Semantic Anchors in In-Context Learning: Why Small LLMs Cannot Flip Their Labels

Anantha Padmanaban Krishna Kumar
Department of Computer Science, Boston University
anantha@bu.edu

Abstract

Can in-context learning (ICL) override pre-trained label semantics, or does it merely refine an existing semantic backbone? We address this question by treating LLMs as prompt-induced classifiers and contrasting their behavior under *natural* demonstrations (with correct labels) and *inverted* demonstrations (systematically flipping label meanings). We decompose ICL behavior into three alignment metrics (truth, prior, and prompt alignment) and introduce a semantic override rate, defined as correctness under flipped semantics. Across eight classification tasks and eight open-source LLMs (1–12B parameters), we find consistent evidence for a semantic anchor view. With natural demonstrations, ICL improves accuracy while maintaining strong prior alignment; most correct predictions coincide with zero-shot behavior, even when the prior is weak. With inverted demonstrations, models cannot learn coherent anti-semantic classifiers: prompt alignment increases only by sacrificing accuracy, and semantic override rates remain exactly zero in our few-shot 1–12B setting. Rather than flexibly remapping label meanings, ICL primarily adjusts how inputs project onto stable semantic directions learned during pre-training, clarifying fundamental limits of few-shot prompting and suggesting that overriding label semantics at these scales requires interventions beyond ICL. All code is available at: `https://github.com/AnanthaPadmanaban-KrishnaKumar/semantic-anchors-icl`.

1 Introduction

Large language models (LLMs) exhibit remarkable in-context learning (ICL): given natural-language instruction and a few input-label examples, frozen models generalize to new inputs without parameter updates [Brown et al., 2020]. This gradient-free adaptation has made ICL the standard paradigm for LLM deployment, sometimes approaching fine-tuned performance at zero training cost. Yet, the mechanism remains contested: does ICL learn new input-label mappings, or merely refine pre-trained behavior?

Two theories compete. The *task learning* view treats ICL as a general learning algorithm that is either implicit Bayesian inference [Xie et al., 2021] or gradient descent simulation [Dai et al., 2023, Akyürek et al., 2022, Von Oswald et al., 2023], and that flexibly adopts any coherent mapping from consistent demonstrations. The *prior refinement* view counters that ICL sharpens existing classifiers rather than learning de novo: random label permutations barely hurt performance [Min et al., 2022b], ICL violates Bayesian consistency [Falck et al., 2024, Kossen et al., 2023], and methods like Self-ICL successfully bootstrap from zero-shot predictions alone.

At stake is whether pre-trained *semantic anchors* can be overridden. Label tokens carry deep semantic commitments: changing `positive` to `great` can swing accuracy by tens of points [Schick and Schütze, 2021, Gao et al., 2021], and models exhibit systematic biases toward common tokens [Zhao et al., 2021, Fei et al., 2023]. Wei et al. [2023b] showed that overriding these anchors (mak-

Preprint.

ing models label positive reviews as `NEG`) requires massive scale: GPT-3 can eventually flip labels, but smaller models cannot. Whether the 1–12B models that dominate open deployments can learn anti-semantic mappings remains unclear for today's open-source families and across tasks in the few-shot regime. Unlike prior flipped-label studies focused on large proprietary models or narrow task sets, we systematically map this effect across small open-source families and quantify it with alignment metrics.

We test semantic flexibility directly. Using *natural* demonstrations (with correct labels) versus *inverted* demonstrations (systematically flipped labels), we decompose ICL into three alignments: truth (accuracy), prior (zero-shot agreement), and prompt (agreement with the demonstrated mapping). Our key metric, the *semantic override rate*, counts predictions that are correct under the inverted mapping. Across eight tasks and eight open-source LLMs (1–12B parameters across the LLaMA, Mistral, Qwen, and Gemma families), we find two consistent patterns. **Natural ICL refines priors:** accuracy improves while maintaining tight zero-shot coupling, even when priors are weak. **Inverted ICL fails to remap semantics:** models partially follow inverted demonstrations but never learn the intended anti-semantic classifier, and the semantic override rate is zero, not near-zero but exactly zero across thousands of predictions.

ICL adjusts how inputs project onto a pre-trained semantic space, but cannot redefine what labels mean. These rigid semantic anchors explain both ICL's effectiveness (when aligned with pre-training) and its fundamental limits (when opposed to it).

2 Related Work

Semantic manipulation in ICL. Prior work has tested ICL's flexibility through label manipulation. Min et al. [2022a] found that random label permutations cause only marginal accuracy drops, though this tests noise robustness rather than coherent remapping. Wei et al. [2023b] directly tested semantic override with flipped labels (positive reviews labeled `NEG`), finding a stark scale dependency: GPT-3-sized models can eventually adopt inverted mappings, but smaller models cannot. Agarwal et al. [2024] showed that even many-shot regimes (hundreds of examples) struggle with semantic override at smaller scales. Related analyses attribute flipped-label failure to demonstration shortcuts or prior dominance in ICL. We extend this inquiry to the 1–12B parameter range with explicit alignment metrics that decompose the failure modes rather than reporting accuracy alone.

Theoretical frameworks. Two mechanistic theories dominate. Bayesian accounts frame ICL as inference over latent tasks [Xie et al., 2021, Panwar et al., 2023], while meta-optimization views argue that transformers simulate gradient descent [Akyürek et al., 2022, Von Oswald et al., 2023, Dai et al., 2023]. Both often model labels as arbitrary symbols, implying that they should be remappable under consistent demonstrations. However, Falck et al. [2024] showed that ICL violates Bayesian consistency, and Kossen et al. [2023] demonstrated that "label relationships inferred from pre-training have a lasting effect that cannot be surmounted by in-context observations." These findings support a prior-constrained learning view rather than flexible remapping.

Label semantics as constraints. Labels carry strong semantic priors: switching `positive` to `great` can change accuracy by tens of points [Schick and Schütze, 2021, Gao et al., 2021, Cui et al., 2022, Mueller et al., 2022]. Zhao et al. [2021] identified systematic biases (majority-label, recency, common-token) that require explicit calibration. Holtzman et al. [2021] showed that semantically equivalent forms compete for probability mass. This evidence frames labels as semantic anchors rather than neutral symbols, a view implicitly acknowledged by methods like Self-ICL that bootstrap from zero-shot predictions. When override is necessary, approaches such as symbol tuning [Wei et al., 2023a] or contrastive decoding [Peng et al., 2025] bypass these constraints through fine-tuning or inference-time interventions, suggesting that standard ICL alone cannot overcome semantic anchors.

3 Problem Setup

We formalize in-context learning as a choice between two classifiers: a zero-shot prior and an in-context classifier induced by demonstrations.

3.1 Zero-shot and In-Context Classifiers

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label set. Each input $x \in \mathcal{X}$ has a ground-truth label $y^*(x) \in \mathcal{Y}$. Given a causal language model and a prompt P, we obtain predictions by greedy decoding.

The **zero-shot prior** $f_0(x) \in \mathcal{Y} \cup \{\text{UNK}\}$ uses only task instructions, capturing pre-trained tendencies. The **in-context classifier** $f_{\text{icl}}(x; S) \in \mathcal{Y} \cup \{\text{UNK}\}$ conditions on a demonstration set $S = \{(x_i, y_i)\}_{i=1}^k$. We evaluate all metrics on $\mathcal{D}_k = \{x \mid f_{\text{icl}}(x; S_k) \neq \text{UNK}\}$.

3.2 Natural vs. Inverted Demonstrations

We contrast two demonstration regimes. **Natural** demonstrations use correct labels: $S_{\text{nat}} = \{(x_i, y^*(x_i))\}_{i=1}^k$. **Inverted** demonstrations apply a permutation $\phi : \mathcal{Y} \to \mathcal{Y}$ to labels: $S_{\text{inv}} = \{(x_i, \phi(y^*(x_i)))\}_{i=1}^k$. For sentiment, ϕ swaps POS $\leftrightarrow$ NEG; for NLI, ϕ cycles ENTAILMENT $\to$ NEUTRAL $\to$ CONTRADICTION $\to$ ENTAILMENT.

The prompt-favored label is $y_{\text{prompt}}(x) = y^*(x)$ under natural demonstrations and $\phi(y^*(x))$ under inverted demonstrations.

3.3 Alignment Metrics

To decompose how demonstrations affect predictions, we measure three alignments:

$$\text{Accuracy}(k) = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \mathbf{1}[f_{\text{icl}}(x; S_k) = y^*(x)] \tag{1}$$

$$\text{Prior Alignment}(k) = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \mathbf{1}[f_{\text{icl}}(x; S_k) = f_0(x)] \tag{2}$$

$$\text{Prompt Alignment}(k) = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \mathbf{1}[f_{\text{icl}}(x; S_k) = y_{\text{prompt}}(x)] \tag{3}$$

Prior alignment reveals whether ICL refines or overrides zero-shot behavior. **Prompt alignment** measures agreement with demonstrated mappings. Under natural demonstrations, prompt alignment equals accuracy. Under inverted demonstrations, these metrics diverge.

Our key metric is the **semantic override rate**: the probability that predictions are both correct and consistent with inverted mappings, $P[f_{\text{icl}}(x; S_k) = y^*(x) \wedge f_{\text{icl}}(x; S_k) = y_{\text{prompt}}(x)]$ under inverted demonstrations. This captures true semantic flexibility: whether models can accurately apply anti-semantic rules.

4 Experimental Setup

4.1 Models

We evaluated eight models in four architectural families that span 1–12B parameters: LLaMA [Grattafiori et al., 2024], Mistral [Jiang et al., 2023], Qwen [Yang et al., 2025], and Gemma [Team et al., 2025]. This diverse set eliminates architectural confounds: Gemma uses different positional encodings, Qwen employs distinct tokenization, and the 12× scale range tests whether semantic anchoring weakens with capacity. The base versus instruction-tuned comparison (LLaMA 3.1 8B) isolates the effect of fine-tuning on semantic rigidity. All models are accessed via Hugging Face with identical evaluation pipelines.

Model Family	Model Name (Abbr.)
LLaMA	LLaMA-3.1-8B-Base (L8B)
	LLaMA-3.1-8B-Inst (L8I)
	LLaMA-3.2-3B-Inst (L3I)
Mistral	Mistral-7B-Inst-v0.3 (M7I)
Qwen	Qwen2.5-7B (Q7)
Gemma	Gemma-3-1B-IT (G1I)
	Gemma-3-4B-IT (G4I)
	Gemma-3-12B-IT (G12I)

Table 1: Models evaluated

4.2 Tasks and Datasets

We evaluated eight classification tasks where the label tokens have strong semantic meaning: **Sentiment Analysis** (Socher et al. 2013, SST-2; Maas et al. 2011, IMDB) with {POS, NEG}; **Natural Language Inference** (Bowman et al. 2015, SNLI; Williams et al. 2018, MNLI) with {ENTAILMENT, NEUTRAL, CONTRADICTION}; **Paraphrase Detection** (Dolan and Brockett 2005, MRPC; Quora 2017, QQP); **Hate Speech Detection** (Mollas et al. 2022, ETHOS); and **Topic Classification** (Zhang et al. 2015, AG News).

4.3 Prompting Conditions

Zero-shot	**Natural k-shot**	**Inverted k-shot**

`You are a sentiment classifier.`
`Classify the sentiment of the following review as POS or NEG.`
 	
`Review: <text>`
`Label:`
	`You are a sentiment classifier.`
	`Classify the sentiment of the following review as POS or NEG.`
	
	`Review: <`x_1`>`
	`Label: POS`
	`Review: <`x_2`>`
	`Label: NEG`
	`...`
	`Review: <text>`
	`Label:`
		`You are a sentiment classifier.`
		`Classify the sentiment of the following review as POS or NEG.`
		
		`Review: <`x_1`>`
		`Label: NEG`
		`Review: <`x_2`>`
		`Label: POS`
		`...`
		`Review: <text>`
		`Label:`

Table 2: Example prompting conditions for sentiment classification. Natural k-shot uses correct labels; inverted k-shot flips labels, shown in red. Other tasks follow the same pattern.

We evaluate three prompting conditions, illustrated in Table 2. For each test example, we sample k demonstrations from the same dataset using fixed random seeds to ensure consistent comparisons across models and conditions.

4.4 Implementation Details

We generate up to 3 tokens after the label stub using greedy decoding (temperature 0) and map the first generated word to labels using task-specific heuristics. Outputs that do not match any label are marked as UNK and excluded from metrics. We test with $k \in \{1, 2, 4, 8\}$ demonstrations, using identical demonstration sets across all models. Input length is capped at 512 tokens for zero-shot and 2,048 for few-shot prompts. All experiments use 5 random seeds (0, 1, 2, 42, 123) for demonstration sampling. We evaluate the three alignment metrics plus the semantic override rate for binary tasks.

5 Results

We analyze in-context learning behavior for LLaMA-3.1-8B-Instruct across eight classification tasks, contrasting natural demonstrations (correct labels) with inverted demonstrations (systematically flipped labels). Results for all eight models are provided in Appendix A.

5.1 Natural ICL improves accuracy, inverted ICL degrades it

Table 3 reveals a fundamental asymmetry. Natural ICL improves average accuracy from 69.5% (zero-shot) to 79.3% at $k = 8$. Gains are largest where zero-shot priors are weakest: QQP jumps 37.8 points (40.6% $\rightarrow$ 78.4%), while NLI tasks gain 14–17 points. Even strong baselines improve: sentiment classification adds 2–3 points despite starting above 90%.

Inverted ICL produces the opposite effect. Average accuracy drops to 52.8% at $k = 8$, with severe degradation on semantically rich tasks. Sentiment classification loses over 40 points; topic classification drops 32 points. The sole exception is QQP, where inverted ICL (71.6%) exceeds the weak zero-shot baseline yet still underperforms natural ICL by 7 points. This suggests that when priors are

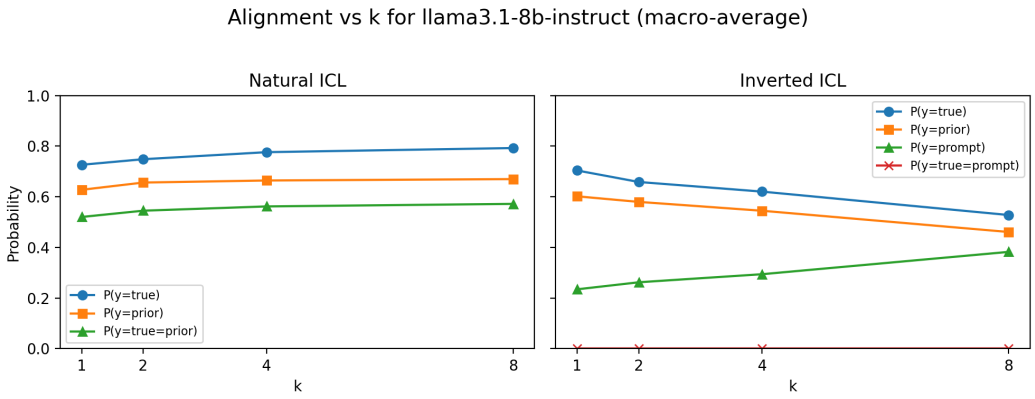

Figure 1: Macro-averaged alignment probabilities for LLaMA-3.1-8B-Instruct. Left: natural ICL increases both accuracy and joint correctness while maintaining prior alignment. Right: inverted ICL degrades accuracy and prior alignment as prompt-following increases, but joint alignment remains zero.

sufficiently weak, even corrupted demonstrations provide task-relevant signal, though never enough to override semantic anchors.

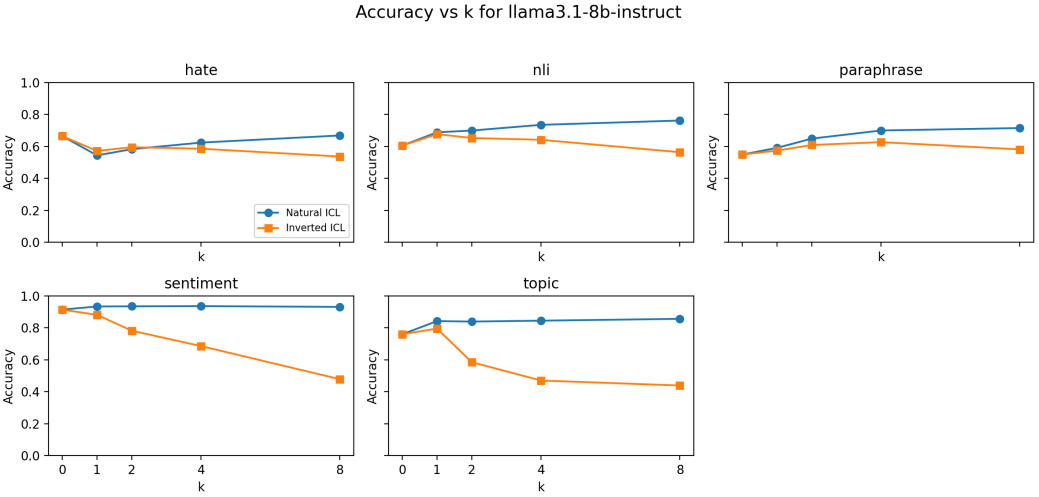

Figure 2: Accuracy vs. demonstrations k for LLaMA-3.1-8B-Instruct. Natural ICL (blue) improves performance; inverted ICL (orange) degrades systematically with more examples.

5.2 ICL operates through prior refinement, not replacement

To understand the mechanism, we decompose ICL behavior through alignment analysis (Table 4). Under natural demonstrations, k-shot predictions remain strongly coupled to zero-shot behavior. SST-2 achieves 92.5% accuracy at $k = 8$ while maintaining 89.8% alignment with zero-shot predictions: 86.4% of examples are both correct and consistent with the prior.

This coupling persists even when ICL dramatically improves accuracy. On QQP, where accuracy increases by 37 points, the model still agrees with its zero-shot prior on 34.9% of examples. Figure 1 (left) shows that as k increases, both accuracy $P(y_{\text{icl}} = y^*)$ and joint correctness $P(y_{\text{icl}} = y^* = y_0)$

Table 3: LLaMA-3.1-8B-Instruct on eight datasets. "Zero-shot" reports accuracy and unknown rate without demonstrations. "Natural" and "Inverted" report k=8 ICL accuracy. All numbers are percentages and use only non-unknown predictions for accuracy.

Task	Dataset	Zero-shot Acc.	Unknown	Natural Acc. (k=8)	Inverted Acc. (k=8)
Hate	ETHOS	66.4	0.2	66.8	53.7
NLI	MNLI	60.9	0.0	77.6	58.9
NLI	SNLI	60.2	0.0	74.8	53.9
Paraphrase	MRPC	69.2	7.1	64.6	44.7
Paraphrase	QQP	40.6	5.9	78.4	71.6
Sentiment	IMDB	92.4	12.2	93.7	48.4
Sentiment	SST-2	90.4	0.1	92.5	47.4
Topic	AG News	76.0	12.9	85.6	43.9

Table 4: Alignment statistics for LLaMA-3.1-8B-Instruct at k=8. $P_{\text{nat}}(y{=}y^{(0)})$ and $P_{\text{nat}}(y{=}y^{(0)}{=}y^*)$ are computed under natural ICL. $P_{\text{inv}}(y{=}\tilde{y})$ is computed under inverted ICL. For all datasets and k, we observe $P_{\text{inv}}(y{=}y^*{=}\tilde{y}) \approx 0$.

Task	Dataset	$P_{\text{nat}}(y{=}y^{(0)})$	$P_{\text{nat}}(y{=}y^{(0)}{=}y^*)$	$P_{\text{inv}}(y{=}\tilde{y})$
Hate	ETHOS	84.7	58.9	46.3
NLI	MNLI	64.8	54.0	27.0
NLI	SNLI	66.7	53.0	23.8
Paraphrase	MRPC	43.6	37.6	55.3
Paraphrase	QQP	34.9	26.5	28.4
Sentiment	IMDB	81.5	78.6	51.6
Sentiment	SST-2	89.8	86.4	52.6
Topic	AG News	70.0	62.7	20.8

rise together, while raw prior alignment $P(y_{\text{icl}} = y_0)$ remains stable. Natural ICL sharpens correct predictions while suppressing errors: refinement rather than replacement.

5.3 The semantic override paradox

The critical test comes with inverted demonstrations. If ICL learned arbitrary input-output mappings, consistent anti-semantic examples should induce coherent label flipping. Instead, we observe a paradox: models partially follow inverted demonstrations but never learn the intended remapping.

Table 4 quantifies this precisely. On sentiment tasks, roughly half of predictions match the demonstrated (incorrect) labels: $P(y = \tilde{y})$ reaches 51.6% on IMDB and 52.6% on SST-2. Yet the semantic override rate, $P(y = y^* = \tilde{y})$, remains exactly zero. No examples simultaneously satisfy both the true label and the inverted mapping.

Figure 1 (right) reveals the mechanism: under inversion, accuracy and prior alignment decay while prompt-following $P(y = \tilde{y})$ increases, but joint alignment $P(y = y^* = \tilde{y})$ stays at zero. The model cannot reconcile conflicting constraints (demonstrated mappings versus semantic priors), producing incoherent outputs that satisfy neither.

5.4 Effects strengthen with demonstration count

The divergence between natural and inverted ICL intensifies with more demonstrations (Figure 2). Natural ICL shows monotonic improvement or stability in all tasks as k increases from 1 to 8. Gains correlate inversely with zero-shot strength: weak priors (QQP) show steep improvement, while strong priors (SST-2) show modest gains.

Inverted ICL exhibits the opposite pattern. Sentiment tasks that barely degrade at $k = 1$ (losing 5–10 points) collapse at $k = 8$ (losing more than 40 points). This is not random variation, but a systematic conflict: each additional inverted example strengthens the contradiction between demonstrated and pre-trained semantics, forcing predictions further from both the prior and ground truth.

6 Conclusion

We asked whether few-shot prompts can override the pre-trained semantics of label tokens in small- to medium-scale LLMs. The answer is decisively no. Across eight classification tasks and eight models (1–12B parameters), inverted demonstrations that systematically flip label meanings fail to induce coherent anti-semantic classifiers. The semantic override rate, our key metric, remains exactly zero: models cannot simultaneously be accurate and follow inverted mappings. They can increase prompt-following only by sacrificing accuracy, producing incoherent outputs that satisfy neither constraint.

This failure reveals ICL's true mechanism. Natural demonstrations reliably improve performance while maintaining tight coupling to zero-shot predictions; most correct outputs coincide with the prior even when the prior is weak. ICL operates as *prior refinement*, not flexible learning. Demonstrations adjust how inputs project onto pre-existing semantic directions but cannot redefine what label tokens mean. The semantic anchors encoded in tokens like `POS`, `NEG`, and `ENTAILMENT` form rigid constraints that few-shot prompting cannot overcome.

These findings have immediate practical implications. Tasks requiring non-standard label semantics need explicit interventions such as symbol tuning, contrastive decoding, or fine-tuning rather than prompt engineering. Tasks with natural label semantics can leverage ICL effectively, because demonstrations will refine rather than fight the prior. Our alignment decomposition provides a diagnostic tool: high prior alignment indicates that ICL will succeed, and attempts to override semantics will fail.

More broadly, our results suggest a geometric interpretation: semantic labels occupy topologically stable regions in the representation manifold, locked in place by millions of pre-training observations. ICL can adjust projections within this learned geometry but cannot reshape the manifold itself, which explains why natural demonstrations succeed (moving along intrinsic gradients) while inverted demonstrations fail (pushing orthogonal to the manifold structure). Future work should test whether this constraint is specific to semantically loaded labels or extends to arbitrary symbol-concept mappings, and explore how model scale affects the flexibility of these geometric constraints. The boundaries of few-shot learning are not computational but semantic.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966, 2024.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075/`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.483. URL `https://aclanthology.org/2022.acl-long.483/`.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, 2023.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002/.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models bayesian? a martingale perspective. *arXiv preprint arXiv:2406.00793*, 2024.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.783. URL https://aclanthology.org/2023.acl-long.783/.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. *arXiv preprint arXiv:2307.12375*, 2023.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022b.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. Label semantic aware pre-training for few-shot text classification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8318–8334, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.570. URL https://aclanthology.org/2022.acl-long.570/.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*, 2023.

Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, Yancheng Yuan, and Dacheng Tao. Enhancing input-label mapping in in-context learning with contrastive decoding. *arXiv preprint arXiv:2502.13738*, 2025.

Quora. Quora question pairs. https://www.kaggle.com/competitions/quora-question-pairs, 2017. Accessed: 2025-11-19.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 255–269, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979, 2023a.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023b.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

Appendix

A Complete Experimental Results

The semantic override rate $P(y = y^* \wedge y = \tilde{y} \mid \text{inverted})$ was exactly zero across all experimental conditions. Tables 5–12 present complete results for all models across all tasks, showing the progression from $k = 1$ to $k = 8$ demonstrations. All values represent mean accuracy $\pm$ standard deviation over five random seeds.

Model abbreviations: **L8B**: LLaMA-3.1-8B-Base; **L8I**: LLaMA-3.1-8B-Instruct; **L3I**: LLaMA-3.2-3B-Instruct; **M7I**: Mistral-7B-Instruct-v0.3; **Q7**: Qwen2.5-7B; **G1I**: Gemma-3-1B-IT; **G4I**: Gemma-3-4B-IT; **G12I**: Gemma-3-12B-IT.

Table 5: SST-2 Sentiment Classification Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	90.5±0.5	82.7	89.9	91.4	91.9	78.5	83.4	66.4	37.4	0.0%
L8I	90.4±0.6	92.1	92.3	92.5	92.5	86.6	76.4	68.6	47.4	0.0%
L3I	63.8±2.9	81.6	82.1	87.9	90.7	76.1	71.7	68.0	55.8	0.0%
M7I	84.2±0.3	88.1	91.1	91.6	91.8	87.3	87.0	80.7	60.9	0.0%
Q7	92.8±0.3	93.7	94.3	94.3	94.0	92.3	78.3	64.7	52.1	0.0%
G1I	48.9±4.6	77.7	83.2	81.9	84.2	84.0	79.2	73.5	66.9	0.0%
G4I	94.1±2.9	87.2	89.3	90.2	90.7	87.0	87.9	84.8	81.6	0.0%
G12I	89.6±0.4	90.2	91.5	92.1	93.1	90.1	85.0	73.0	63.0	0.0%

Table 6: IMDB Sentiment Classification Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	77.2±0.9	75.3	94.2	93.2	92.2	75.0	89.0	73.7	41.1	0.0%
L8I	92.4±1.1	94.7	94.6	94.7	93.7	89.8	80.1	68.6	48.4	0.0%
L3I	64.6±1.3	91.9	91.9	92.4	90.8	91.9	84.6	79.2	52.0	0.0%
M7I	93.4±0.7	91.2	93.2	90.9	76.0	92.4	85.1	70.4	37.0	0.0%
Q7	94.1±0.8	94.4	94.6	94.4	94.0	92.2	80.3	66.1	65.9	0.0%
G1I	48.9±1.0	82.1	79.8	81.1	83.2	77.0	66.8	64.0	59.2	0.0%
G4I	95.9±1.1	90.4	91.6	92.4	92.2	92.4	91.8	90.4	81.1	0.0%
G12I	93.6±0.6	94.7	94.7	94.8	95.0	93.0	89.7	72.3	57.4	0.0%

Table 7: SNLI Natural Language Inference Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	55.2±2.0	56.2	56.7	65.8	70.4	55.4	53.4	60.3	54.8	0.0%
L8I	60.2±1.1	66.8	67.5	72.1	74.8	65.8	62.8	63.2	53.9	0.0%
L3I	43.5±3.2	55.2	52.9	59.0	63.6	52.2	51.5	50.2	45.3	0.0%
M7I	58.9±2.7	68.0	70.4	71.8	73.4	67.1	67.6	62.7	53.5	0.0%
Q7	84.7±0.9	86.5	86.9	86.6	87.4	86.9	84.4	78.1	72.5	0.0%
G1I	38.0±3.6	43.4	41.8	41.6	43.3	42.2	40.4	38.0	37.1	0.0%
G4I	57.3±2.1	70.7	73.6	74.9	75.8	67.2	67.8	65.3	57.5	0.0%
G12I	69.3±2.2	71.7	75.4	78.1	82.0	71.2	72.9	68.6	61.7	0.0%

Table 8: MNLI Natural Language Inference Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	55.0±1.6	54.8	60.7	67.0	71.3	55.4	56.2	60.7	57.4	0.0%
L8I	60.9±1.5	70.8	72.3	74.9	77.6	69.5	67.8	65.0	58.9	0.0%
L3I	56.8±1.6	58.2	58.1	60.6	62.0	58.1	55.9	55.2	48.5	0.0%
M7I	72.5±1.2	69.3	72.4	74.6	76.2	67.2	67.1	63.5	54.7	0.0%
Q7	84.5±0.9	85.1	85.1	85.3	86.6	83.1	80.3	75.3	70.5	0.0%
G1I	44.3±1.8	43.6	45.2	42.8	42.6	44.0	43.6	40.6	39.6	0.0%
G4I	69.3±1.7	71.9	71.4	70.5	70.6	70.3	68.2	65.4	61.6	0.0%
G12I	80.1±1.1	80.8	81.9	83.1	83.9	79.9	80.7	78.1	69.7	0.0%

Table 9: MRPC Paraphrase Detection Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	66.0±1.0	69.8	70.2	68.6	71.1	68.8	64.7	60.8	54.4	0.0%
L8I	69.2±1.8	46.5	55.0	62.8	64.6	47.2	49.9	51.0	44.7	0.0%
L3I	67.6±0.9	33.6	33.7	36.1	42.9	33.5	33.5	34.5	34.8	0.0%
M7I	75.9±2.8	64.7	68.5	65.7	65.8	69.9	62.7	51.8	40.4	0.0%
Q7	75.2±0.3	76.2	75.8	74.2	74.5	75.8	74.9	71.5	69.9	0.0%
G1I	63.9±3.2	60.3	63.8	65.3	65.7	48.5	51.1	54.3	49.5	0.0%
G4I	72.4±1.2	71.5	72.2	72.3	72.5	68.6	69.1	64.3	59.3	0.0%
G12I	83.6±1.1	78.6	78.1	78.0	78.0	77.1	76.7	75.4	72.2	0.0%

Table 10: QQP Paraphrase Detection Results. Note the anomalously weak zero-shot performance across all models.

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	40.3±0.8	55.4	69.6	73.2	77.3	61.8	61.9	66.0	60.1	0.0%
L8I	40.6±1.7	71.7	74.7	77.1	78.4	67.6	71.8	74.3	71.6	0.0%
L3I	42.4±1.1	64.6	65.3	67.2	69.3	65.2	64.1	65.5	65.0	0.0%
M7I	84.2±0.4	85.3	88.2	86.3	85.8	85.3	86.8	81.4	77.1	0.0%
Q7	80.7±0.9	78.8	81.0	81.3	82.5	81.6	80.6	75.7	68.0	0.0%
G1I	24.7±15.0	68.9	66.0	62.9	65.6	68.7	61.3	58.4	55.3	0.0%
G4I	65.2±1.3	78.2	78.4	78.2	80.0	76.1	78.3	77.9	74.7	0.0%
G12I	85.0±1.6	80.8	80.8	81.1	81.4	79.4	79.8	74.8	62.7	0.0%

Table 11: ETHOS Hate Speech Detection Results

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	40.5±0.1	45.5	52.3	69.1	75.2	63.6	62.9	50.1	34.6	0.0%
L8I	66.4±0.2	54.4	58.3	62.4	66.8	57.2	59.5	58.6	53.7	0.0%
L3I	41.0±0.4	48.0	52.8	62.1	68.4	53.8	58.9	60.7	59.4	0.0%
M7I	67.8±0.1	64.7	65.4	68.4	72.4	63.7	61.7	59.4	56.2	0.0%
Q7	41.3±0.1	53.2	61.5	71.3	78.1	48.0	49.6	45.6	36.4	0.0%
G1I	30.2±0.2	42.5	55.0	65.9	73.0	71.0	53.1	45.8	35.9	0.0%
G4I	59.9±0.3	60.6	60.6	64.5	69.1	62.1	56.3	57.7	57.3	0.0%
G12I	62.8±0.2	67.5	69.3	71.8	75.1	65.4	64.2	59.6	43.1	0.0%

Table 12: AG News Topic Classification Results. Uses 4-way cyclic label permutation rather than binary inversion.

Model	Zero-shot	Natural ICL (%)				Inverted ICL (%)				Override
	Acc (%)	k=1	k=2	k=4	k=8	k=1	k=2	k=4	k=8	Rate
L8B	80.9±1.7	88.2	86.2	87.6	87.9	74.9	54.1	50.3	38.0	0.0%
L8I	76.0±1.3	84.2	83.9	84.5	85.6	79.6	58.5	47.1	43.9	0.0%
L3I	84.8±3.0	80.3	79.4	80.6	82.5	74.9	64.3	68.5	69.9	0.0%
M7I	84.7±0.8	86.6	85.6	84.1	85.3	84.7	80.8	64.2	48.8	0.0%
Q7	82.8±1.2	82.6	82.6	84.5	86.1	78.7	66.8	57.3	44.8	0.0%
G1I	47.1±2.3	74.6	68.4	74.6	77.7	74.1	68.9	72.6	68.4	0.0%
G4I	83.3±1.3	80.6	80.7	82.9	85.0	79.6	79.0	74.4	66.2	0.0%
G12I	84.7±1.1	85.1	85.2	86.2	86.7	81.3	72.2	59.6	46.9	0.0%

A.1 Key Observations

Several patterns emerge consistently across all models and tasks. **(1) Universal Zero Override:** The semantic override rate remains exactly 0.0% across all 320 experimental conditions (8 models $\times$ 8 tasks $\times$ 5 seeds), confirming that no model successfully learned to apply inverted label mappings. **(2) Monotonic Degradation:** Under inverted demonstrations, accuracy degrades monotonically as k increases from 1 to 8, with the strongest effects on sentiment tasks (often dropping more than 40 points at $k = 8$). **(3) QQP Anomaly:** The QQP dataset exhibits unusually weak zero-shot performance across all models (24.7%–85.0%), yet models still cannot override label semantics despite the weak prior. **(4) Scale Independence:** The pattern holds consistently across the twelvefold parameter range (1B to 12B), with no evidence that larger models within this range can overcome semantic anchors.

B Prompt Templates and Demonstration Structure

B.1 Implementation Parameters

All experiments used identical hyperparameters for reproducibility. Greedy decoding (temperature = 0) was used with a maximum of three new tokens generated per prediction. The input sequences were truncated to 512 tokens for zero-shot prompts and 2,048 tokens for k-shot prompts, with left-sided padding and truncation. All experiments were carried out with five random seeds $\{0, 1, 2, 42, 123\}$ to ensure statistical validity. The demonstrations for each test example were randomly sampled without replacement from the training set, excluding the query example itself.

B.2 Complete Prompt Templates

Table 13: Exact prompt templates used for all tasks. Identical instructions are used for both natural and inverted conditions.

Task	Template	Inversion

Sentiment	`You are a sentiment classifier.`
	`Classify the sentiment of the following`
	`review as POS or NEG.`
	
	`Review: [TEXT]`
	`Label:`
		POS $\leftrightarrow$ NEG

NLI	`You are a natural language inference (NLI)`
	`classifier. Given a premise and a`
	`hypothesis,`
	`decide whether the relationship is`
	`ENTAILMENT,`
	`NEUTRAL, or CONTRADICTION.`
	
	`Premise: [TEXT1]`
	`Hypothesis: [TEXT2]`
	`Label:`
		Cyclic: $(y+1) \bmod 3$
		ENT$\rightarrow$NEU$\rightarrow$CON$\rightarrow$ENT

Paraphrase	`You are a paraphrase detector.`
	`Determine if these two sentences are`
	`SIMILAR or DIFFERENT in meaning.`
	
	`Sentence 1: [TEXT1]`
	`Sentence 2: [TEXT2]`
	`Label:`
		SIMILAR $\leftrightarrow$ DIFFERENT

Hate	`You are a hate speech detector.`
	`Classify whether the following text contains`
	`hate speech as HATE or NOT_HATE.`
	
	`Text: [TEXT]`
	`Label:`
		HATE $\leftrightarrow$ NOT_HATE

Topic	`You are a news topic classifier.`
	`Classify this article into one of these`
	`categories: WORLD, SPORTS, BUSINESS,`
	`or TECHNOLOGY.`
	
	`Article: [TEXT]`
	`Topic:`
		Cyclic: $(y+1) \bmod 4$
		W$\rightarrow$S$\rightarrow$B$\rightarrow$T$\rightarrow$W

B.3 Demonstration Example

Table 14: Example of natural vs inverted demonstrations for sentiment classification ($k = 2$)

Natural Demonstrations	Inverted Demonstrations
`Review: Amazing film!` `Label: POS`	`Review: Amazing film!` `Label: NEG` ← flipped
`Review: Waste of time.` `Label: NEG`	`Review: Waste of time.` `Label: POS` ← flipped
`Review: [QUERY]` `Label:`	`Review: [QUERY]` `Label:`