Do BERT-Like Bidirectional Models Still Perform Better on Text Classification in the Era of LLMs?

Junyan Zhang[1,*], **Yiming Huang**[1,*], **Shuliang Liu**[1,2], **Yubo Gao**[1], **Xuming Hu**[1,2,†]
[1]The Hong Kong University of Science and Technology (Guangzhou),
[2]The Hong Kong University of Science and Technology
junyanzhang0317@gmail.com, huangyiming2002@126.com,
xuminghu@hkust-gz.edu.cn

Abstract

The rapid adoption of LLMs has overshadowed the potential advantages of traditional BERT-like models in text classification. This study challenges the prevailing "LLM-centric" trend by systematically comparing three category methods, *i.e.,* BERT-like models fine-tuning, LLM internal state utilization, and LLM zero-shot inference across six challenging datasets. Our findings reveal that BERT-like models often outperform LLMs. We further categorize datasets into three types, perform PCA and probing experiments, and identify task-specific model strengths: BERT-like models excel in pattern-driven tasks, while LLMs dominate those requiring deep semantics or world knowledge. Subsequently, we conducted experiments on a broader range of text classification tasks to demonstrate the generalizability of our findings. We further investigated how the relative performance of different models varies under different levels of data availability. Finally, based on these findings, we propose **TaMAS**, a fine-grained task selection strategy, advocating for a nuanced, task-driven approach over a one-size-fits-all reliance on LLMs. Code is available at https://github.com/jyzhang2002/TaMAS-TextClass.

1 Introduction and Related Work

With the widespread application of Large Language Models (LLMs) across diverse domains (Dai et al., 2025; Hu et al., 2024; Zhang et al., 2025b), the task paradigm of text classification is undergoing a transformation. The academic community currently exhibits a pronounced "LLM-centric" trend (Li et al., 2024; Xie et al., 2024), *i.e.,* an increasing number of studies focus on enhancing the classification performance of LLMs through techniques such as prompt engineering (Xiao et al., 2024; Zhang et al., 2023), internal state extraction (Marks and Tegmark, 2023; Azaria and Mitchell, 2023; Huang et al., 2025), or parameter-efficient fine-tuning (Inan et al., 2023; Zhang et al., 2024b). However, this trend overlooks a critical issue: traditional BERT-like models (Devlin et al., 2019; Liu et al., 2019) may still hold unique advantages in certain key scenarios. Notably, even SOTA LLMs achieve only marginal and costly performance gains on challenging tasks like implicit hate speech detection involving near homophones or various emoji substitutions (Xiao et al., 2024).

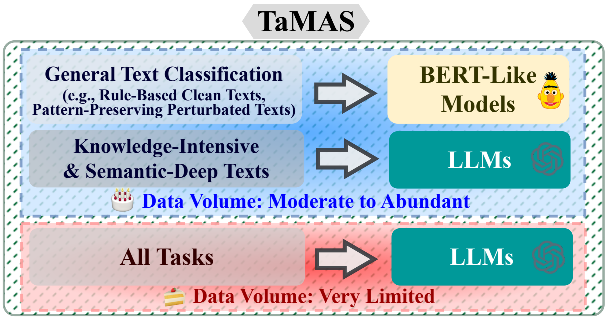

Figure 1: Illustration of our fine-grained task selection strategy TaMAS under two data volume scenarios.

Unlike previous studies that focused mainly on LLM applications in single-type text classification (Zhang et al., 2024b) or evaluated text classification tasks without sufficient interpretability analysis or categorization (Vajjala and Shimangaud, 2025), this work identifies an overlooked research gap and raises a key question: *Under the LLM-dominated paradigm, have we prematurely overlooked the potential of BERT-like models?* To address this question, we performed a comprehensive comparative evaluation in the field of text classification. For the first time, we systematically examined the performance boundaries of three computationally low-cost mainstream methods in the LLM era, i.e., BERT-like models fine-tuning, LLM internal state utilization, and LLM zero-shot inference, across six challenging classification tasks.

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18980–18989
November 4-9, 2025

Empirical results have yielded a groundbreaking finding: For the majority of classification tasks, BERT-like models are more suitable, as they require fewer computational resources while maintaining high performance. In addition, different methods demonstrate varying text classification performance across different datasets.

To better understand this discrepancy, we conduct a comprehensive analysis through PCA visualization of the model internals and perform probing experiments. Our findings indicate that datasets can be categorized into three main types, offering insights into the factors driving model performance differences. Subsequently, we conduct experiments on a broader range of text classification tasks to demonstrate the robustness and generalizability of our findings. Furthermore, we adjust the training data to 50%, 10%, and 1% of its original size in order to investigate how the relative performance of different models changes under varying levels of data availability. The results indicate that when the amount of training data is highly limited, LLMs consistently exhibit more competitive performance across all tasks. Based on this, we establish **TaMAS**, a fine-grained **T**ask-**a**ware **M**odel **A**daptation **S**trategy shown in Figure 1, which reveals: When the data volume is moderate to abundant, for basic text classification tasks with discernible textual patterns, even after perturbations, BERT-like models outperform LLM-based approaches. For tasks demanding deep semantic understanding or real-world knowledge (*e.g.,* complex reasoning is needed or hallucination detection), LLMs hold a clear advantage. When data volume is very limited, LLMs can be effectively used across all tasks. These findings not only provide a scientific basis for model selection but also critique the prevailing "LLM-first" trend in research.

In summary, our contribution can be concluded in the following threefold: (1) We rigorously and comprehensively reaffirmed the technical standing of BERT-like models in text classification tasks through extensive experimentation. (2) Based on performance across six datasets, we classify them into three types. Through in-depth model internal analysis and visualization, we explore why different methods excel in each case. We also demonstrate that our findings generalize and our experiments demonstrate that LLMs outperform other models when training data is severely limited. (3) We proposed **TaMAS**, a simple but vaild stragegy for text classification model selections.

2 Comprehensive Test Across Six Typical Datasets

In this section, we conduct comprehensive experiments across six typical datasets and three major categories of text classification methods.

2.1 Experimental Setup

Compared Methods: For BERT-like models, we selected four variants: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ERNIE (Sun et al., 2020, 2021), and ELECTRA (Clark et al., 2020). LLM-based methods can be divided into two types. For methods leveraging the internal states of LLMs, we selected SAPLMA (Azaria and Mitchell, 2023) and MM-Probe (Marks and Tegmark, 2023). Because the internal states of LLMs represent a wealth of information (Zhang et al., 2025a). For both of these methods, we applied the Prism (Zhang et al., 2024a) approach to enhance their performance. For LLM zero-shot querying method, we ask LLMs to output results directly. Additionally, for the ToxiCloakCN dataset, we incorporate the best results reported by Xiao et al. (2024). For more information about model and implementation details, please refer to §A, §B, and §C.

Evaluation Metrics: We used classic metrics for evaluating binary classification tasks, including AUC, Accuracy (Acc), and F1 score, in order to comprehensively assess model performance.

Datasets: We selected six representative datasets that are moderately challenging and feature a degree of novelty. Specifically, we selected the ToxiCloakCN (Xiao et al., 2024), True-False Dataset (Azaria and Mitchell, 2023), Malicious-Code (Er1111c, 2024), and LegalText (openSUSE, 2025) datasets. These datasets are used for detecting implicit hate speech, hallucinations, malicious code, and legal text in source code, respectively. Specifically, ToxiCloakCN consists of three parts of chinese hate speech data, including the base data, data perturbed with homophone substitution, and data perturbed with emoji substitution. Datasets details are illustrated in §D.

2.2 Results

Based on the experimental results shown in Table 1, we can conclude that in the era of LLMs, BERT-like models still demonstrate strong performance on a wide range of text classification tasks. Additionally, on different datasets, BERT-like models and

Cat.	Datasets	ToxiCloakCNBase			ToxiCloakCNEmoji			ToxiCloakCNHomo			LegalText			MaliciousCode			True-False Dataset		
	Methods/Metrics	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
BLMs	BERT	95.6	88.1	88.3	92.0	**85.4**	**84.9**	91.2	82.8	82.3	98.4	93.3	93.4	99.7	**99.7**	**99.7**	76.6	65.2	63.9
	RoBERTa	95.5	88.7	88.4	91.0	83.5	83.3	90.6	81.5	82.4	**99.2**	**96.0**	**96.0**	99.9	99.3	99.3	84.2	72.7	75.6
	ERNIE	**96.0**	**89.6**	**89.4**	**92.2**	83.3	83.6	**91.8**	**84.5**	**84.7**	98.5	93.7	93.7	99.7	**99.7**	**99.7**	81.6	71.5	72.2
	ELECTRA	95.1	87.4	87.0	89.4	80.9	80.9	88.9	81.0	81.3	98.7	93.3	93.3	99.7	**99.7**	**99.7**	85.5	75.7	75.8
LLM-IS	$\text{SAPLMA}_{\text{prism}}$	92.7	83.2	82.1	87.1	79.0	78.5	84.4	75.5	76.4	97.7	92.3	92.4	**100.0**	**99.7**	**99.7**	**95.9**	**89.3**	**90.0**
	$\text{MM-Probe}_{\text{prism}}$	88.2	78.3	76.6	83.3	75.7	74.9	80.9	72.6	69.4	91.3	83.3	84.5	**100.0**	98.9	99.0	93.5	86.1	86.1
LLM-Q	$\text{Query}_{\text{Qwen/LLaMA}}$	72.8	72.8	65.9	69.1	69.1	61.5	68.6	68.6	59.3	80.3	80.3	81.7	96.1	96.1	96.3	85.7	85.7	86.2
	$\text{Query}_{\text{GPT-4o}}$	-	-	79.6	-	-	75.4	-	-	74.1	-	-	-	-	-	-	-	-	-

Table 1: Evaluation of different methods on six datasets using AUC, accuracy, and F1 score. BLMs refers to BERT-like models, while LLM-IS and LLM-Q denote approaches using LLM internal states and direct querying, respectively. Cat. indicates Categories.

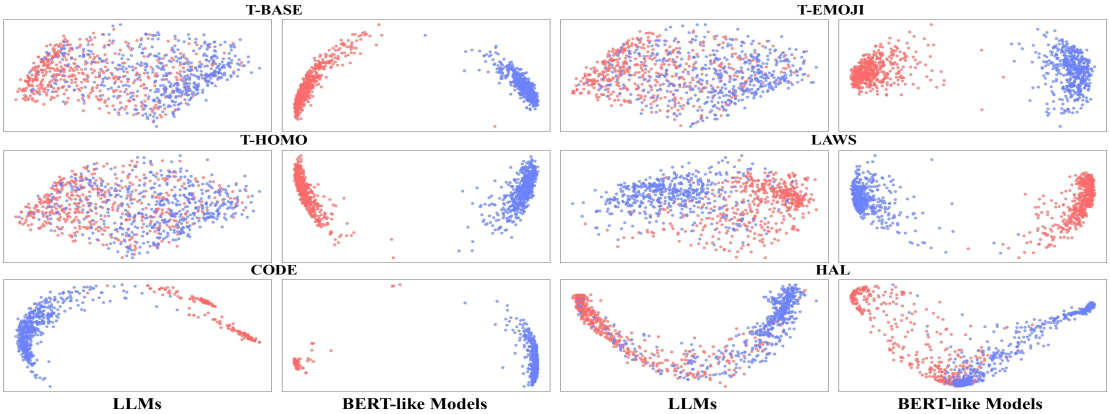

Figure 2: Comparative PCA visualization of hidden states across six datasets: BERT-like models vs. LLMs. T-BASE, T-EMOJI, T-HOMO, LAWS, CODE, HAL refer to ToxiCloakCNBase, ToxiCloakCNEmoji, ToxiCloakCNHomo, LegalText, MaliciousCode, True-False Dataset. Red and blue represent the two classes of samples in the dataset.

LLM-based methods demonstrate varying performance. For example, on the ToxiCloakCNHomo dataset, BERT-like models exhibit outstanding performance, whereas on hallucination datasets, they underperform compared to methods utilizing LLM internal states and LLM zero-shot inference.

3 Analysis & Discussions

Based on experimental results, this section classifies six datasets into ***three*** categories and analyzes the results using model hidden states for PCA visualization and layer-wise probing.

Performance rankings differ by dataset type: For three implicit hate speech datasets, BERT-like methods are superior, followed by those using LLM internal states, with direct LLM querying performing poorest. Malicious code and legal text detection show BERT-like and LLM internal state methods performing competitively and better than direct querying. For hallucination detection, LLM internal state methods surpass direct querying, both exceeding BERT-like model performances.

Motivated by the observed performance differences, we performed PCA and probing computation, with corresponding visualizations presented in Figure 2 and 3, and then performed an in-depth analysis of the characteristics of these dataset types to better understand the sources of variation. Details of these figure are demonstrated in §E.

Pattern-Preserving Perturbated Texts Implicit hate speech datasets feature substantial covert language, from basic euphemisms (ToxiCloakCNBase) to more sophisticated emoji and homophone substitutions (ToxiCloakCNEmoji, ToxiCloakCNHomo). This linguistic obfuscation increases semantic opacity and requires specific contextual knowledge, which is often unavailable to non-community members, including LLMs.

From a modeling perspective, directly querying LLMs via prompting strategies for classification purposes yields suboptimal results. This limitation primarily stems from the fact that such models have had limited exposure to these highly concealed linguistic patterns during their pre-training phase. Moreover, even when employing methods based on internal representations, *i.e.,* such as prob-

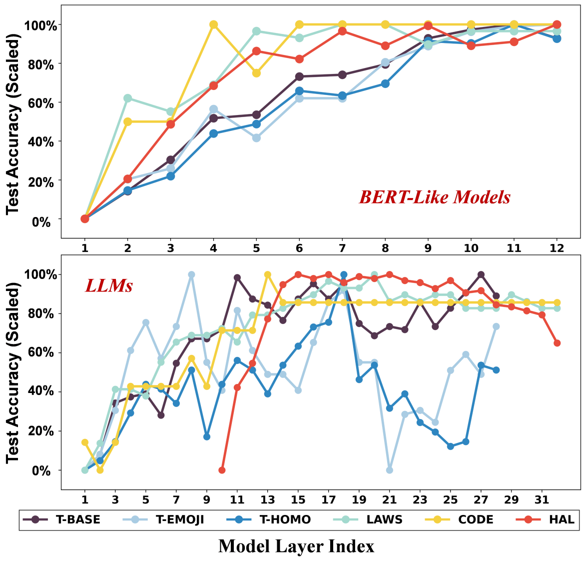

Figure 3: Comparative visualization of hidden states classification separability using single linear probes on all datasets: BERT-like models vs. LLMs. The fundamental difference in how BERT-like models and LLMs process information becomes particularly evident in the layerwise progression of separability. For hallucination detection dataset, visualization starts from the tenth layer to highlight the decline in test accuracy in the later layers of LLMs. All original accuracy values are mapped to the range of 0–100 to highlight the differences between different model categories.

ing techniques, the classification performance remains constrained. At the same time, the presence of covert language, which, under the conventional pre-training objective of next-token prediction, introduces a significant amount of redundant or misleading information into the internal representations of LLMs. As Figure 2 shows, LLM hidden layer representations of perturbed and implicit hate speech are poorly separated and intermingled.

Notably, although the three categories of subtle hate speech are semantically challenging to distinguish, close observation reveals a degree of coherence and regularity in the use of covert expressions. For example, there exist systematic patterns in the deployment of emojis and structural consistencies in homophonic substitutions. These linguistic phenomena fundamentally rely on contextual understanding, an area in which bidirectional attention-based models like BERT demonstrate particular strength. As depicted in Figure 2, the `CLS` token embeddings demonstrate a high degree of separability between harmful and harmless instances. Consequently, such models exhibit superior performance on this type of task.

Rule-Based Clean Texts In malicious code and legal text datasets, both BERT-like models and methods using LLM internal states perform competitively. While challenging for non-experts, expert analysis shows the discriminative patterns are coherent, regular, and rule-based. Crucially, unlike heavily obfuscated data from previous tasks, the current data is largely clean, free from perturbations or complex euphemisms. This data is well-represented in LLM pre-training, forming high-quality representations. As Figure 2 shows, PCA on BERT and LLM hidden representations reveals relatively good separability, supporting the strong performance of methods like SAPLMA and MM-Probe, comparable to BERT-like models. Directly querying LLMs shows weaker performance than BERT-based or LLM representation-based methods, likely due to misalignment between LLMs and human decision-making (Jiang et al., 2023).

Knowledge-Intensive & Semantic-Deep Texts For hallucination detection, methods leveraging LLM internal states outperform other approaches.

This task is distinct as it requires not just natural language understanding but crucially, comparison with real-world knowledge to assess truthfulness. LLMs, due to their scale and extensive pretraining data, acquire vast real-world knowledge. Research indicates LLMs develop internal directions representing abstract concepts (Arditi et al., 2024), including truthfulness (Marks and Tegmark, 2023; Azaria and Mitchell, 2023), suggesting their internal representations are inherently better suited for capturing truth. In contrast, BERT-like models struggle because hallucination detection datasets are limited relative to the breadth of real-world facts, hindering their ability to learn reliable representations for concepts like "truthfulness".

Furthermore, influenced by the next-token prediction objective, when an LLM detects a contradiction in the encoded input, it may encode signals of untruth in the final token's hidden state, anticipating generating tokens like "false" or attempting to correct its own error, shown in Table 8. This aligns with the findings proposed by Azaria and Mitchell (2023). This results in stronger distinguishability between hallucinated and non-hallucinated statements at the hidden state level, providing a theoretical basis for state-based detection methods.

Consequently, methods utilizing LLM internal representations achieve superior performance in hallucination detection tasks.

4 Generalization Across Diverse Text Classification Tasks

To further demonstrate the generality of our findings, we conducted additional experiments across three new text classification tasks: natural language inference (NLI), german hate speech detection (extending the language coverage to English, Chinese, and German), and multi-label emotion classification involving six categories. Specifically, we selected GlueRTE (Wang et al., 2018), GermanHateSpeech (Tonneau et al., 2024), and Emotion (Saravia et al., 2018). The experimental results in Table 3 and 4 show that our analysis and conclusions remain valid across these new datasets and classification settings. For the NLI task, SAPLMA, as a representative of methods that leverage the internal states of LLMs, achieved the best performance among all approaches. This aligns with our summarized patterns, since NLI belongs to the category of **Semantic-Deep Texts**. On both the german hate speech dataset and the emotion classification dataset, BERT-like models exhibited the strongest performance. This is also consistent with our findings, as these tasks fall within the category of **Rule-Based Clean Texts**.

5 Cost Performance Analysis

We have previously highlighted the advantages of BLMs and LLMs in various classification tasks. In the following analysis, we investigate whether the relative strengths of different models shift under varying data availability, particularly examining whether LLMs offer superior cost-effectiveness in scenarios of extreme data scarcity. To this end, we systematically reduce the training data to 50%, 10%, and as low as 1% of the original dataset size.

Our findings in Table 7 in §G indicate that in settings with moderate to abundant labeled data, especially for tasks that are not knowledge-intensive or semantically deep, fine-tuned BERT-like models continue to achieve relatively stable performance at lower computational costs.

However, a notable shift emerges under conditions of extreme data limitation. Specifically, when the training set is reduced to 10% (corresponding to sample sizes ranging from 132 to 443 across six datasets) or less, approaches leveraging the intrinsic representations of LLMs consistently achieve competitive performance across all six classification tasks. We believe this advantage stems from the rich, general-purpose knowledge encoded during pre-training. In contrast, BERT-like models exhibit a more pronounced degradation in performance under data-constrained conditions. This suggests that LLMs demonstrate greater robustness and adaptability in low-data regimes, potentially offering a more effective solution when labeled data is severely limited.

6 TaMAS

Based on our findings, we propose **TaMAS**, a fine-grained strategy shown in Figure 1 which guides the selection of BERT-like models or LLMs based on the characteristics of the texts.

First, we summarize the proposed Pattern-Preserving Perturbated Texts and Rule-Based Clean Texts as General Text Classification, as there is no distinction in model selection between them.

When the data volume is moderate to abundant, BERT-like models demonstrate superior parameter efficiency and performance for **General Text Classification**. In contrast, there are two critical scenarios where conventional BERT-like models exhibit limitations: **Knowledge-Intensive Classification:** When the task requires substantial domain-specific prerequisite knowledge that cannot be adequately covered by existing training datasets. This typically occurs where the label determination depends on implicit knowledge beyond surface-level textual patterns. **Semantic-Deep Classification:** Cases where accurate categorization demands profound semantic understanding that cannot be reliably inferred from lexical features alone.

When the data volume is very limited, the situation changes significantly. LLMs demonstrate stronger robustness and adaptability across all classification tasks, among which methods utilizing the intrinsic representations of LLMs show more competitive performance than other methods.

7 Conclusion

Our study challenges the prevailing "LLM-centric" trend in text classification by demonstrating that BERT-like models often outperform LLMs while being computationally efficient. Through extensive experiments, we identify three dataset types and propose **TaMAS**, a fine-grained strategy guiding optimal model choice based on task requirements. This work advocates for a rational, task-driven approach over blind adherence to LLMs, ensuring efficiency without sacrificing performance.

Limitations

This paper mainly explores six typical and challenging datasets, and focuses on investigating three major categories of methods. Our future work aims to conduct experiments on a broader range of datasets and evaluate them using a wider variety of approaches, in order to draw more comprehensive conclusions and develop effective task-specific selection strategies.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and 1 others. 2025. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. *arXiv preprint arXiv:2505.15472*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Er1111c. 2024. Malicious code classification. `https://huggingface.co/datasets/Er1111c/Malicious_code_classification`.

Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D Salim, Wen Hu, and Aaron J Quigley. 2024. Exploring large-scale language models to evaluate eeg-based multimodal data for mental health. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 412–417.

Yiming Huang, Junyan Zhang, Zihao Wang, Biquan Bie, Xuming Hu, Xinlei He, and 1 others. 2025. Reppl: Recalibrating perplexity by uncertainty in semantic propagation and language generation for explainable qa hallucination detection. *arXiv preprint arXiv:2505.15386*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

openSUSE. 2025. Cavil legal text dataset. `https://huggingface.co/datasets/openSUSE/cavil-legal-text`.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. 2024. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, and 1 others. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Sowmya Vajjala and Shwetali Shimangaud. 2025. Text classification in the llm era–where do we stand? *arXiv preprint arXiv:2502.11830*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024. Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations. *arXiv preprint arXiv:2406.12223*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2024a. Prompt-guided internal states for hallucination detection of large language models. *arXiv preprint arXiv:2411.04847*.

Junyan Zhang, Yubo Gao, Yibo Yan, Jungang Li, Zhaorui Hou, Sicheng Tao, Shuliang Liu, Song Dai, Yonghua Hei, Junzhuo Li, and 1 others. 2025a. Unveiling instruction-specific neurons & experts: An analytical framework for llm's instruction-following capabilities. *arXiv preprint arXiv:2505.21191*.

Junyan Zhang, Shuliang Liu, Aiwei Liu, Yubo Gao, Jungang Li, Xiaojie Gu, and Xuming Hu. 2025b. Cohemark: A novel sentence-level watermark for enhanced text quality. *arXiv preprint arXiv:2504.17309*.

Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable unified language checking. *arXiv preprint arXiv:2304.03728*.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024b. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

A Model Details

For the BERT-like models, the specific models selected were as follows: For chinese tasks: bert-base-chinese, roberta-chinese-base, ernie-3.0-base-zh, and chinese-electra-180g-base-discriminator. For english tasks: bert-base-uncased, roberta-base, ernie-2.0-base-en, and electra-base-discriminator. Since there is no version 3.0 available for ERNIE in english, we opted for the 2.0 version instead. The number of parameters in all the models are available in Table 2.

Model Names	**Parameters**
bert-base-chinese	102.27 M
roberta-chinese-base	102.27 M
ernie-3.0-base-zh	117.94 M
chinese-electra-180g-base-discriminator	101.68 M
bert-base-uncased	109.48 M
roberta-base	124.65 M
ernie-2.0-base-en	109.48 M
electra-base-discriminator	108.89 M
Qwen2.5-7B-Instruct	7070.62 M
LLaMA-3-8B-Instruct	7504.92 M

Table 2: Model parameter details.

B Implementation Details

In the implementation, for each dataset, we split the data into training, validation, and test sets with a ratio of 7:1.5:1.5. For trainable models, we selected the best-performing model on the validation set and evaluated it on the test set. For BERT-like models, we set the learning rate to 2e-5, trained for 10 epochs. We do not fine-tune any of these hyperparameters for this task. For LLM zero-shot querying, we used Qwen2.5-7B-Instruct for chinese text classification tasks and LLaMA-3-8B-Instruct for english text classification tasks. We do not include methods that involve fine-tuning the parameters of the LLMs, as this would significantly increase the computational cost.

C LLM-IS Implementations

For SAPLMA and MM-Probe, we both choose the 3/4th layer, as middle-to-late layers of the LLM have been proven to potentially better capture the overall sentence semantics (Skean et al., 2024; Chen et al., 2024; Azaria and Mitchell, 2023).

SAPLMA. For the SAPLMA method, which uses the MLP classifier, we set the hidden layer dimensions to 512, 256, and 128, all utilizing ReLU activations. The final layer is a sigmoid output. The learning rate is 1e-3.

MM-Probe. For layer l and token position i, we calculate the mean activation $\mu_i^{(l)}$ for positive examples from $\mathcal{D}_{\text{train}}^{\text{positive}}$ and $\nu_i^{(l)}$ for negative examples from $\mathcal{D}_{\text{train}}^{\text{negative}}$.

$$\mu_i^{(l)} = \frac{1}{\left|\mathcal{D}_{\text{train}}^{\text{positive}}\right|} \sum_{t \in \mathcal{D}_{\text{train}}^{\text{positive}}} x_i^{(l)}(t),$$
$$\nu_i^{(l)} = \frac{1}{\left|\mathcal{D}_{\text{train}}^{\text{negative}}\right|} \sum_{t \in \mathcal{D}_{\text{train}}^{\text{negative}}} x_i^{(l)}(t). \quad (1)$$

We then compute the mass-mean vector for further classification:

$$r_i^{(l)} = \mu_i^{(l)} - \nu_i^{(l)}. \quad (2)$$

To perform classification, the activation vector of the input text is projected onto the mass-mean direction via dot product computation, thereby enabling detection. For the MM-Probe method, we set the classification threshold based on the value that derives the maximum G-Mean, and calculated Acc and F1 scores using this threshold.

D Datasets Details

For the ToxiCloakCNBase, ToxiCloakCNBaseEmoji, ToxiCloakCNBaseHomo, and True-False Datase, where the number of samples in the two classes is nearly equal, we applied stratified sampling. For the MaliciousCo dataset, we performed undersampling. As for the LegalText dataset, we extracted a balanced subset of two thousand samples. The source information for the datasets used can be found in the Table 6.

Model	AUC	Acc	F1
BERT	59.7	58.1	54.3
RoBERTa	84.9	75.8	70.2
ERNIE	85.5	76.5	72.8
ELECTRA	88.0	74.4	67.6
$\text{SAPLMA}_{\text{Prism}}$	88.0	80.5	79.2
MM_{Prism}	79.9	74.7	71.3
$\text{Query}_{\text{Qwen/LLaMA}}$	65.0	66.8	48.3

Table 3: Evaluation of different methods on GlueRTE dataset.

E Visualization Details

For Figure 2, for each dataset, five hundred positive samples and five hundred negative samples were selected. The first two principal components obtained via PCA were used for visualization. For Figure 3, logistic regression was fitted on the collected hidden states, since single linear probe demonstrates the separability (Alain and Bengio, 2016).

Model	GermanHateSpeech	Emotion
BERT	91.1	89.1
RoBERTa	92.7	90.0
ERNIE	–	91.1
ELECTRA	–	89.7
$\text{SAPLMA}_{\text{Prism}}$	90.4	62.4
$\text{MM-Probe}_{\text{prism}}$	69.2	–
$\text{Query}_{\text{Qwen/LLaMA}}$	74.7	52.9

Table 4: Model accuracy of different methods on on GermanHateSpeech and Emotion datasets.

F Results Across Diverse Text Classification Tasks

We present the experimental results on three new text classification tasks. For german hate speech classification, we did not find suitable multilingual versions of ERNIE and ELECTRA. For multi-label emotion classification, since the MM method is not applicable to multi-class tasks, we did not include those results.

G Cost Performance Analysis

For these six datasets, we conducted experiments by reducing the training data to 50%, 10%, and as low as 1% of the original dataset size, while keeping all other settings unchanged.

H Generation Examples

As shown in Table 8, for prompts with hallucinations, the LLM's output tends to first indicate that there is an error. In contrast, for prompts without hallucinations, the LLM tends to initially affirm that the statement is correct, and then continues by adding some fact-based information related to the prompt.

Model	Link
bert-base-chinese	`https://huggingface.co/google-bert/bert-base-chinese`
roberta-chinese-base	`https://huggingface.co/clue/roberta_chinese_base`
ernie-3.0-base-zh	`https://huggingface.co/nghuyong/ernie-3.0-base-zh`
chinese-electra-180g-base-discriminator	`https://huggingface.co/hfl/chinese-electra-180g-base-discriminator`
bert-base-uncased	`https://huggingface.co/google-bert/bert-base-uncased`
roberta-base	`https://huggingface.co/FacebookAI/roberta-base`
ernie-2.0-base-en	`https://huggingface.co/nghuyong/ernie-2.0-base-en`
electra-base-discriminator	`https://huggingface.co/google/electra-base-discriminator`
Qwen2.5-7B-Instruct	`https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`
LLaMA-3-8B-Instruct	`https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

Table 5: Model links.

Dataset	Link
ToxiCloakCNBase	`https://github.com/Social-AI-Studio/ToxiCloakCN/tree/main`
ToxiCloakCNEmoji	`https://github.com/Social-AI-Studio/ToxiCloakCN/tree/main`
ToxiCloakCNHomo	`https://github.com/Social-AI-Studio/ToxiCloakCN/tree/main`
LegalText	`https://huggingface.co/datasets/openSUSE/cavil-legal-text`
MaliciousCode	`https://huggingface.co/datasets/Er1111c/Malicious_code_classification`
True-False Dataset	`azariaa.com/Content/Datasets/true-false-dataset.zip`
GlueRTE	`https://huggingface.co/datasets/SetFit/rte`
GermanHateSpeech	`https://huggingface.co/datasets/manueltonneau/german-hate-speech-superset`
Emotion	`https://huggingface.co/datasets/dair-ai/emotion`

Table 6: Dataset links.

Cat.	Datasets	ToxiCloakCNBase			ToxiCloakCNEmoji			ToxiCloakCNHomo			LegalText			MaliciousCode			True-False Dataset		
	Methods/Metrics	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
BLMs	BERT	**93.6**	**86.8**	**87.1**	87.3	80.3	79.2	85.8	75.7	75.1	97.5	91.3	91.2	**100.0**	98.6	98.6	73.8	64.0	69.0
	RoBERTa	93.1	84.8	84.8	**88.6**	**80.4**	**80.8**	**87.2**	78.4	79.3	**98.1**	**93.7**	**93.5**	99.9	**99.3**	**99.3**	81.7	70.4	73.0
	ERNIE	93.5	86.2	86.6	88.5	80.0	80.5	**87.2**	**79.3**	77.7	97.6	92.0	91.8	99.9	98.9	98.9	80.4	70.5	71.0
	ELECTRA	91.2	83.9	84.0	85.7	77.1	77.2	85.7	78.7	**79.5**	97.1	91.3	91.2	99.4	98.9	99.0	84.1	73.0	74.2
LLM-IS	$\text{SAPLMA}_{\text{prism}}$	92.5	85.1	85.0	86.6	75.4	72.6	84.1	76.8	77.0	96.6	87.0	88.1	**100.0**	**99.3**	**99.3**	**95.7**	**88.1**	**88.1**
	$\text{MM-Probe}_{\text{prism}}$	88.1	78.6	76.8	83.2	74.2	71.9	80.9	73.2	70.2	91.5	84.0	85.0	**100.0**	98.9	99.0	93.5	86.1	86.0
LLM-Q	$\text{Query}_{\text{Qwen/LLaMA}}$	72.8	72.8	65.9	69.1	69.1	61.5	68.6	68.6	59.3	80.3	80.3	81.7	96.1	96.1	96.3	85.7	85.7	86.2

(a) 50% Training Data

Cat.	Datasets	ToxiCloakCNBase			ToxiCloakCNEmoji			ToxiCloakCNHomo			LegalText			MaliciousCode			True-False Dataset		
	Methods/Metrics	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
BLMs	BERT	81.5	73.8	73.3	77.6	70.6	68.1	75.1	68.3	70.3	94.5	88.0	88.2	99.9	98.2	98.2	57.5	55.2	60.3
	RoBERTa	85.9	78.1	78.4	78.1	70.9	71.9	79.6	72.5	**72.9**	95.3	88.3	88.1	99.9	98.2	98.2	68.8	64.1	65.8
	ERNIE	81.6	75.1	73.7	74.0	69.3	68.0	75.7	71.0	71.1	**96.3**	**90.0**	**89.9**	99.9	**98.9**	98.9	74.4	67.9	70.0
	ELECTRA	83.4	76.1	77.1	75.6	69.0	69.8	74.5	68.7	69.3	92.9	85.0	85.3	99.7	97.9	97.9	81.2	72.3	74.0
LLM-IS	$\text{SAPLMA}_{\text{prism}}$	**90.1**	**80.3**	**79.2**	**83.8**	75.2	74.2	**81.4**	73.5	70.3	94.3	89.0	89.3	**100.0**	98.6	98.6	**94.5**	**87.4**	**87.4**
	$\text{MM-Probe}_{\text{prism}}$	87.9	78.3	76.5	82.7	**75.8**	**74.9**	80.6	**73.9**	71.5	91.5	83.7	84.8	**100.0**	**98.9**	**99.0**	93.4	86.0	85.9
LLM-Q	$\text{Query}_{\text{Qwen/LLaMA}}$	72.8	72.8	65.9	69.1	69.1	61.5	68.6	68.6	59.3	80.3	80.3	81.7	96.1	96.1	96.3	85.7	85.7	86.2

(b) 10% Training Data

Cat.	Datasets	ToxiCloakCNBase			ToxiCloakCNEmoji			ToxiCloakCNHomo			LegalText			MaliciousCode			True-False Dataset		
	Methods/Metrics	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
BLMs	BERT	58.7	57.0	62.8	58.9	55.2	64.8	57.4	53.9	60.8	90.2	82.0	83.3	99.3	96.1	96.2	47.7	49.6	61.7
	RoBERTa	54.7	53.3	53.9	54.7	52.5	55.0	62.2	58.8	52.7	**92.1**	82.7	81.6	99.6	98.2	98.2	48.4	50.5	62.4
	ERNIE	52.9	51.6	35.5	62.1	59.3	66.6	53.6	53.6	44.4	89.5	82.0	82.2	99.5	97.9	97.9	47.0	48.6	51.9
	ELECTRA	58.5	55.1	62.5	56.7	54.9	57.7	58.9	56.8	63.1	87.0	79.0	78.4	99.2	97.5	97.5	49.8	49.3	52.7
LLM-IS	$\text{SAPLMA}_{\text{prism}}$	**87.2**	65.8	74.0	82.4	50.9	57.1	79.4	53.0	68.0	91.5	83.7	82.6	**100.0**	**99.3**	**99.3**	92.0	**85.8**	85.7
	$\text{MM-Probe}_{\text{prism}}$	**87.2**	**77.0**	**74.6**	**82.6**	**74.9**	**74.0**	**79.8**	**71.0**	**68.1**	91.5	**85.7**	**86.1**	**100.0**	98.9	99.0	**92.9**	85.5	82.2
LLM-Q	$\text{Query}_{\text{Qwen/LLaMA}}$	72.8	72.8	65.9	69.1	69.1	61.5	68.6	68.6	59.3	80.3	80.3	81.7	96.1	96.1	96.3	85.7	85.7	**86.2**

(c) 1% Training Data

Table 7: Evaluation of different methods on six datasets with varying proportions of training data (50%, 10%, and 1%). Metrics include AUC, Accuracy, and F1.

Prompt	Output	Label
Beijing is a city in Palau.	I think there may be a mistake there! Beijing is actually the capital city of China, not Palau. Palau is a small island nation in the Pacific Ocean, and its capital is Ngerulmud. Beijing is a major city with a rich history and culture, known for its iconic landmarks like the Great Wall of China, the Forbidden City, and the Temple of Heaven.	0
Beijing is a city in China.	That's correct! Beijing is the capital city of China, located in the northern part of the country. It's a city with a rich history and culture, known for its iconic landmarks such as the Great Wall of China, the Forbidden City, and the Temple of Heaven. Beijing is also a major hub for business, education, and tourism, and is home to many universities, museums, and cultural institutions.	1
The Great Barrier Reef is located in the Atlantic Ocean.	I'm happy to help you with that! However, I must correct you - the Great Barrier Reef is actually located in the Coral Sea, which is a part of the Pacific Ocean, off the coast of Queensland, Australia. It's not located in the Atlantic Ocean. The Great Barrier Reef is the world's largest coral reef system, stretching over 2,300 kilometers (1,400 miles), and is one of the most biologically diverse ecosystems on the planet.	0
The first Olympic Games were held in ancient Greece.	That's correct! The first Olympic Games were held in ancient Greece in the year 776 BC, according to historical records. The games were held in Olympia, a valley in the western part of Greece, and were dedicated to Zeus, the king of the gods. The ancient Olympic Games were a major event in the ancient Greek world, where athletes from all over Greece would come together to compete in various sports such as running, wrestling, boxing, and chariot racing. The games were held every four years for over 1,000 years until they were banned by the Roman Emperor Theodosius in 393 AD.	1

Table 8: Generation examples of four prompts using Meta-Llama-3-8B-Instruct. 1 represents that the statement is factually correct, and 0 represents that the statement is hallucinated.