

Do Language Models Prefer Better Demonstrations or Easier Output Tokens?

Jiaye Zhang*

School of Information Science and Technology, Shanghai Tech University, Shanghai, China

**Corresponding author: Jiaye Zhang.*

Abstract

Model feedback is often used to choose in-context demonstrations, but a high probability for the correct answer token may reflect token compatibility rather than useful example content. We audit this issue with Qwen2.5-0.5B-Instruct on BoolQ using corrected inputs, frozen samples, and query-level inference. Under the original Yes/No labels, the preferred same-label demonstration improves the held-out margin over a frozen random baseline by 0.2526. Late-layer residual interventions, however, show no reliable advantage for a matched donor over an unrelated donor with the same label. Counterbalanced A/B mappings then make the raw preferred-minus-random effect follow the output token. A new frozen extension compares raw correct-token likelihood, a binary logit margin, and content-free calibration. The raw selector is harmful on average and strongly favors B. Binary-margin selection gives the largest average held-out gain but leaves a smaller bias toward A. Content-free calibration reduces the raw token gap but does not meet the preset equivalence rule. These results show that model-based demonstration preference depends on the full prompt, label vocabulary, and scoring rule. It should not be read as a direct measure of semantic example quality.

Keywords

in-context learning, demonstration selection, label bias, calibration, activation patching

1. Introduction

In-context learning (ICL) enables a language model to perform a task from examples in the prompt without weight updates [1]. However, the selection and order of these demonstrations can change model confidence and accuracy [2, 3]. This sensitivity has motivated semantic retrieval, learned retrieval, and model-feedback methods that seek better demonstrations for individual queries [4–6].

A common approach scores a candidate demonstration by the probability that the model assigns to the known correct label after conditioning on that demonstration. Although intuitive, this score conflates several factors. It may increase because the demonstration helps answer the query, makes a particular output token easier to generate, or better matches the prompt wording. Calibration studies and work on surface-form competition show that such effects can be substantial [7, 16, 17]. Recent work also shows that a token-based decision rule may be a poor proxy for task-level performance [19].

This study asks a narrow question: what does model-based demonstration preference actually measure? We first test whether a preferred same-label demonstration outperforms ordinary same-label baselines. We then

test whether its internal effect is specific to the matched query or shared across queries with the same answer. Finally, we reverse the mapping between task meanings and output tokens and compare selection scores to determine whether preference effects follow semantic meaning or output tokens.

This study makes three contributions. First, it separates a genuine behavioral gain from the stronger claim of semantic matching. Second, it uses same-label, opposite-label, shuffled-content, and norm-matched controls to constrain the interpretation of residual interventions. Third, it shows that changing only the selection score can reverse both average utility and output-token preference. The aim is not to propose a deployable selector, but to audit a widely used measurement choice.

The analysis is organized around four research questions. **RQ1** asks whether a preferred demonstration provides a stable held-out behavioral gain over frozen same-label baselines. **RQ2** asks whether the residual intervention effect is specific to the matched query rather than shared across queries with the same label. **RQ3** asks whether the preference effect follows task meaning or the output token under counterbalanced A/B verbalizers. **RQ4** asks whether a selector can improve held-out utility while removing dependence on the output token.

2. Related Work

2.1 In-Context Demonstration Selection and Calibration

Semantic retrieval, learned retrievers, and language-model feedback can all improve few-shot prompts [4–6]. These methods primarily ask which demonstrations work better. Our question is different: whether a model-preference score measures useful content or merely an easier output token. Contextual calibration removes content-free answer priors [7], while label-mapping studies show that arbitrary labels compete with semantic knowledge [8, 17]. Label words can also serve as information anchors inside the model [9]. In-context fixation provides a recent example of demonstrated labels overriding task semantics [21].

2.2 Mechanistic Interpretability and Causal Interventions in LLMs

Task-vector, function-vector, and in-context-head studies show that demonstrations can create internal states that transfer across inputs [10, 11, 18]. Activation patching and causal mediation can test whether such a state changes an output [12, 13]. A large intervention effect, however, does not by itself identify what the state represents. Best-practice work stresses the need for careful metrics and controls [20], and recent evidence suggests that output templates can be distributed rather than stored at a single position [22]. We therefore compare matched donors with unrelated donors that share or oppose the answer label.

3. Methodology and Causal Auditing Framework

3.1 Problem Formulation and Notation

Each BoolQ instance is a binary query $q = (p, u)$, where p is the passage, u is the question, and $y \in \{\text{No}, \text{Yes}\}$ is the gold semantic label. A candidate demonstration $d = (q_d, y_d)$ contains its own BoolQ item and answer and is eligible for q only when $y_d = y$. A verbalizer mapping m maps each semantic label to a single output token; $t_y^m = m(y)$ and $t_{\bar{y}}^m = m(\bar{y})$ denote the correct and competing output tokens.

Given selection template T_S , the selector function $s(d, q, y; m)$ assigns a scalar score to every candidate d in the frozen pool $\mathcal{D}_q$ and returns the highest-scoring demonstration d^* . Evaluation is deliberately separated from selection: $M_E(d, q, y; m)$ is the correct-versus-competing logit margin under held-out template T_E . Thus, s determines the ranking, whereas M_E is the common outcome used for preferred-versus-baseline comparisons.

3.2 Prompt Templates

We use separate prompt templates for selection and evaluation to avoid selecting and evaluating a demonstration with identical prompt wording. The selection template ranks candidates, whereas the held-out evaluation template changes the wording and measures whether the selected demonstration transfers.

The selection template uses the form

Example: {demo_text};
 Answer: {demo_label};
 Question to answer: {query_text};
 Answer:
 The held-out evaluation template uses
 Demonstration: {demo_text};
 Answer: {demo_label};
 BoolQ item: {query_text};
 Answer:

The stored prompts preserve the corresponding field boundaries and line breaks.

A candidate may be favored because it is unusually compatible with the selection-template wording. A gain that persists under the held-out evaluation template is less likely to be a template-specific artifact.

3.3 Selector Scoring Rules

The three scoring rules below instantiate the selector function s , while all reported held-out effects use the evaluation metric M_E defined above.

Immediately after the final “Answer:”, $z(t \mid T)$ denotes the next-token logit for token t under template T . Raw selection uses the correct token's log probability over the full vocabulary. We define three selector functions:

$$s_{\text{raw}}(d, q, y; m) = \log p(t_y^m \mid T_S(d, q; m)), \quad (1a)$$

$$s_{\text{margin}}(d, q, y; m) = z(t_y^m \mid T_S) - z(t_y^m \mid T_S), \quad (1b)$$

$$s_{\text{cal}}(d, q, y; m) = s_{\text{margin}}(d, q, y; m) - s_{\text{margin}}(d, q_{cf}, y; m). \quad (1c)$$

where T_S denotes the selection template. The fixed content-free query q_{cf} is defined as:

Passage: N/A

Question: N/A?

The raw score represents conventional model-based preference. The binary-margin score isolates the competition between the correct and competing output tokens. The content-free calibrated score subtracts the same demonstration's preference under a query with no task content. Comparing these selectors reveals how demonstration rankings depend on the scoring rule, output token, and prompt-level prior.

Evaluation uses the following held-out correct-versus-competing binary logit margin:

$$M_E(d, q, y; m) = z(t_y^m \mid T_E) - z(t_y^m \mid T_E). \quad (2)$$

$\mathcal{D}_q$ is the fixed same-label candidate set, and $\mathcal{R}_q^n$ contains the n candidates drawn and frozen before model scoring. For selector s , the preferred candidate and its query-level excess over the frozen random candidates are:

$$d_{s,q,m}^* = \arg \max_{d \in \mathcal{D}_q} s(d, q, y; m); \quad (3)$$

$$e_{s,q,m} = M_E(d_{s,q,m}^*, q) - \frac{1}{n} \sum_{d \in \mathcal{R}_q} M_E(d, q). \quad (4)$$

3.4 Counterbalanced Label Mapping and Token-Effect Contrasts

For the counterbalanced cells $e(\text{semantic class}, \text{token})$, we define the following frozen contrasts for the binary task:

$$C_{\text{sem}} = \frac{e(\text{Yes}, A) + e(\text{Yes}, B)}{2} - \frac{e(\text{No}, A) + e(\text{No}, B)}{2}, \quad (5)$$

$$C_{\text{tok}} = \frac{e(\text{Yes}, A) + e(\text{No}, A)}{2} - \frac{e(\text{Yes}, B) + e(\text{No}, B)}{2}. \quad (6)$$

These contrasts measure whether preference effects follow semantic class or output token, respectively.

Thus, C_{tok} is the $A - B$ contrast. The selector extension uses the opposite $B - A$ sign convention:

$$G_s = E[e_s|B] - E[e_s|A] = -C_{\text{tok}} \quad (7)$$

3.5 Layer-Wise Residual Intervention

Donor k at layer l supplies the preferred-minus-worst residual difference:

$$\Delta h_{k,l} = h_{k,l}^{\text{preferred}} - h_{k,l}^{\text{worst}}. \quad (8)$$

where $h_{k,l}$ denotes the donor residual state at the final prompt token under the indicated demonstration condition.

For receiver r , let w_r and p_r denote its worst and preferred runs, and let $M_r(\cdot) = \text{logit}(\text{correct}) - \text{logit}(\text{competing})$ denote its correct-versus-competing margin after intervention. The forward and reverse margin changes and their bidirectional average are:

$$F_{r,k,l} = M_r(w_r + \Delta h_{k,l}) - M_r(w_r); \quad (9)$$

$$R_{r,k,l} = M_r(p_r) - M_r(p_r - \Delta h_{k,l}). \quad (10)$$

$$I_{r,k,l} = \frac{F_{r,k,l} + R_{r,k,l}}{2}. \quad (11)$$

A positive value of I means that the residual direction from donor k increases the receiver's correct-versus-competing margin.

Let S_r denote the receiver's cross-query same-label donors. The receiver-level matched excess is:

$$X_r = E_I[I_{r,r,l}] - E_{k \in S_r}[I_{r,k,l}]. \quad (12)$$

where $I_{r,r,l}$ is the mean intervention effect across the three tested layers when the receiver uses its matched donor; $I_{r,k,l}$ is averaged across the two unrelated same-label donors and the three tested layers.

We also conduct a projected analysis. After estimating a unit label direction $\hat{v}_{\text{label}}$, we replace each donor difference by $\Delta h_{\perp} = \Delta h - (\Delta h^{\top} \hat{v}_{\text{label}}) \hat{v}_{\text{label}}$, before computing the receiver-level matched excess.

4. Experimental Setup

4.1 Model Checkpoint and Dataset

All experiments use the same fixed local Qwen2.5-0.5B-Instruct weights and tokenizer configuration [15]. Queries come from the corrected BoolQ validation split, and candidate demonstrations come from the corrected training split [14]. Preprocessing preserves each original passage and question, retains the binary gold label, and renders valid items with fixed prompt field boundaries. Development-only and formal samples are disjoint, and development outputs are excluded from formal inference.

4.2 Candidate Pool Construction and Sample Freezing

Each formal query has a materialized pool of 12 training demonstrations with the same gold semantic label. Candidate pools and four-candidate random subsets are constructed by deterministic SHA-256 ordering and frozen before any model score is observed. Paired mappings and selector conditions reuse the same candidate pool and random subset; reserved direction-estimation items and all earlier-stage items are excluded. Frozen

candidate order resolves ties, while source hashes, partitions, seeds, prompts, model outputs, and audit records are stored with the project. Table I summarizes the formal analysis units and sample sizes.

Table 1: Experiment overview

Stage	Formal unit	Size	Main question
Behavior	Query	256	Does preferred beat ordinary same-label examples?
Mechanism	Receiver	64	Is residual transfer query-specific?
A/B mapping	Query	256	Does heterogeneity follow meaning or token?
Selector extension	Query	256	Can a selector improve utility and reduce token dependence?

4.3 Statistical Inference and Query-Level Bootstrap Protocol

Behavioral, mapping, and selector analyses use 50,000 query-bootstrap draws. Paired mappings and selectors remain paired, while No and Yes strata are resampled separately and weighted equally. RQ3 uses Bonferroni simultaneous 95% intervals for its two confirmatory contrasts; RQ4 uses paired 95% intervals and a prespecified 90% equivalence interval of $[-0.20, 0.20]$. Mechanism intervals use a receiver-level bootstrap after donors and layers are aggregated within each receiver. Candidate, mapping, donor, layer, and intervention cells are never treated as independent observations.

5. Empirical Results and Analysis

5.1 RQ1: Held-Out Behavioral Utility of Preferred Demonstrations

RQ1 compares the raw preferred candidate with the numerical-median candidate, the worst candidate, a TF-IDF lexical nearest neighbor, and the frozen random subset defined in Section IV. The raw selector ranks candidates under the selection template, and every condition is evaluated under the differently worded held-out template.

The **RQ1** outcome is M_E . With the raw selector and the original Yes/No verbalizer, the preferred effect is $e_{s,q,m}$. The random-four mean is a Monte Carlo estimate of a uniformly chosen random single demonstration. Candidate-level accuracy is $1[M_E > 0]$. The random-single expectation is the mean of the four correctness indicators. Classifying only after averaging four margins is reported separately as the random-four margin ensemble.

Table 2: RQ1 held-out behavioral effects in logit-margin units

Preferred minus baseline	Mean margin difference	95% query-bootstrap CI
Frozen random-four mean	0.2526	[0.1948, 0.3103]
Numerical median	0.2583	[0.1776, 0.3382]
TF-IDF nearest	0.2543	[0.1809, 0.3283]
Worst candidate	0.3943	[0.2984, 0.4906]

Note: Correct-vs-competing logit margins are computed with the held-out evaluation template. The random row uses the within-query mean of the frozen four, which estimates the expected margin of a random single demonstration.

Preferred minus random is the primary confirmatory comparison; median and TF-IDF are frozen secondary comparisons, and worst is descriptive. All confidence intervals are pointwise 95% intervals.

Table II reports **RQ1**. The primary preferred-minus-random margin is 0.2526, and its pointwise 95% query-bootstrap confidence interval lies entirely above zero. Preferred single-demonstration accuracy is 76.56%, compared with a random-single expectation of 69.53%; the paired gain is 7.03 percentage points (95% CI [3.91, 10.25]). No-demonstration accuracy is 63.67%. The random-four margin ensemble reaches 72.27%, but it uses four demonstrations and is therefore not a budget-matched baseline for a single preferred demonstration.

The frozen random set contains the preferred candidate for 90 of the 256 queries. Preferred minus the mean of the full 12-candidate pool is 0.2561, close to the primary estimate. The positive result therefore does not depend on an unusually weak random subset. The margin gain is larger for Yes than for No, which motivated the subsequent mapping experiment.

5.2 RQ2: Query Specificity in Late-Layer Residual Representations

RQ2 tests query specificity by applying preferred-minus-worst residual differences at the final prompt token in layers 19, 21, and 22. Each receiver is evaluated with its matched donor, two unrelated same-label donors, and two opposite-label donors; shuffled-content and norm-matched random vectors serve as controls. Forward and reverse margin changes are averaged to form the bidirectional intervention effect.

A one-dimensional Yes-minus-No direction is estimated from a separate balanced sample and removed from each donor difference before the projected comparison. This projection tests whether any matched excess remains beyond an independently estimated label direction.

Table 3: RQ2 Receiver-Level Residual Intervention Effects

Condition or contrast	Mean bidirectional margin change	95% receiver-bootstrap CI
Matched donor, full	0.6948	[0.5698, 0.8229]
Cross-query same-label, full	0.6671	[0.5746, 0.7613]
Matched excess, full	0.0278	[-0.1268, 0.1894]
Matched donor, label direction removed	0.6789	[0.5557, 0.8057]
Cross-query same-label, direction removed	0.6617	[0.5681, 0.7575]
Matched excess, direction removed (primary)	0.0172	[-0.1387, 0.1769]
Opposite-label donor, full	-0.6815	[-0.7821, -0.5865]
Norm-matched random vector	-0.0197	[-0.0798, 0.0407]

Note: Effects are logit-margin changes computed with the held-out evaluation template after donor and layer aggregation within each receiver. Matched excess equals the matched-donor effect minus the mean cross-query same-label donor effect. Direction-removed matched excess is the confirmatory RQ2 estimand; the other rows are controls or descriptive results.

Table III reports **RQ2**. The matched intervention is strong, but an unrelated same-label donor is nearly as strong. Using the receiver-level matched-excess definition, the full matched excess is 0.0278 with a 95% receiver-bootstrap confidence interval of [-0.1268, 0.1894]. After removing the separately estimated label direction, the preregistered matched excess is 0.0172 with an interval of [-0.1387, 0.1769]. The **RQ2** success criterion is therefore not met.

Opposite-label donors produce a similar-sized effect in the opposite direction, whereas a random vector with the same norm has little effect. This finding argues against vector norm alone as the main explanation. The one-dimensional label direction explains only a small part of the transfer, a pattern consistent with a distributed state or a mixture of label, confidence, and prompt-format information.

5.3 RQ3: Disentangling Semantic Meaning from Output Token Bias

RQ3 evaluates every formal query under two counterbalanced mappings: Yes-to-A with No-to-B, and the reverse. Preferred is selected separately under each mapping, while paired evaluation reuses the identical frozen candidate pool and random subset. The semantic-class-by-token cells are then aggregated into semantic and output-token contrasts.

Table 4. RQ3 counterbalanced a/b effects and confirmatory contrasts

Cell or contrast	Mean preferred-random effect	Simultaneous 95% CI
Yes; correct token A	-0.6029	Not estimated
No; correct token A	-0.8328	Not estimated
Yes; correct token B	+0.4662	Not estimated
No; correct token B	+0.4779	Not estimated
C_{sem} : Yes-No	+0.1091	[-0.0493, 0.2660]
C_{tok} : A-B	-1.1899	[-1.3623, -1.0197]

Note: All effects are logit-margin differences computed with the held-out evaluation template. Cell means are descriptive. The two confirmatory contrasts use Bonferroni simultaneous 95% intervals (97.5% marginal intervals).

Table IV reports **RQ3**. The semantic contrast C_{sam} is 0.1091 with a Bonferroni simultaneous 95% interval [-0.0493, 0.2660]. The token contrast C_{tok} is -1.1899 with interval [-1.3623, -1.0197]. The effect changes sign with A versus B but remains similar across semantic classes. Under the frozen decision rule, the result is token-following rather than semantic-following.

5.4 RQ4: Evaluating Selector Calibration on Disjoint Extensions

After observing **RQ3**, we froze a disjoint confirmatory extension that compares raw correct-token log probability, binary margin, and content-free calibration while keeping selection and held-out evaluation templates separate. The primary test asks whether calibration reduces the raw B -minus- A token gap. Practical elimination requires the calibrated 90% confidence interval to lie wholly inside the preset range $[-0.20, 0.20]$.

Both **RQ1** and **RQ4** call a candidate correct when $M_E > 0$, so the candidate-level prediction rule is the same. Their absolute accuracies are not a controlled comparison: they use disjoint formal queries and candidate banks, different output vocabularies (Yes/No versus A/B), and separately frozen stages. Both pools were fixed before formal model outcomes and were not screened by observed accuracy; nevertheless, the new sample and disjoint oracle same-label pool may differ in difficulty. Because no common anchor set was run, the source of any level difference cannot be separated. Cross-stage claims therefore compare within-stage paired effects and token gaps, not raw accuracy levels.

Table 5: RQ4 frozen selector results on the disjoint extension

Selector	Overall effect [95% CI]	B-A gap [95% CI]	Preferred single-demo accuracy
Raw log probability	-0.1640 [-0.2343, -0.0927]	+1.0793 [0.9190, 1.2376]	99.41%
Binary margin	+0.9006 [0.8353, 0.9655]	-0.2920 [-0.4199, -0.1663]	99.61%
Content-free calibrated	+0.2149 [0.1446, 0.2860]	+0.3025 [0.1508, 0.4530]	99.22%

Note: Effect and gap are held-out evaluation-template logit-margin quantities with paired, stratified query-bootstrap 95% confidence intervals. Accuracy is for the selected single demonstration; the random-single expectation is 93.46%. This confirmatory extension follows a separately frozen protocol. The calibrated equivalence test uses a 90% confidence interval rather than the 95% interval shown here.

Table V reports **RQ4**. Raw selection is harmful on average under the held-out binary-margin metric and strongly favors B . Binary-margin selection produces the largest average effect, but its smaller remaining gap favors A . Content-free calibration produces a positive average effect and reduces the preference for B . The primary raw-minus-calibrated gap reduction is 0.7769 with a paired 95% confidence interval of $[0.5977, 0.9551]$, so the primary reduction test passes.

Calibration does not eliminate token dependence: its $B - A$ gap is 0.3025, and the 90% confidence interval $[0.1758, 0.4293]$ does not lie entirely within the preset equivalence range $[-0.20, 0.20]$. Preferred accuracy is close to 99% for all selectors, but the random-single expectation is already 93.46%, so accuracy is nearly saturated. These levels should not be compared directly with the **RQ1** value of 76.56% because the samples, candidate banks, and output tokens differ. Without a shared anchor set, the difference cannot be attributed to query difficulty, candidate-pool composition, or the A/B verbalizer. Cross-stage evidence therefore rests on paired margins, token gaps, and selector differences that accuracy would obscure.

5.5 Post-Hoc Exploratory Analysis: Symmetric Calibrated Selector

We used the saved candidate-level logits for a post hoc decomposition; this analysis was not part of the frozen extension. Holding the raw candidate fixed still leaves a $B - A$ gap of 0.8687 when the candidate is chosen under A and 0.6862 when chosen under B . Thus, candidate switching is not the whole explanation. At the same time, fixed-candidate results for the other selectors depend strongly on the mapping used for selection, showing a candidate-by-mapping interaction.

An exploratory symmetric calibrated score averages calibrated selection margins over both counterbalanced mappings and then uses one candidate in both evaluations. On the existing formal set, it yields an average effect of 0.2656 with a 95% interval from 0.1989 to 0.3326, and a $B - A$ gap of 0.0020 with an interval from -0.1693 to 0.1693. This is a promising design clue, not confirmatory evidence, because it was discovered after inspecting the formal data.

6. Discussion

6.1 What Demonstration Preference Scores Truly Capture

The four research questions yield a consistent but qualified answer. **RQ1** supports held-out utility; **RQ2** does not support stable query specificity at the tested intervention sites; **RQ3** is token-following; and **RQ4**

identifies no selector that both maximizes utility and removes token dependence. We can say the behavioral result is therefore genuine but narrow. Within a fixed Yes/No prompt, the raw score identifies candidates that transfer better to held-out wording. The *A/B* and selector experiments show why this result should not be interpreted as semantic example quality. Raw correct-token probability depends on the full output distribution, not only on competition between the two task labels. Replacing it with a binary margin greatly improves average transfer but reverses the remaining token preference. A score can therefore be useful while remaining verbalizer-dependent.

Content-free calibration removes part of the prompt-level answer prior, but a single N/A query is insufficient to establish invariance. This result is consistent with work showing that label words and decision criteria shape few-shot behavior [7, 9, 17, 19]. A practical audit should report both utility and stability under counterbalanced label mappings.

6.2 Scope, Boundaries, and Mechanistic Interpretations

The intervention results support a shared label-conditioned state at the tested layer, final token locations, but they do not support a stable matched query advantage there. This is a boundary result, not evidence that query-specific information is absent from the model. Such information may appear at other positions, across a subspace, or through nonlinear interactions. The finding also illustrates why matched interventions require cross-query and opposite-label controls [20, 22].

6.3 Limitations and Threats to Validity

The evidence comes from one small model, one binary task, one candidate-pool design, two primary output tokens per mapping, and a single content-free query. Candidate demonstrations are also reused across some formal queries, so the query bootstrap should be interpreted as inference over queries conditional on a fixed demonstration bank. The mechanism study covers only selected responders, three layers, and the final prompt token.

The highest-priority next step is an independent confirmation of the symmetric calibrated selector. The protocol should be published before model outcomes are observed, assign non-overlapping candidate pools across queries, use several content-free prompts, and hold one candidate fixed across counterbalanced mappings. A minimum generalization study should add one larger model and one additional binary task; a stronger study should also add a different model family and several single-token label pairs. Primary success criteria should jointly require positive held-out utility and token-gap equivalence.

6.4 Practical Recommendations for Demonstration Selection

This study is an oracle diagnostic: it uses the query's gold label both to construct the same-label pool and to score the correct output, so it is not a deployable retrieval algorithm. However, the evidence supports three concrete rules for NLP and RAG/prompt pipelines for labeled development and audit sets.

Avoid raw correct-token probability as a stand-alone demonstration-retrieval score. It mixes useful transfer with full-vocabulary token compatibility and was harmful on average under the counterbalanced *A/B* evaluation. Always compare it with a frozen random baseline under held-out prompt wording.

Use the binary competition margin whenever two labels define the decision. Rank candidates by the correct-minus-competing logit margin and evaluate with the same decision quantity under an independently worded template; report margin utility as well as accuracy when accuracy is near saturation.

Require symmetry under reversed label mappings. Counterbalance the verbalizer, keep candidates fixed across mappings, and predefine an equivalence bound for the output-token gap. Treat the symmetric calibrated selector as a promising audit pattern that still requires independent confirmation before deployment.

7. Conclusion

A preferred demonstration can improve behavior under a fixed label setting, but the raw preference score is not a direct measure of semantic match. Much of its variation is related to the output token and changes with the selector. Binary-margin selection improves average held-out utility, whereas content-free calibration

reduces but does not eliminate token dependence. Late-layer residual transfer is strong but largely shared across queries with the same label. These results support a careful interpretation that demonstration preference belongs to the full prompt, verbalizer, and scoring rule, not to the demonstration alone.

References

- [1] T. Brown et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877-1901.
- [2] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity,” in *Proc. ACL*, 2022, pp. 8086-8098, doi: 10.18653/v1/2022.acl-long.556.
- [3] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, “Rethinking the role of demonstrations: What makes in-context learning work?” in *Proc. EMNLP*, 2022, pp. 11048-11064, doi: 10.18653/v1/2022.emnlp-main.759.
- [4] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, “What makes good in-context examples for GPT-3?” in *Proc. DeeLIO*, 2022, pp. 100-114, doi: 10.18653/v1/2022.deelio-1.10.
- [5] O. Rubin, J. Herzig, and J. Berant, “Learning to retrieve prompts for in-context learning,” in *Proc. NAACL-HLT*, 2022, pp. 2655-2671, doi: 10.18653/v1/2022.naacl-main.191.
- [6] Z. Zhang, S. Lan, L. Song, J. Bian, Y. Li, and K. Ren, “Learning to select in-context demonstration preferred by large language model,” in *Findings of ACL*, 2025, pp. 11345-11360, doi: 10.18653/v1/2025.findings-acl.592.
- [7] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, “Calibrate before use: Improving few-shot performance of language models,” in *Proc. ICML*, PMLR 139, 2021, pp. 12697-12706.
- [8] J. Wei et al., “Larger language models do in-context learning differently,” arXiv:2303.03846, 2023.
- [9] L. Wang, L. Li, D. Dai, D. Chen, H. Zhou, F. Meng, J. Zhou, and X. Sun, “Label words are anchors: An information flow perspective for understanding in-context learning,” in *Proc. EMNLP*, 2023, pp. 9840-9855, doi: 10.18653/v1/2023.emnlp-main.609.
- [10] R. Hendel, M. Geva, and A. Globerson, “In-context learning creates task vectors,” in *Findings of EMNLP*, 2023, pp. 9318-9333, doi: 10.18653/v1/2023.findings-emnlp.624.
- [11] E. Todd, M. Li, A. Sen Sharma, A. Mueller, B. C. Wallace, and D. Bau, “Function vectors in large language models,” in *Proc. ICLR*, 2024.
- [12] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, “Locating and editing factual associations in GPT,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 17359-17372.
- [13] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, S. Sakenis, J. Huang, Y. Singer, and S. Shieber, “Causal mediation analysis for interpreting neural NLP: The case of gender bias,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12388-12401.
- [14] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, “BoolQ: Exploring the surprising difficulty of natural yes/no questions,” in *Proc. NAACL-HLT*, 2019, pp. 2924-2936, doi: 10.18653/v1/N19-1300.
- [15] Qwen Team, “Qwen2.5 technical report,” arXiv:2412.15115, 2024.
- [16] A. Holtzman, P. West, V. Shwartz, Y. Choi, and L. Zettlemoyer, “Surface form competition: Why the highest probability answer is not always right,” arXiv:2104.08315, 2021.
- [17] Y. Reif and R. Schwartz, “Beyond performance: Quantifying and mitigating label bias in LLMs,” in *Proc. NAACL*, 2024, arXiv:2405.02743.
- [18] Z. Yu and S. Ananiadou, “How do large language models learn in-context? Query and key matrices of in-context heads are two towers for metric learning,” in *Proc. EMNLP*, 2024, arXiv:2402.02872.

- [19] H. Cho, Y. Sakai, M. Kato, K. Tanaka, A. Ishii, and N. Inoue, “Token-based decision criteria are suboptimal in in-context learning,” in Proc. NAACL-HLT, 2025, pp. 5378-5401, doi: 10.18653/v1/2025.naacl-long.278.
- [20] F. Zhang and N. Nanda, “Towards best practices of activation patching in language models: Metrics and methods,” in Proc. ICLR, 2024.
- [21] M. Liu, “In-context fixation: When demonstrated labels override semantics in few-shot classification,” arXiv:2605.08295, 2026.
- [22] B. Cheng and J. Zhang, “Single-position intervention fails: Distributed output templates drive in-context learning,” arXiv:2605.04061, 2026.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).