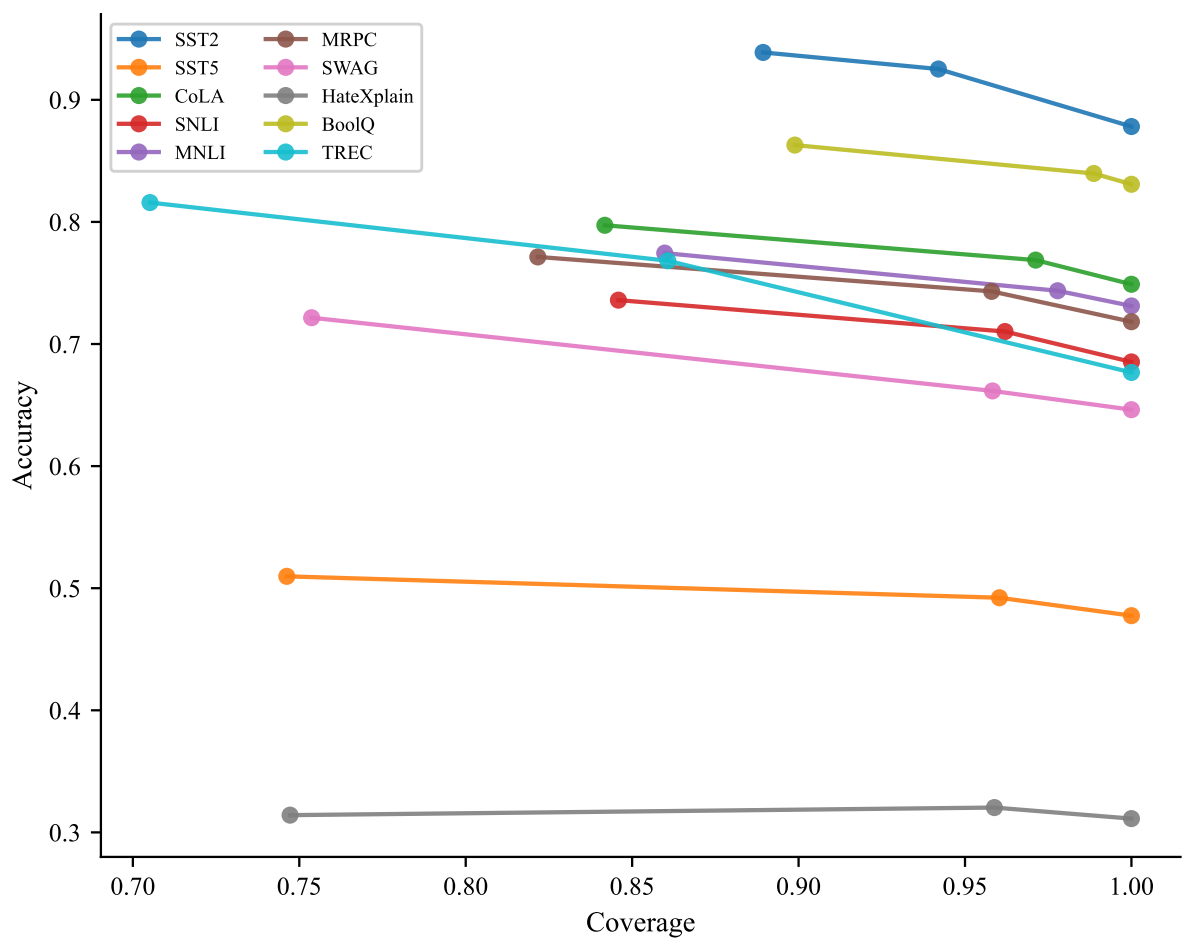


Figure 55: Accuracy-Coverage Tradeoff at SC Agreement Thresholds

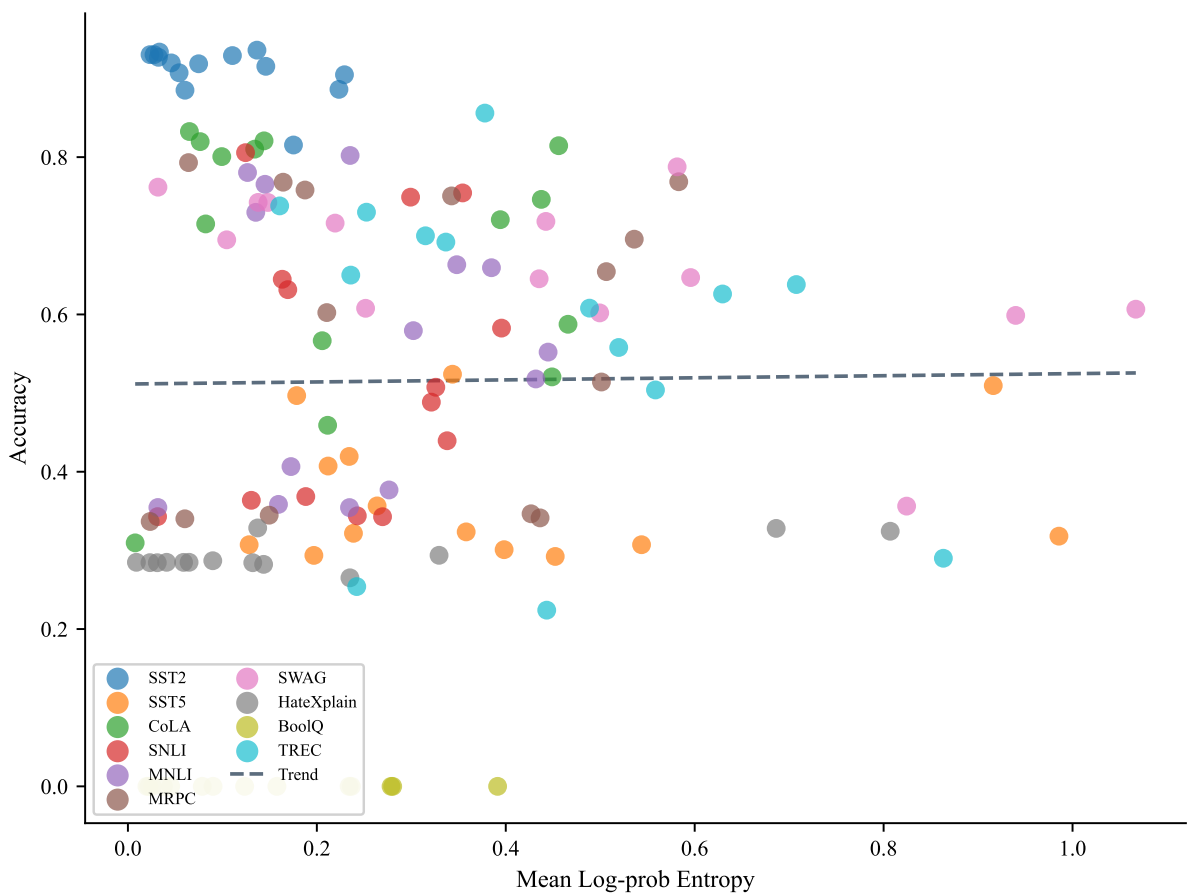


Figure 56: Entropy vs Accuracy

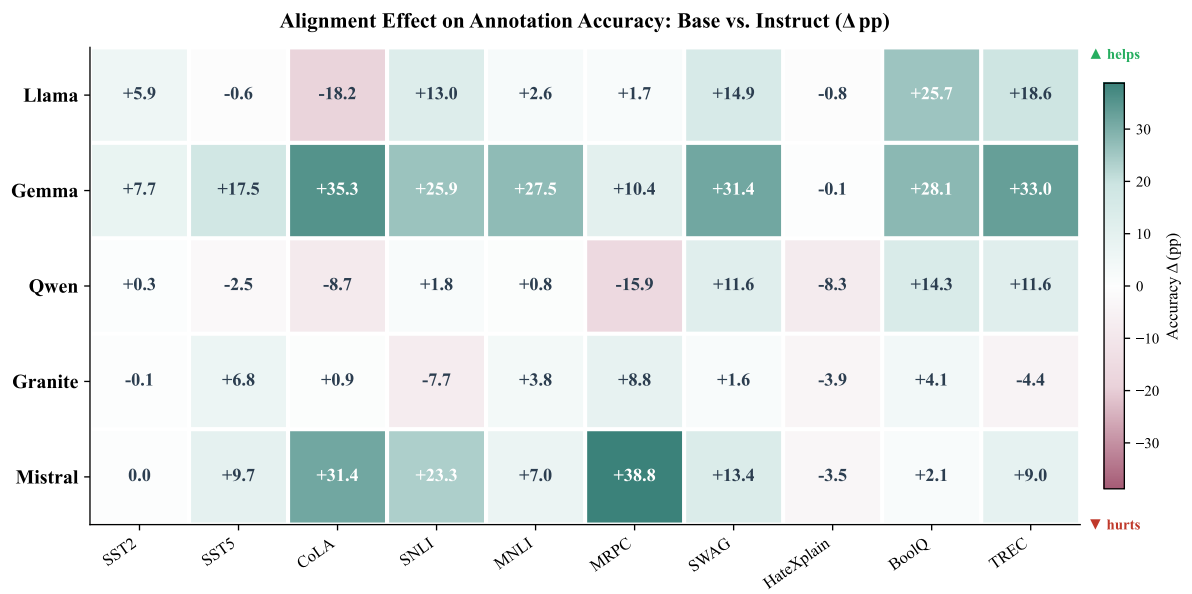


Figure 57: Alignment Effect on Annotation Accuracy



Figure 58: Failure Mode Attribution: Alignment Induced vs Base Model Failures

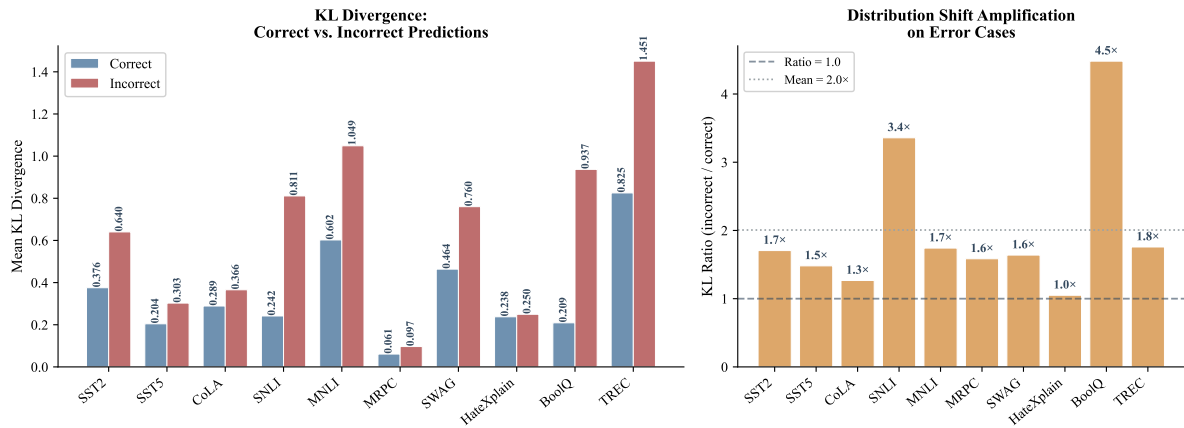


Figure 59: KL divergence: alignment-associated distribution shifts are correlated with errors. Ratios are computed per dataset and reported as an unweighted macro-average across the ten datasets, so the larger instance count of SWAG does not disproportionately drive the aggregate.

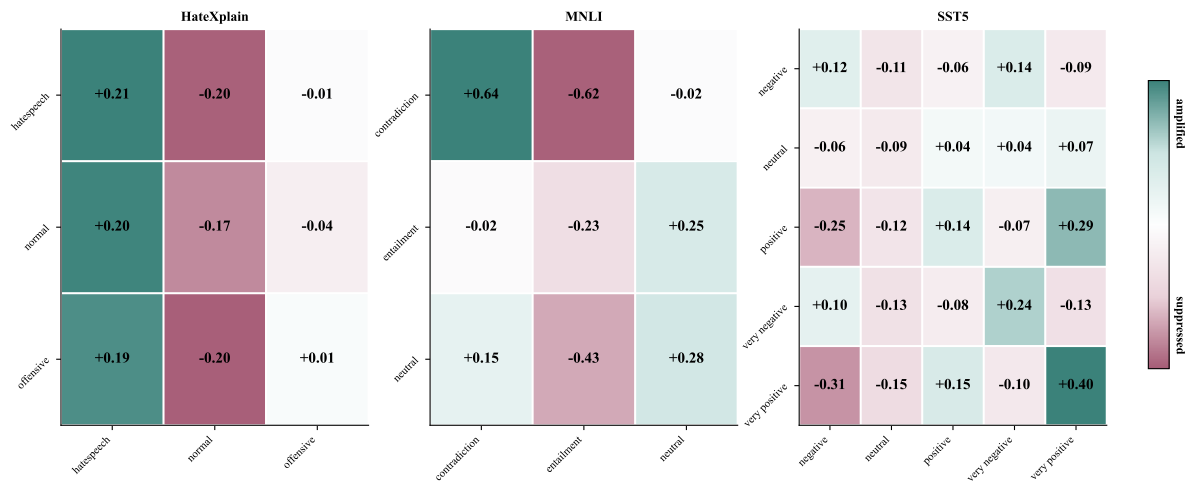


Figure 60: Label Probability Shift (LPS) of Gemma on Key Failure Mode Datasets

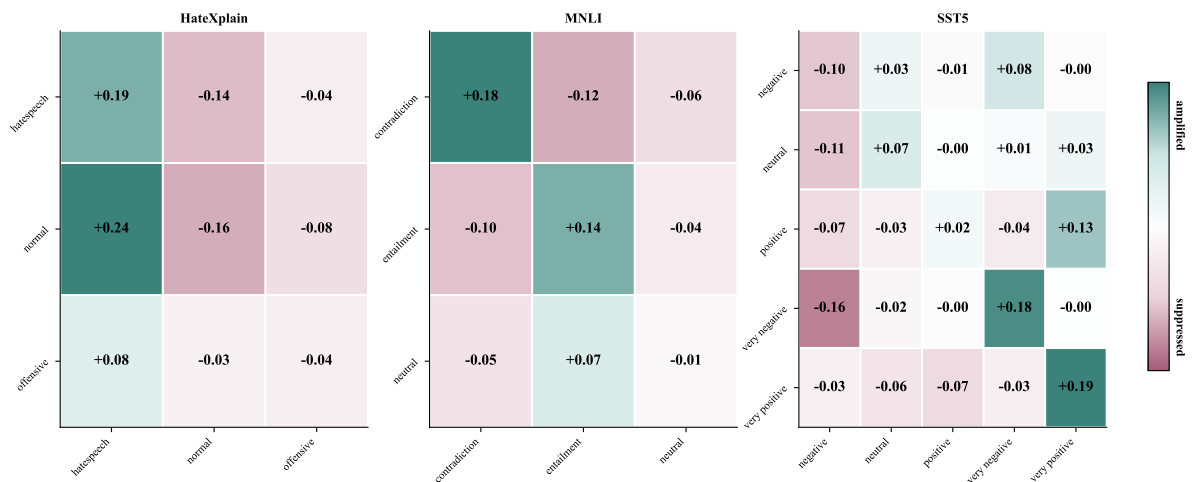


Figure 61: Label Probability Shift (LPS) of Llama on Key Failure Mode Datasets

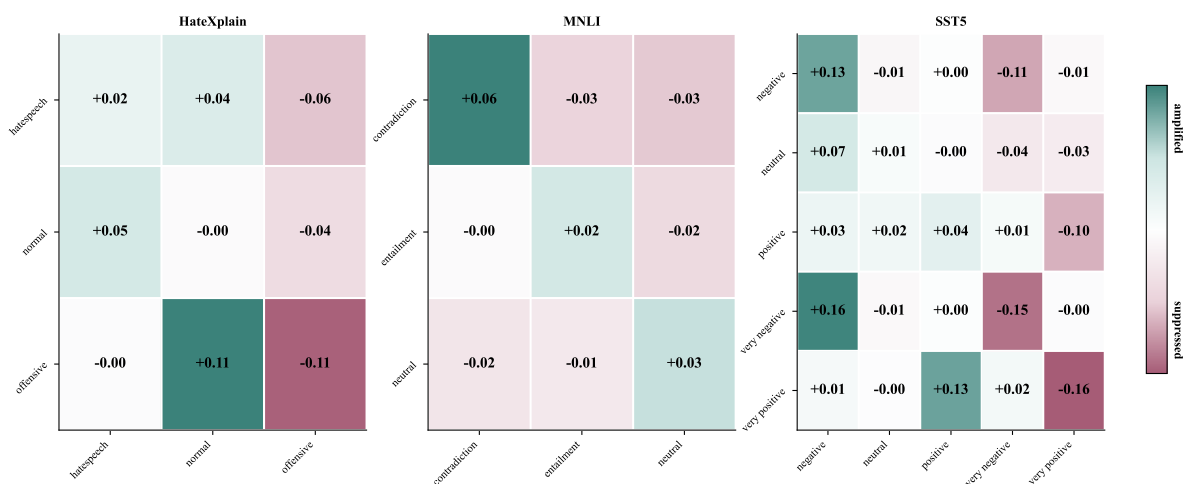


Figure 62: Label Probability Shift (LPS) of Qwen on Key Failure Mode Datasets

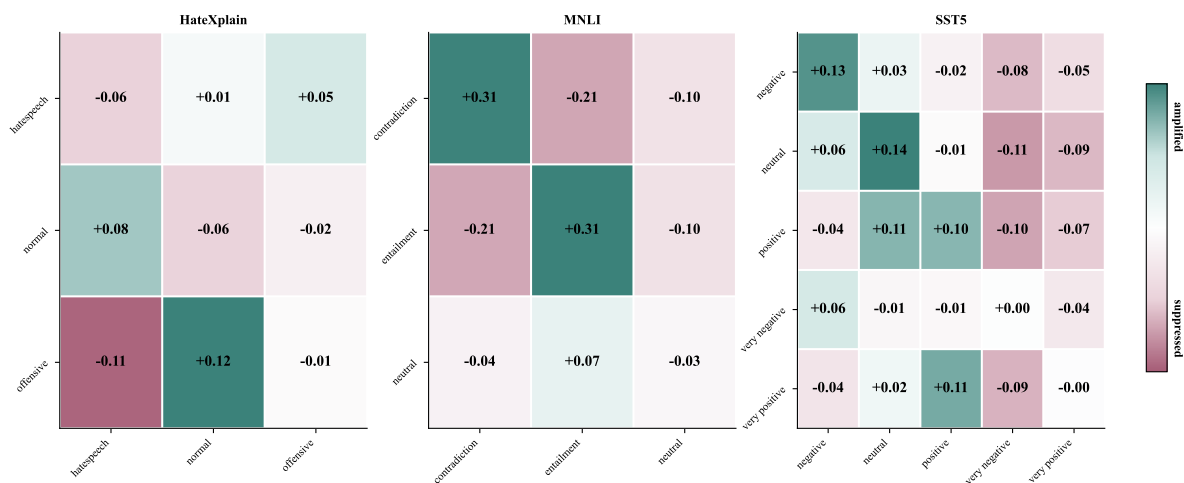


Figure 63: Label Probability Shift (LPS) of Mistral on Key Failure Mode Datasets

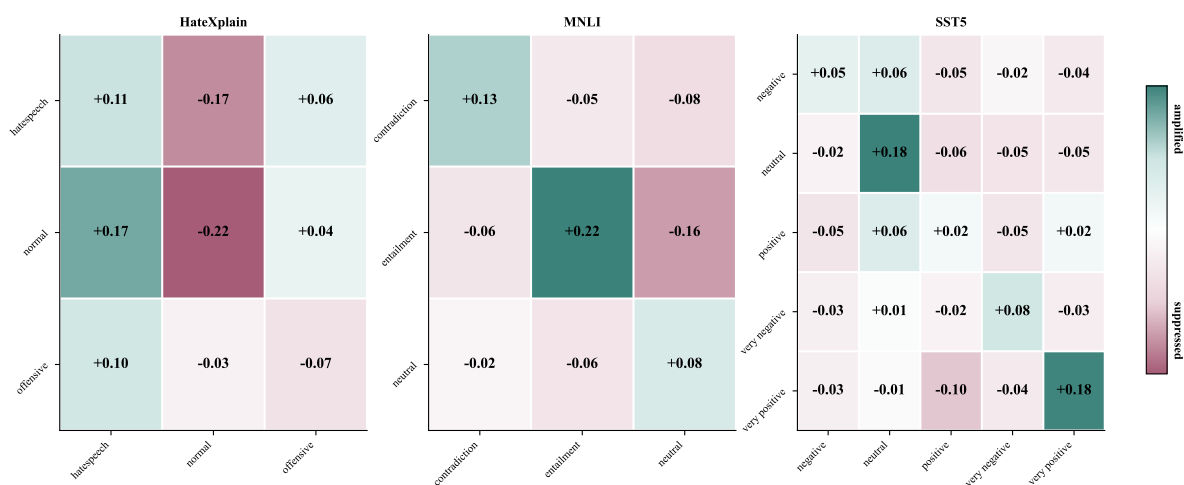


Figure 64: Label Probability Shift (LPS) of Granite on Key Failure Mode Datasets

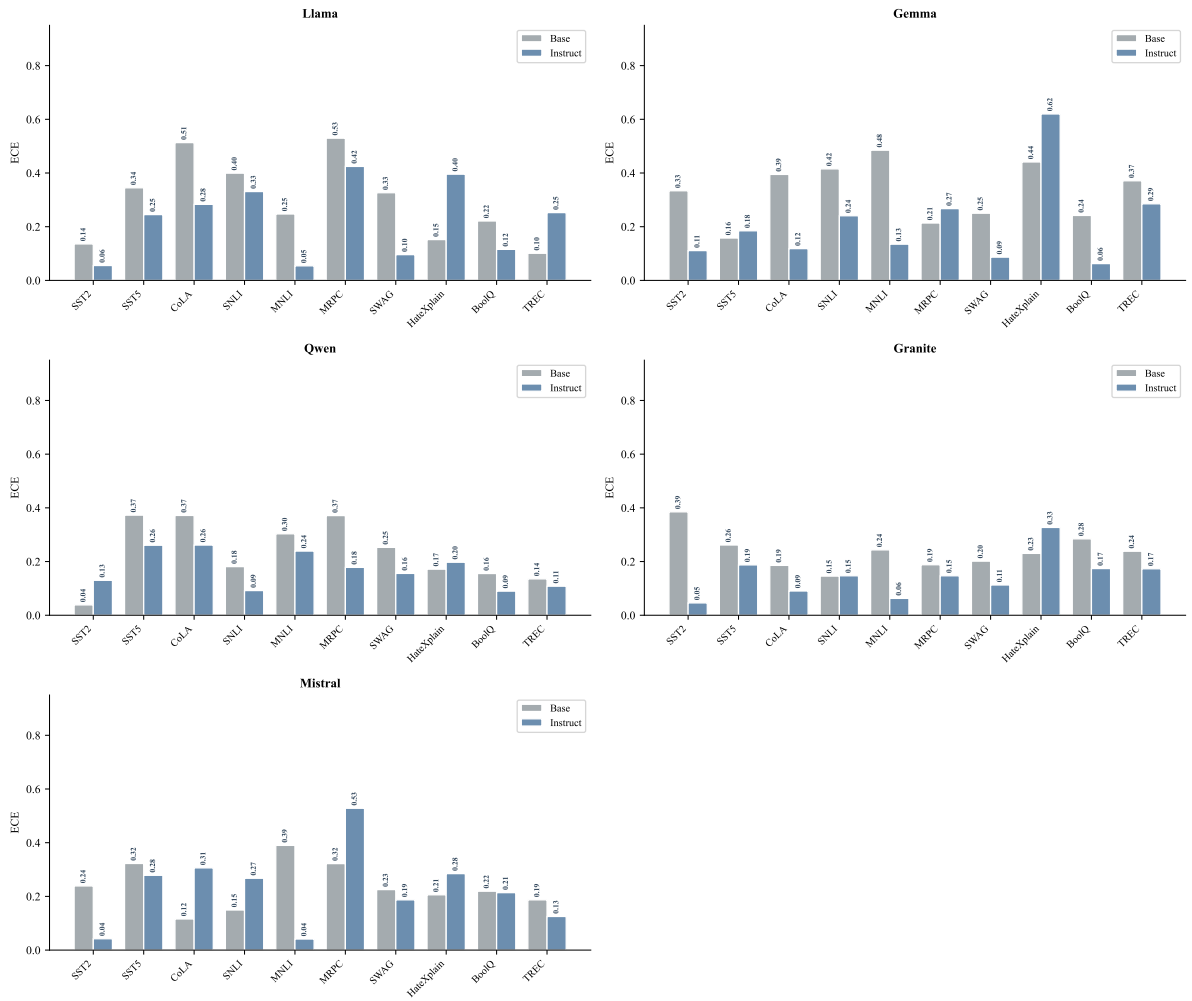


Figure 65: ECE: Base vs Instruct Models

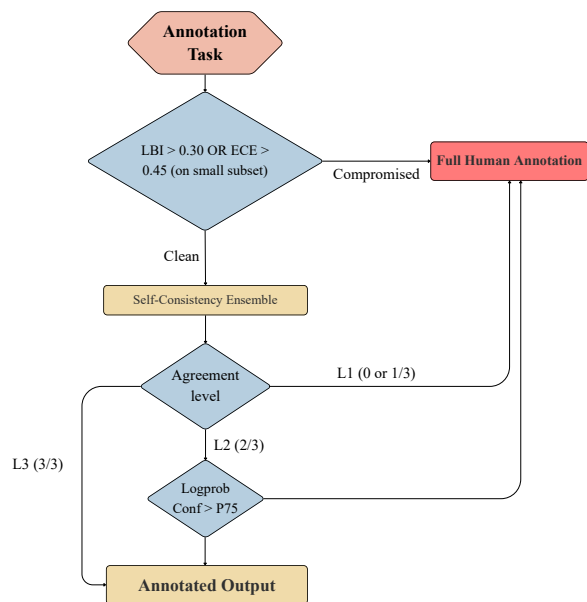


Figure 66: Ground Truth Free Annotation Pipeline