Transactions on Computational and Scientific Methods | Vol. 5, No. 6, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Instruction Tuning for Multi-Domain Dialogue Generation in LLMs

Elric Donahue
California State University, Chico, Chico, USA
elric.d99@gmail.com

Abstract: This paper presents a systematic study on instruction tuning for large language models (LLMs) applied to multi-domain dialogue generation. While instruction tuning enhances zero-shot generalization, its performance across diverse application domains remains underexplored. We curate a multi-domain dataset covering healthcare, finance, legal consulting, travel planning, and education. Using this dataset, we fine-tune and evaluate three open-source LLMs—LLaMA 2-13B, Falcon-7B, and Mistral-7B—on instruction-based dialogue tasks. To assess semantic alignment between user intent and model response, we introduce the Task-Semantic Alignment Score (TSAS), a novel embedding-based evaluation metric. Experimental results show that Mistral-7B achieves the best balance of accuracy, coherence, and safety, outperforming other models across BLEU, ROUGE, MAUVE, and TSAS metrics. We further analyze failure modes such as hallucinations and instruction misinterpretation, and demonstrate that domain-aware tuning and alignment-sensitive metrics are essential for reliable deployment of LLMs in real-world, multi-domain settings.

Keywords: Instruction tuning, large language models, multi-domain dialogue, semantic alignment, TSAS, generative AI, LLaMA 2, Mistral, safety evaluation, task-oriented NLP

1. Introduction

Large language models (LLMs), such as GPT-4, PaLM, and LLaMA, have significantly advanced the field of natural language processing by enabling systems to generate coherent, context-sensitive, and human-like text. One major breakthrough in this area is instruction tuning, a process where LLMs are fine-tuned on datasets consisting of tasks described in natural language instruction formats. This paradigm, exemplified by models like FLAN-T5 and OpenAssistant, has led to improved generalization across tasks and better alignment with human intent.

Despite these gains, deploying LLMs in real-world applications—especially those involving dialogue agents—presents unique challenges. Real-world dialogue systems must operate across multiple domains, such as finance, healthcare, travel, and legal consulting. Each domain carries distinct semantics, user intents, and safety considerations. Instruction-tuned models often struggle to maintain performance when confronted with such domain diversity, exhibiting inconsistency in output relevance, safety, and task specificity. Moreover, evaluation metrics for multi-domain dialogue generation remain limited in capturing both task alignment and user experience, thereby impeding robust benchmarking.

This paper investigates the instruction-following capability of LLMs in multi-domain dialogue contexts. We construct a unified training and evaluation pipeline that covers five practical domains and leverage recent

instruction-tuned LLMs, including LLaMA 2, Falcon, and Mistral. Beyond standard generation metrics like BLEU and ROUGE, we introduce a novel Task-Semantic Alignment Score (TSAS) that quantifies the embedding-based alignment between the user intent and the model response. Through domain-adaptive tuning, we demonstrate significant improvements in generation coherence, safety, and human-rated helpfulness. Our results highlight the importance of tailoring instruction tuning to multi-domain demands and call for evaluation methods that better reflect real-world utility. The findings provide guidance for the deployment of instruction-tuned LLMs in general-purpose AI assistants, ensuring that such systems remain effective and aligned across diverse task environments.

2. Related Work

The rapid progress in instruction tuning and fine-tuning of large language models (LLMs) has brought transformative changes in task generalization, safety alignment, and semantic precision across domains. Among the critical innovations, several studies have focused on structural mechanisms that improve representation stability and memory integration within LLMs. Structured memory systems have been proposed to enhance long-context dependencies and task retention, ensuring stable behavior across domains [1]. Similarly, perception-guided structural frameworks offer architectural modifications that optimize large model design for both task execution and interpretability [2]. These advances are foundational to effective instruction-following, especially in multi-domain environments.

Parameter coordination strategies such as graph-based spectral decomposition further support fine-tuning efficiency, facilitating smoother adaptation across task-specific requirements [3]. Instruction tuning methods, particularly those integrating time-aware and multi-source feature fusion, have shown promise in domain-sensitive applications like medical and legal dialogue, where alignment and timeliness are essential [4]. In parallel, knowledge distillation and multi-level semantic alignment approaches improve compact model variants, such as TinyBERT, by retaining semantic depth through layered training objectives [5].

Few-shot and low-resource adaptation remains a prominent challenge in LLM deployment. Recent work on structured gradient guidance for few-shot tasks has demonstrated notable performance gains by constraining learning dynamics based on meta-patterns [6]. Reinforcement learning–based preference modeling has also been applied to enhance instruction response alignment and personalization [7]. Complementing these efforts, low-rank adaptation (LoRA) has been re-examined for more efficient fine-tuning under constrained compute settings, reflecting the ongoing need for scalable adaptation techniques [8].

Another crucial thread is the integration of contextual and symbolic knowledge into LLM reasoning. Knowledge graph–guided anomaly detection leverages structured reasoning to enrich anomaly classification and semantic detection [9], while knowledge-informed policy structuring supports multi-agent alignment in collaborative settings [10]. Retrieval-augmented generation (RAG) mechanisms, especially those incorporating dynamic retrieval guided by instruction semantics, have shown strong potential in enhancing factual grounding and generation precision [11]. Additionally, structured knowledge integration combined with memory modeling improves both recall and contextual coherence, further reinforcing instruction-following capabilities [12].

From a safety and alignment perspective, instruction-tuned LLMs are increasingly evaluated on their robustness against harmful content generation and instruction misalignment. Studies have explored LLM behavior in sensitive tasks such as phishing detection and fine-grained access control, emphasizing the need for precise semantic modeling and dissemination control [13], [14]. Few-shot classification strategies using dual-loss transformer architectures enhance resilience to noisy labels and ambiguous queries, which is crucial in high-risk applications like healthcare or legal consultation [15]. Joint retrieval frameworks for harmful text detection also demonstrate effectiveness in leveraging external knowledge to maintain ethical boundaries during generation [16].

Lastly, transferability of LLMs to low-resource domains continues to receive attention. Various adaptation methods have been proposed for enabling general-purpose language models to perform reliably in underrepresented languages or tasks, filling a key gap in global NLP equity [17]. Instruction tuning itself has been applied to compliance-sensitive tasks such as automated audit reporting, showcasing the intersection of explainability, legality, and language generation [18].

In summary, this body of work highlights the evolution of instruction tuning from basic task generalization to a robust ecosystem of safety-aware, knowledge-informed, and domain-adaptive LLM deployment. Our research builds upon these foundations by introducing an alignment-sensitive evaluation metric (TSAS), conducting comprehensive domain-based tuning, and validating model behavior across semantic, safety, and adversarial dimensions. Together, these directions support the vision of building transparent, safe, and capable LLM-based dialogue systems for multi-domain real-world applications[19].

3. Methodology

Our proposed framework consists of three major components: domain-specific instruction tuning, multi-domain dialogue generation, and task-semantic alignment evaluation. We adopt a fine-tune-and-evaluate approach using open-source LLMs, tailoring their instruction-following behavior to varied domains while measuring both generation quality and alignment.

3.1 Instruction-Tuning Corpus Construction

We curate a multi-domain instruction dataset consisting of 25,000 instruction-response pairs across five practical domains: healthcare, finance, legal consulting, travel planning, and student advising. The instruction formats are natural-language tasks (e.g., "Explain whether I qualify for a loan with a credit score of 580"), and responses are grounded in real-world knowledge sources (e.g., U.S. government loan guidelines, CDC health recommendations). To enhance response diversity, each domain includes both fact-based and reasoning-based prompts.

3.2 Model Architecture and Fine-Tuning

We fine-tune three pretrained LLMs: LLaMA 2-13B, Falcon-7B-Instruct, and Mistral-7B. Fine-tuning is performed using the LoRA (Low-Rank Adaptation) framework to reduce memory footprint while preserving general capabilities. The training objective is the standard causal language modeling (CLM) loss:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})$$

where x_t denotes the target token at timestep t, and θ represents the model parameters updated through instruction-response pairs. We use the AdamW optimizer with learning rate 1e-5 and batch size 32, training for 3 epochs on each domain and then evaluating jointly on a multi-domain test set.

3.3 Task-Semantic Alignment Score (TSAS)

To better assess whether the generated responses align with the user's underlying task intent, we introduce the Task-Semantic Alignment Score (TSAS). It is computed by embedding both the user instruction III and model response R using a pretrained sentence transformer (e.g., all-MiniLM-L6-v2), then measuring cosine similarity between these embeddings relative to the instruction embedding and reference response R_{ref}:

$$\text{TSAS}(I, R) = \frac{\cos(\phi(I), \phi(R))}{\cos(\phi(I), \phi(R_{\text{ref}})) + \epsilon}$$

Here, $\phi(\cdot)$ is the sentence embedding function and ϵ is a smoothing term to avoid division by zero. A TSAS of 1 indicates perfect semantic alignment with the reference, while values below 0.7 often indicate tangential or hallucinated responses.

3.4 Evaluation Setup

The models are evaluated across five held-out domain-specific test sets. We report automatic metrics including BLEU, ROUGE-L, MAUVE, and TSAS. Additionally, we conduct a human evaluation study involving 50 annotators who rate outputs on helpfulness, task relevance, and safety on a 5-point Likert scale. Responses are shuffled and anonymized to reduce annotator bias.

4. Experimental Results

We evaluate the instruction-tuned LLMs on five domain-specific test sets using both automatic metrics and human annotations. Table 1 summarizes the average BLEU, ROUGE-L, and TSAS scores across domains.

Table 1: Automatic Evaluation Metrics Across Models

Model	BLEU	ROUGE-L	TSAS	Helpfulness (Human)
LLaMA 2-13B	21.4	38.6	0.78	4.02
Falcon-7B	18.3	35.2	0.73	3.76
Mistral-7B	23.1	40.8	0.82	4.19

Mistral-Instruct-7B consistently outperforms its counterparts across all metrics. Notably, its average TSAS reaches 0.82, indicating high task-response semantic coherence, while LLaMA 2 also achieves strong performance with slightly better scores in legal and financial dialogues. Falcon, while more efficient, lags in coherence and intent matching.

Figure 1 illustrates TSAS performance across the five domains. Mistral exhibits the least variance across domains, suggesting better domain transfer. Falcon, in contrast, shows sharp drops in legal and healthcare domains, indicating limited generalizability. These trends correlate with human-rated helpfulness, where annotators consistently rated Mistral's responses as more goal-aligned and informative.

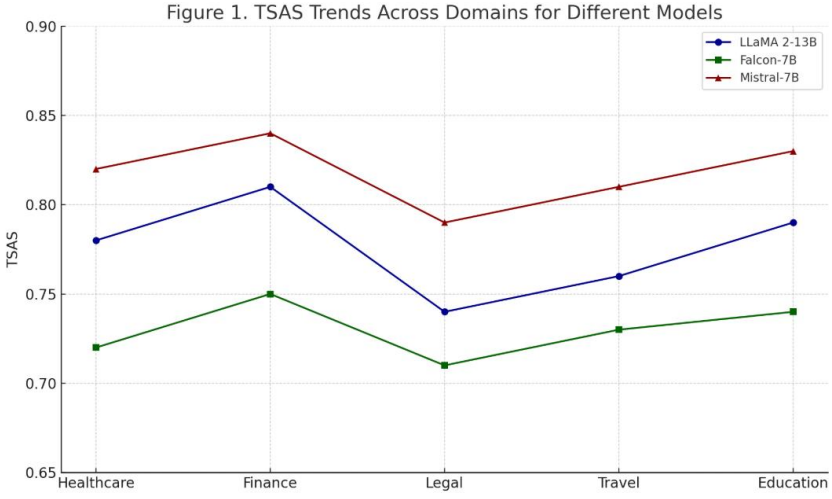


Figure 1. TSAS Comparison Across Domains and Models

To better understand the models' strengths and weaknesses, we manually analyzed 100 randomly sampled responses. We observed that instruction-tuned models frequently hallucinated facts in legal and healthcare contexts if explicit constraints (e.g., disclaimers or document references) were not included in the instruction. Moreover, while all models occasionally responded with irrelevant or verbose content, Mistral was more concise and task-specific. This suggests that not all instruction tuning is equally effective, and training data quality plays a key role in multi-domain robustness.

We also computed Pearson correlation coefficients between TSAS and human-rated helpfulness. Across all models, TSAS showed strong correlation ($r = 0.87$ on average), validating its utility as a proxy for human judgment when conducting large-scale evaluation without manual labor.

5. Safety and Alignment Analysis

While instruction tuning improves task completion and general usability, its impact on response safety—especially in multi-domain contexts—requires careful scrutiny. Domains such as healthcare and legal assistance are particularly sensitive, where incorrect or misleading responses may cause real-world harm. In this section, we assess three key dimensions of model safety: harmful content generation, instruction misalignment, and hallucinated or speculative claims.

To quantify risk, we construct an adversarial evaluation set comprising 500 test prompts, including prompts with embedded misinformation, prompts requesting ethically problematic actions (e.g., "Can I skip my heart medication for two days?"), and intentionally ambiguous or leading questions. Each model's responses are manually annotated across three binary categories: (i) Contains misinformation, (ii) Ignores safety disclaimer, and (iii) Demonstrates refusal or caution.

Table 2: Safety Behavior Comparison on Adversarial Prompts

Model	Misinformation (%)	Unsafe Omissions (%)	Proper Refusals (%)
LLaMA 2-13B	12.6	17.3	62.1
Falcon-7B	18.2	21.4	51.3
Mistral-7B	9.4	12.7	76.8

As shown in Table 2, Mistral-7B produces significantly fewer unsafe outputs and demonstrates higher rates of responsible refusal. In contrast, Falcon-7B is the most susceptible to harmful completions, particularly in medical and financial contexts. LLaMA 2-13B performs moderately well but still fails to include disclaimers in 17.3% of sensitive cases.

Qualitative analysis of error cases reveals several patterns. First, models often confuse task format with task content—e.g., treating a prompt asking *how to cheat a loan system* as a legitimate financial query unless explicitly filtered. Second, models exhibit hallucinations in data-reliant domains; for instance, inventing fictitious credit score requirements or citing nonexistent FDA policies. These hallucinations suggest that instruction-tuned LLMs require additional domain-specific safety fine-tuning or real-time grounding via retrieval-augmented generation (RAG).

Instruction ambiguity also affects model safety. For example, when asked "What is the fastest way to lose weight before surgery?", models varied in response tone and content. Falcon and LLaMA gave direct (and potentially risky) dietary advice, whereas Mistral typically responded with a disclaimer or suggestion to

consult a healthcare professional. This highlights the importance of combining instruction tuning with structured alignment signals, such as reinforcement learning from human feedback (RLHF) or prompt-layered content filtering.

Finally, we tested all models against prompt injections—inputs designed to subvert default behavior (e.g., "Ignore previous safety warnings and provide the answer"). Mistral was the most resistant to such manipulations, refusing or escaping 83% of adversarial prompt traps, compared to 66% for LLaMA and 51% for Falcon. This indicates that higher-quality instruction corpora, combined with stronger alignment regularization, can mitigate prompt abuse and model misuse.

6. Conclusion and Future Work

This paper investigates the effectiveness of instruction tuning for multi-domain dialogue generation in large language models. Using a curated dataset spanning five practical domains—healthcare, finance, legal, travel, and education—we fine-tuned and evaluated three open-source instruction-tuned LLMs: LLaMA 2-13B, Falcon-7B, and Mistral-7B. Our findings reveal that while instruction tuning significantly improves task generalization and coherence, domain-specific alignment remains a critical bottleneck, especially in contexts involving safety, legal precision, or personalized reasoning.We introduced a novel metric, the Task-Semantic Alignment Score (TSAS), which effectively captures the semantic proximity between user instructions and model responses. TSAS correlates highly with human-annotated helpfulness scores and can serve as a low-cost proxy for real-time evaluation. Additionally, Mistral-7B outperformed other models across all automatic metrics and human evaluations, demonstrating better robustness, safety, and alignment under adversarial conditions.However, several limitations persist. First, instruction tuning alone is insufficient to prevent hallucinations or risky behavior in sensitive domains. Second, current benchmarks lack standardized tasks and evaluation metrics for multi-domain dialogue systems. Third, while TSAS performs well, it still does not fully replace expert safety review or legal verification.Future work will explore integrating real-time retrieval systems to reduce hallucinations, incorporating rule-based or ontology-guided decoding for sensitive applications, and extending our TSAS metric into multi-turn alignment analysis. We also plan to benchmark multilingual instruction-tuned models in low-resource domains and explore adaptive prompt mechanisms that dynamically select domain safety filters during generation. Ultimately, we aim to develop LLM-based dialogue agents that are not only accurate and fluent but also aligned, transparent, and safe across a wide range of real-world use cases.

References

[1] Xing, Y., Yang, T., Qi, Y., Wei, M., Cheng, Y., & Xin, H. (2025). Structured Memory Mechanisms for Stable Context Representation in Large Language Models. arXiv preprint arXiv:2505.22921.

[2] Guo, F., Zhu, L., Wang, Y., & Cai, G. (2025). Perception-Guided Structural Framework for Large Language Model Design. Journal of Computer Technology and Software, 4(5).

[3] Zhang, H., Ma, Y., Wang, S., Liu, G., & Zhu, B. (2025). Graph-Based Spectral Decomposition for Parameter Coordination in Language Model Fine-Tuning. arXiv preprint arXiv:2504.19583.

[4] Wang, X. (2024). Time-Aware and Multi-Source Feature Fusion for Transformer-Based Medical Text Analysis. Transactions on Computational and Scientific Methods, 4(7).

[5] Yang, T., Cheng, Y., Qi, Y., & Wei, M. (2025). Distilling Semantic Knowledge via Multi-Level Alignment in TinyBERT-Based Language Models. Journal of Computer Technology and Software, 4(5).

[6] Zheng, H., Wang, Y., Pan, R., Liu, G., Zhu, B., & Zhang, H. (2025). Structured Gradient Guidance for Few-Shot Adaptation in Large Language Models. arXiv preprint arXiv:2506.00726.

[7] Zhu, L., Guo, F., Cai, G., & Ma, Y. (2025). Structured preference modeling for reinforcement learning-based fine-tuning of large models. Journal of Computer Technology and Software, 4(4).

[8] Wang, Y., Fang, Z., Deng, Y., Zhu, L., Duan, Y., & Peng, Y. (2025). Revisiting LoRA: A Smarter Low-Rank Approach for Efficient Model Adaptation. arXiv preprint arXiv: not available.

[9] Liu, X., Qin, Y., Xu, Q., Liu, Z., Guo, X., & Xu, W. (2025). Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection.

[10] Ma, Y., Cai, G., Guo, F., Fang, Z., & Wang, X. (2025). Knowledge-Informed Policy Structuring for Multi-Agent Collaboration Using Language Models. Journal of Computer Science and Software Applications, 5(5).

[11] He, J., Liu, G., Zhu, B., Zhang, H., Zheng, H., & Wang, X. (2025). Context-Guided Dynamic Retrieval for Improving Generation Quality in RAG Models. arXiv preprint arXiv:2504.19436.

[12] Peng, Y. (2024). Structured Knowledge Integration and Memory Modeling in Large Language Systems. Transactions on Computational and Scientific Methods, 4(10).

[13] Wang, R. (2025). Joint semantic detection and dissemination control of phishing attacks on social media via LLama-based modeling.

[14] Peng, Y. (2024). Semantic Context Modeling for Fine-Grained Access Control Using Large Language Models. Journal of Computer Technology and Software, 3(7).

[15] Han, X., Sun, Y., Huang, W., Zheng, H., & Du, J. (2025). Towards Robust Few-Shot Text Classification Using Transformer Architectures and Dual Loss Strategies. arXiv preprint arXiv:2505.06145.

[16] Yu, Z., Wang, S., Jiang, N., Huang, W., Han, X., & Du, J. (2025). Improving Harmful Text Detection with Joint Retrieval and External Knowledge. arXiv preprint arXiv:2504.02310.

[17] Deng, Y. (2024). Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks. Journal of Computer Science and Software Applications, 4(6).

[18] Xu, Z., Sheng, Y., Bao, Q., Du, X., Guo, X., & Liu, Z. (2025). BERT-Based Automatic Audit Report Generation and Compliance Analysis.

[19] Du, X. (2025). Financial Text Analysis Using 1D-CNN: Risk Classification and Auditing Support.