ProBA: Prompt Brittleness Mitigation via Layer-wise Localization and Representation Alignment

Kexiang Qiao[1], Yang Li[2,*], Suge Wang[1,3], Jian Liao[1], Jianxing Zheng[1,3], Deyu Li[1,3]
[1]*School of Computer and Information Technology, Shanxi University, Taiyuan, China*
[2]*School of Finance, Shanxi University of Finance and Economics, Taiyuan, China*
[3]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China*
Email: qiao.kexiang@qq.com, liyang@sxufe.edu.cn, wsg@sxu.edu.cn, liaoj@sxu.edu.cn, jxzheng@sxu.edu.cn, lidy@sxu.edu.cn

***Abstract*—Large Language Models (LLMs) demonstrate strong performance across diverse tasks, yet remain susceptible to prompt brittleness: minor perturbations in input prompts can degrade output quality and undermine model reliability. However, existing approaches rely on data augmentation or input adjustments, failing to address representation misalignment that causes the inference trajectory drift. To address these limitations, we draw inspiration from how humans maintain cognitive consistency: we localize sources of uncertainty within the reasoning process and realign internal representations to ensure stable reasoning across varying contexts. Motivated by this cognitive paradigm, we propose Prompt Brittleness Mitigation via Layer-wise Localization and Representation Alignment (ProBA), a general framework designed to ensure inference trajectory consistency. ProBA employs Layer-wise Brittleness Localization (LBL), guided by cosine divergence-based stability analysis, to localize the brittle layers responsible for inference drift. The representations within these layers are then rectified through Intra-layer Representation Alignment (IRA), which employs selective parameter freezing and MSE geometric constraints. Extensive experiments on LLMs with various architectures and different tasks show that our ProBA outperforms existing methods, achieving improvements in effectiveness and stability.**

***Index Terms*—Large Language Models, Prompt Brittleness, Representation Alignment**

I. Introduction

In recent years, LLMs have exhibited remarkable progress in multi-task generalization and generative capabilities. However, prompt brittleness critically undermines their performance stability [1]. In some cases, models even rely primarily on surface-level syntactic patterns or specific formatting templates to respond, rather than a genuine understanding of the prompt semantics [2], [3]. This phenomenon manifests macroscopically as the instability of outputs and microscopically as the divergence of internal inference paths. As illustrated in Fig. 1, two semantically equivalent prompts, P1 and P2, follow coincident inference trajectories in the early layers. However, due to the failure of representation alignment in specific intermediate layers (e.g., $Layer\_q$, $Layer\_k$), the inference path of P2 undergoes significant drift within the representation space, ultimately leading to erroneous output results.

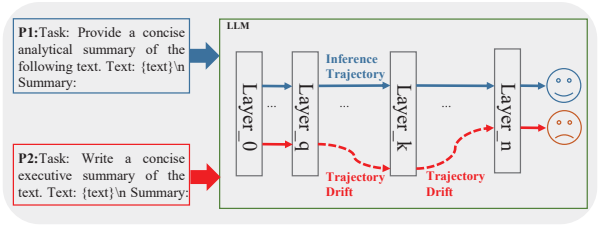

Fig. 1. Illustration of inference path divergence caused by representation misalignment. The variant prompt (red path) drifts away from the stable anchor trajectory (blue path) after passing through specific layers (Layer_q, Layer_k), ultimately leading to incorrect outputs.

Existing methods adopt various strategies to mitigate prompt brittleness. (1) Prompt engineering and discrete optimization methods attempt to mitigate prompt brittleness by automatically searching or generating optimal prompts in the discrete input space [4], [5], [6]. However, these strategies focus solely on "surface-level input adaptation" and fail to rectify the misalignment of the internal representations of the model. (2) Data augmentation and context enhancement improve stability by introducing diverse synthetic prompt samples or utilizing voting mechanisms across multiple inference paths [1], [7]. However, these methods rely on data fitting without explicitly rectifying the representational misalignment that drives inference-trajectory drift. (3) LoRA and its variants achieve model adaptation by updating a small number of parameters, but their optimization objectives are usually limited to the final prediction loss [8], [9]. Optimizing the output probability distribution often fails to capture and rectify the geometric divergence occurring inside the deep network [10].

To address these challenges, we draw inspiration from the cognitive paradigm by which humans maintain conceptual consistency. This process typically begins with locating the

This work is supported by the National Natural Science Foundation of China (NO.U24A20335, 62376143, 62473241, 62476162, 62272286) and Fundamental Research Program of Shanxi Province (No.202503021211239, 202303021211021).

source of cognitive inconsistency, followed by calibrating internal understanding to a stable standard to ensure stable reasoning. Motivated by this observation, we propose Prompt Brittleness Mitigation via Layer-wise Localization and Representation Alignment (ProBA), a general framework designed to ensure inference trajectory consistency. ProBA incorporates Layer-wise Brittleness Localization (LBL) to precisely locate brittle layers via cosine divergence analysis. Subsequently, the Intra-layer Representation Alignment (IRA) strategy rectifies the representation within these layers using selective parameter freezing and targeted MSE constraints, effectively rectifying the inference trajectory drift and ensuring the consistency of the inference path. In a nutshell, our major contributions are as follows:

- We propose ProBA, a general framework that locates brittle layers and performs targeted interventions on divergent inference trajectories, effectively mitigating the prompt brittleness problem in large language models.
- We propose the LBL method to locate the exact positions of these brittle layers that cause inference-path drift. Furthermore, we introduce the IRA strategy to rectify the divergent inference trajectories.
- Extensive experiments across both multi-modal and text-only benchmarks show that our ProBA consistently improves large language model performance on multiple downstream tasks, verifying its superior efficacy.

II. Related Work

A. Prompt Brittleness and Representation Analysis

LLMs exhibit brittleness to non-semantic changes in input prompts [11], [12]. Sclar et al. provided a quantitative analysis of this phenomenon, revealing that even minor formatting alterations can induce severe performance instability [13]. Sun et al. showed that prompt variations erode zero-shot stability, revealing overfitting to specific syntactic templates in the training data [14]. Zou et al. proposed the concept of Representation Engineering, proving that high-level behaviors like honesty and safety are encoded as "directions" in specific layers, and intervening in these local representations can effectively control model behavior [3]. Polo et al. proposed Prompt Eval, advocating for a "multi-prompt" rather than "single-prompt" evaluation protocol to ensure more stable performance estimation [15]. Furthermore, Li et al. utilized Inference-Time Intervention to confirm that critical reasoning capabilities are often localized within specific attention heads rather than being diffused throughout the network [16]. Similarly, Hendel et al. showed that In-Context Learning operates by constructing specific task vectors internally [2], while Wolf et al. pointed out that information flow mainly propagates across layers via specific "anchor" tokens; once the geometric structure of these key layers is misaligned, the inference chain is disrupted [17].

B. Fine-Tuning Strategies

Parameter-Efficient Fine-Tuning (PEFT) aims to adapt models by updating a small number of parameters. LoRA and its variants such as QLoRA and DoRA, although performing excellently on specific tasks, usually use end-to-end cross-entropy loss for optimization, ignoring the geometric consistency of intermediate layers [8], [9], [18]. Wei et al. proposed PAFT, which injects diverse synthetic prompt samples during fine-tuning to force the model to learn core task semantics instead of clinging to specific prompt expressions [1]. Similarly, Ngweta et al. proposed the MOF strategy to mitigate the prompt brittleness problem by introducing multiple prompt formats in few-shot exemplars [19]. Liu et al. optimized prompt content and format simultaneously, improving performance consistency under diverse prompt conditions by iteratively exploring combinations of content and format [20]. Hu et al. adopted a localized search strategy [5]. Although these works improve model stability to prompt variations from different angles, they mainly focus on operations in the prompt input space and fail to directly analyze and intervene in the instability caused by prompts at the level of model internal representations.

III. Method

In this section, we present the ProBA framework in detail. As illustrated in Fig. 2, the workflow begins with a Prompt Generation and Anchor Selection phase. Inspired by Hu et al. [5], we first evaluate a diverse set of GPT-generated prompt candidates. The highest-performing prompt is designated as the anchor (P_{anc}), while the remaining prompts serve as variants (P_{var}). Following this prompt construction, the framework executes two steps: (i) locating the brittle layers of the model via LBL and (ii) rectifying the inference trajectory drift in these layers via the IRA fine-tuning strategy.

A. Problem Formulation

Let $\mathcal{M}$ denote a large model and $\mathcal{D}$ represent the dataset for a downstream task. For an arbitrary input instance $x \in \mathcal{D}$ and a given prompt p, the intermediate hidden state representation at layer l is formalized as:

$$h_l = \mathcal{M}_l(x, p) \tag{1}$$

where $l \in \{1, 2, \ldots, L\}$ is the index of the layer, and L is the total number of layers.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote the set of n semantically equivalent prompt candidates. To distinguish the anchor from variants, we evaluate each candidate on the training data $\mathcal{D}_{train}$ using a task-specific metric $\mathcal{G}$ (e.g., Accuracy or ROUGE-L). We formally define the prompt space as follows:

(1) Anchor Prompt (p_{anc}): Let p_{anc} denote the candidate that maximizes the expected performance metric, formalized as:

$$p_{anc} = \arg\max_{p \in \mathcal{P}} \mathbb{E}_{(x,y) \in \mathcal{D}_{train}} [\mathcal{G}(f(x, p), y)] \tag{2}$$

where $f(x, p)$ is the predicted output of the model given input x and prompt p, and y represents the corresponding ground truth label. p_{anc} serves as the geometric reference for the optimal inference trajectory.

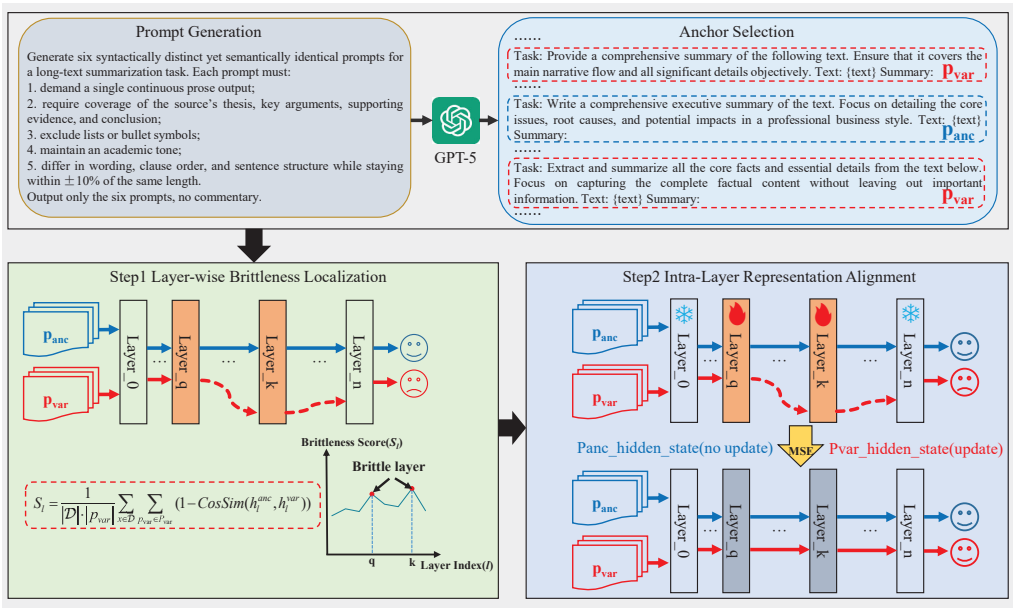

Fig. 2. Overview of the ProBA framework. The workflow begins with anchor selection, where GPT-5 generates diverse prompt candidates. The optimal candidate is selected as the anchor (p_{anc}), while the rest serve as variants (p_{var}).

(2) Variant Prompts ($p_{var} \in \mathcal{P}_{var}$): Let P_{var} denote the set of remaining candidates after removing the optimal anchor, formalized as:

$$\mathcal{P}_{var} = \mathcal{P} \setminus \{p_{anc}\} \tag{3}$$

where $p_{var} \in \mathcal{P}_{var}$ retains semantic equivalence to p_{anc} while introducing syntactic variations to simulate input perturbations.

The anchor p_{anc} and the variant p_{var}, which are semantically equivalent but syntactically different, cause the internal representation of the model at layer l to undergo significant drift, i.e., $h_l(x, p_{anc}) \neq h_l(x, p_{var})$, subsequently leading to divergence in downstream inference results. Our objective is to derive a parameter update strategy $\Delta\theta$ that minimizes this inference trajectory drift while preserving the general capabilities of the model:

$$\min_{\Delta\theta} \mathbb{E}_{x \in D}[Dist(h_l(x, p_{var}; \theta + \Delta\theta), h_l(x, p_{anc}; \theta))] \tag{4}$$

where $Dist(\cdot)$ is a representation distance metric function, and θ represents the original pre-trained parameters of the model.

B. Layer-wise Brittleness Localization(LBL)

To precisely locate the brittle layers responsible for inference inconsistency, we propose the LBL method. Unlike conventional approaches that focus solely on final output probabilities, LBL is designed to directly measure the geometric divergence of hidden states across different layers.

For an arbitrary sample $x \in \mathcal{D}$, we construct inputs by combining x with p_{anc} and p_{var}, respectively. These inputs are fed into the model, and for each layer $l \in \{1, 2, \ldots, L\}$, we extract the hidden state vector corresponding to the last token:

$$h_l^{anc} = \mathcal{M}_l(x, p_{anc}) \tag{5}$$

$$h_l^{var} = \mathcal{M}_l(x, p_{var}) \tag{6}$$

To quantify the degree of representation deviation between h_l^{anc} and h_l^{var}, we define the brittleness score of the l-th layer, adopting the expectation of cosine dissimilarity, specifically defined as follows:

$$\text{CosSim}(h_l^{anc}, h_l^{var}) = \frac{(h_l^{anc} \cdot h_l^{var})}{\|h_l^{anc}\| \|h_l^{var}\|} \tag{7}$$

$$S_l = \frac{1}{|\mathcal{D}| \cdot |p_{var}|} \sum_{x \in \mathcal{D}} \sum_{p_{\text{var}} \in P_{\text{var}}} (1 - CosSim(h_l^{anc}, h_l^{var})) \tag{8}$$

where $\text{CosSim}(\cdot)$ is the standard cosine similarity function, and S_l is the brittleness score for the l-th layer.

A higher S_l implies that the layer generates more severe representation drift under semantically equivalent inputs. Based on this brittleness assessment, we define the set $\mathcal{S}$ of brittle layers as the Top-K layers exhibiting the highest brittleness scores:

$$\mathcal{S} = TopK_l(\{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_L\}) \tag{9}$$

where L is the total number of Transformer layers in the large model.

C. Intra-layer Representation Alignment(IRA)

Upon locating the set of brittle layers $\mathcal{S}$, we propose the Intra-layer Representation Alignment (IRA) fine-tuning strategy to rectify the inference trajectory drift. The principle of IRA utilizes an anchor-guided alignment mechanism to

enforce explicit geometric constraints, compelling the model's internal states to align with the stable activation patterns of the anchor prompt when processing variant prompts.

To mitigate brittleness while maximally preserving the general knowledge of the model and preventing catastrophic forgetting, we implement a rigorous Parameter-Efficient Fine-Tuning (PEFT) strategy. Specifically, we freeze the entire large model $\mathcal{M}$ and only unfreeze the self-attention projection matrices of the brittle layers in $\mathcal{S}$. Let $W_l = \{W_Q, W_K, W_V\}_l$ denote the weights of the self-attention module at layer l. The set of trainable parameters θ_{train} is defined as:

$$\theta_{train} = \bigcup_{l \in S} W_l \tag{10}$$

For every training instance x, we construct input pairs consisting of p_{anc} and p_{var}, computing their respective hidden state representations at each brittle layer l. To formulate the optimization objective for representation alignment, we define the process as follows:

(1) Anchor Path: The pair (x, p_{anc}) is fed into the model to compute the hidden state h_l^{anc} at brittle layer l. Serving as the stable geometric reference, we apply a Stop-Gradient operation to this representation to prevent the intrinsic representations of the anchor from being altered during optimization.

(2) Variant Path: The pair (x, p_{var}) is fed into the model to compute the corresponding hidden state h_l^{var}. Unlike the anchor path, this path maintains the computational graph to enable gradient backpropagation, allowing the parameters to be updated to align h_l^{var} with the reference h_l^{anc}.

Our optimization objective is to minimize the MSE between these two outputs across all brittle layers:

$$\mathcal{L}_{IRA} = \frac{1}{|S|} \sum_{l \in \mathcal{S}} \|h_l^{var} - sg(h_l^{anc})\|_2^2 \tag{11}$$

where $\mathcal{L}_{IRA}$ is the loss function, and $\mathrm{sg}(\cdot)$ is the stop-gradient operation.

During training, gradients are backpropagated exclusively through the variant path to update the parameters in θ_{train}. The specific update rule is formulated as:

$$\theta_{train} \leftarrow \theta_{train} - \eta \nabla_{\theta_{train}} \mathcal{L}_{IRA} \tag{12}$$

where η is the learning rate, and $\nabla_{\theta_{train}}$ is the gradient with respect to the trainable parameters.

By employing this optimization process, IRA explicitly rectifies the inference trajectories of variant prompts, compelling them to align within the representation space with the stable representations established by the anchor prompt.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. We use four diverse datasets covering both multimodal and pure text tasks: **Weibo** [25], **MVSA-Single** [26], **GovReport** [27], and **Qasper** [27]. Accuracy is reported for Weibo and MVSA-Single; ROUGE-L and F1 for GovReport and Qasper.

Models and Baselines. we select four open-source architectures with varying parameter scales to verify cross-architecture generalizability: **GLM-4V-9B** [21], **Qwen2.5-VL-7B** [22], **Llama-3.2-11B-Vision** [24], **Qwen3-8B** [23], **GLM-4-9B** [21] and **Llama-3-8B** [24]. To validate the superiority of our framework, we compare ProBA against four methods: **LoRA** [8], **ZOPO** [5], **PAFT** [1] and **Vanilla Zero-Shot (Vanilla)**, which serves as the lower bound to assess intrinsic brittleness.

Implementation Details. All experiments were implemented using PyTorch on NVIDIA RTX 4090, RTX A6000, A100, and H100 GPUs. To prevent overfitting to the anchor prompt, we randomly sample 60 instances for the training set. The hyperparameter K for brittle layer localization was set to 2. We employed the AdamW optimizer with a learning rate of 5×10^{-5}, a batch size of 1, and 8 gradient accumulation steps. Training was conducted for 2 epochs. The maximum input sequence length was truncated to 2,048 tokens for multimodal tasks and 6,000 tokens for pure text tasks.

B. Main Results and Analysis

Table I and Table II show the performance of ProBA compared with four baseline methods on multimodal discrimination and text-only generation tasks. Our ProBA performs well on all metrics across different modalities and architectures, consistently surpassing the second-best baselines.

For multimodal discrimination tasks, we outperform existing methods on the MVSA dataset across GLM-4V, Qwen2.5-VL, and Llama3.2-V, surpassing the second-best baselines by 1.11%, 12.12%, and 14.04%, respectively. It demonstrates that our proposed framework ProBA can align internal representation misalignment and rectify divergent inference trajectories through the collaborative application of LBL and IRA. For the Weibo dataset, ProBA surpasses all baselines under Qwen2.5-VL and Llama3.2-V. We note that ZOPO achieves marginally higher accuracy on GLM-4V, however, this is attributed to the simpler semantic structure of short social media texts, where discrete prompt optimization tends to converge to locally optimal patterns but lacks generalization.

For text-only generation tasks, ProBA consistently achieves the highest R-L and F1 scores on the GovReport and Qasper datasets across GLM-4, Qwen3, and Llama3. It should be emphasized that ProBA maintains stable performance even under complex reasoning architectures. Specifically, under Qwen3, ProBA preserves an R-L score of 19.29 on GovReport and an F1 score of 27.56 on Qasper. Compared to the ZOPO method, which suffers catastrophic performance drops as evidenced by its F1 score falling to 6.42 on Qasper, our approach is more effective. This further demonstrates the effectiveness of ProBA in handling long-context dependencies and complex logical reasoning, validating its broad generalizability across diverse textual scenarios.

Fig. 3 illustrates the specific performance trajectories of each method under twelve distinct prompt variations.

For multimodal discrimination tasks, ProBA markedly suppresses performance fluctuations induced by prompt variations, especially on MVSA with Qwen2.5-VL and on Weibo

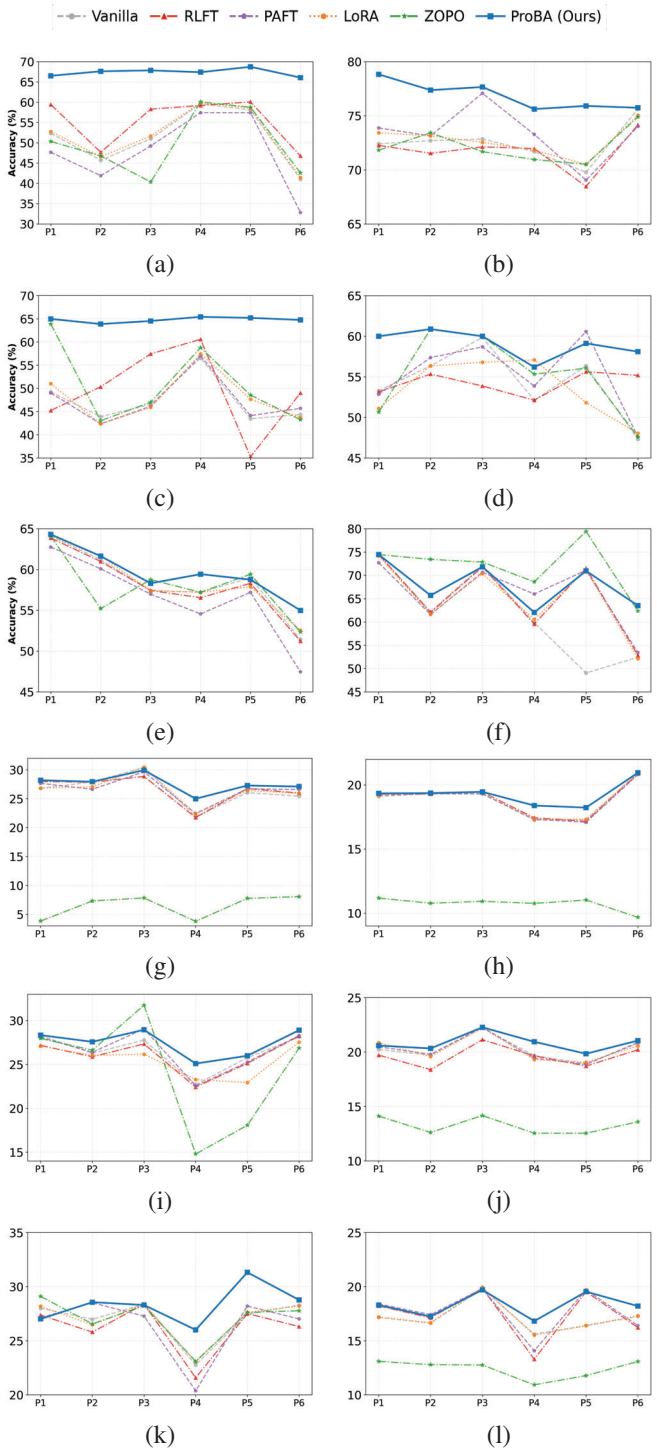

Fig. 3. Performance distributions of different models across twelve prompt variations. The results are reported for (a) Qwen2.5-VL on MVSA, (b) Qwen2.5-VL on Weibo, (c) Llama3.2-V on MVSA, (d) Llama3-V on Weibo, (e) GLM-4V on MVSA, (f) GLM-4V on Weibo, (g) Qwen3 on Qasper, (h) Qwen3 on GovReport, (i) Llama3 on Qasper, and (j) Llama3 on GovReport, (k) GLM4 on Qasper, and (l) GLM4 on GovReport.

TABLE I
MAIN RESULTS ON MULTI-MODAL TASKS

Model	Method	MVSA Acc	Weibo Acc
GLM-4V	Vanilla	58.39	61.68
	LoRA	58.46	65.04
	ZOPO	57.84	**71.85**
	PAFT	56.51	65.91
	ProBA(Ours)	**59.57**	68.08
Qwen2.5-VL	Vanilla	51.30	72.51
	LoRA	51.81	72.75
	ZOPO	49.81	72.21
	PAFT	47.75	73.41
	ProBA(Ours)	**67.37**	**76.86**
Llama3.2-V	Vanilla	47.34	54.21
	LoRA	48.01	53.53
	ZOPO	50.74	55.09
	PAFT	47.38	55.16
	ProBA(Ours)	**64.78**	**59.05**

TABLE II
MAIN RESULTS ON PURE TEXT GENERATION TASK

Model	Method	GovReport R-L	Qasper F1
GLM4	Vanilla	17.14	26.97
	LoRA	17.17	26.99
	ZOPO	12.40	27.06
	PAFT	17.63	26.41
	ProBA(Ours)	**18.30**	**28.33**
Qwen3	Vanilla	18.87	26.41
	LoRA	18.93	26.50
	ZOPO	10.72	6.42
	PAFT	18.84	26.60
	ProBA(Ours)	**19.29**	**27.56**
Llama3	Vanilla	20.25	26.48
	LoRA	20.26	25.50
	ZOPO	13.26	24.34
	PAFT	20.29	26.60
	ProBA(Ours)	**20.84**	**27.47**

with Llama3.2-V. Unlike baseline methods that continue to exhibit pronounced oscillations, our ProBA maintains high stability. It demonstrates that intra-layer representation alignment via IRA on brittle layers identified by LBL effectively mitigates prompt brittleness.

For text-only generation tasks, ProBA likewise maintains a stable trajectory. It should be emphasized that the ZOPO method suffers catastrophic collapse on Qwen3 Qasper and GovReport subplots. This visually demonstrates that ProBA rectifies the inference trajectory drift to deliver superior generalization under long-context dependencies and complex logic.

C. Ablation Study

Table III compares ProBA with the Vanilla baseline, which performs standard inference without any optimization strategies, and with the w/o LBL variant that replaces layer selection guided by LBL with random selection. Our ProBA performs well across all datasets, consistently achieving the highest performance. On MVSA with Qwen2.5-VL, ProBA surpasses

w/o LBL by 12.1%. It demonstrates that our proposed LBL method effectively locates the specific layers responsible for inference trajectory drift. It should be emphasized that this effectiveness extends to long-context scenarios. Compared to the w/o LBL variant, which assumes uniform brittleness across the network, our ProBA is more effective. This further demonstrates that prompt brittleness is not uniformly scattered but concentrated in specific brittle layers.

TABLE III
ABLATION RESULTS

Model	Method	MVSA Acc	Weibo Acc	GovReport R-L	Qasper F1
GLM	Vanilla	58.39	61.68	17.14	26.97
	w/o LBL	58.06	65.38	17.37	26.16
	ProBA(Ours)	**59.57**	**68.08**	**18.30**	**28.33**
Qwen	Vanilla	51.30	72.51	18.81	26.41
	w/o LBL	55.25	71.75	18.92	26.52
	ProBA(Ours)	**67.37**	**76.86**	**19.29**	**27.56**
Llama	Vanilla	47.34	54.21	20.25	26.48
	w/o LBL	49.63	54.21	19.63	26.02
	ProBA(Ours)	**64.78**	**59.05**	**20.84**	**27.47**

V. CONCLUSION

We introduce ProBA, a general framework that mitigates prompt brittleness by combining LBL with IRA. Unlike previous approaches, ProBA locates brittleness layers and rectifies divergent inference trajectories, markedly reducing prediction variance under prompt perturbations and improving downstream performance of large models. Despite these advances, the anchor-selection mechanism still has limitations: anchors currently rely on labeled data. In the future, we will focus on adaptive and unsupervised anchor mining algorithms to automatically locate stable anchors.

REFERENCES

[1] C. Wei, M. Ou, Y. He, Y. Shu, and F. Yu, "PAFT: Prompt-agnostic fine-tuning," in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 2025, pp. 694–717.

[2] R. Hendel, M. Geva, and A. Globerson, "In-context learning creates task vectors," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023, pp. 9318–9333.

[3] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren et al., "Representation engineering: A top-down approach to AI transparency," arXiv:2310.01405, 2023.

[4] M. Seleznyov, M. Chaichuk, G. Ershov, A. Panchenko, E. Tutubalina, and O. Somov, "When punctuation matters: A large-scale comparison of prompt robustness methods for LLMs," in Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, 2025, pp. 20370–20385.

[5] W. Hu, Y. Shu, Z. Yu, Z. Wu, X. Lin, Z. Dai, S.-K. Ng, and B. K. H. Low, "Localized zeroth-order prompt optimization," in Advances in Neural Information Processing Systems, vol. 37, 2024, pp. 86309–86345.

[6] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan et al., "Large language models are human-level prompt engineers," in Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.

[7] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in Proceedings of the 11th International Conference on Learning Representations (ICLR), 2023.

[8] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Advances in Neural Information Processing Systems, vol. 36, 2023, pp. 10088–10115.

[10] Y. Gu, L. Dong, F. Wei, and M. Huang, "MiniLLM: Knowledge distillation of large language models," in Proceedings of the International Conference on Learning Representations (ICLR), 2024, pp. 32694–32717.

[11] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. A. Hasan, "Does prompt formatting have any impact on LLM performance?," arXiv:2411.10541, 2024.

[12] A. Voronov, L. Wolf, and M. Ryabinin, "Mind your format: Towards consistent evaluation of in-context learning improvements," in Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2024, pp. 6287–6310.

[13] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, "Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting," in Proceedings of the International Conference on Learning Representations (ICLR), 2024, pp. 25055–25083.

[14] J. Sun, C. Shaib, and B. Wallace, "Evaluating the zero-shot robustness of instruction-tuned language models," in Proceedings of the International Conference on Learning Representations (ICLR), 2024, pp. 48103–48141.

[15] F. M. Polo, R. Xu, L. Weber, M. Silva, O. Bhardwaj, L. Choshen et al., "Efficient multi-prompt evaluation of LLMs," in Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS), vol. 37, 2024, pp. 22483–22512.

[16] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," in Advances in Neural Information Processing Systems, vol. 36, 2023, pp. 41451–41530.

[17] L. Wolf, T. Pimentel, E. Fedorenko, R. Cotterell, A. Warstadt, E. G. Wilcox et al., "Quantifying the redundancy between prosody and text," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023, pp. 9765–9784.

[18] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-decomposed low-rank adaptation," arXiv:2402.09353, 2024.

[19] L. Ngweta, K. Kate, J. Tsay, and Y. Rizk, "Towards LLMs robustness to changes in prompt format styles," in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 4: Student Research Workshop), Albuquerque, USA, 2025, pp. 529–537.

[20] Y. Liu, J. Xu, L. L. Zhang, Q. Chen, X. Feng, Y. Chen et al., "Beyond prompt content: Enhancing LLM performance via content-format integrated prompt optimization," arXiv:2502.04295, 2025.

[21] A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang et al., "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," arXiv:2406.12793, 2024.

[22] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang et al., "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," in Proceedings of the 37th AAAI Conference on Artificial Intelligence, 2023, pp. 1–9.

[23] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng et al., "Qwen3 technical report," arXiv:2505.09388, 2025.

[24] A. Grattafiori, A. Dubey, A. Jauhri, A.Pandey, A. Kadian, A.-A Dahle et al., "The Llama 3 herd of models," arXiv:2407.21783, 2024.

[25] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in Proceedings of the 25th ACM International Conference on Multimedia (MM '17), Mountain View, CA, USA, 2017, pp. 795–816.

[26] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in MultiMedia Modeling, ser. Lecture Notes in Computer Science, vol. 9516, Cham: Springer, 2016, pp. 15–27.

[27] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang et al., "LongBench: A bilingual, multitask benchmark for long context understanding," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 2024, pp. 3119–3137.