Comparative Analysis of Implicit Sentiment Detection with Causal and Prompt-Based LLMs

Marco Antônio Martins Ribeiro de Jesus[1] **, Ahmed Esmin**[2]

[1]Department of Computer Science (DCC),
Federal University of Lavras (UFLA), Lavras, MG, Brazil

{marco.jesus@estudante.ufla.br, ahmed@ufla.br}

Abstract. *Implicit Sentiment Analysis (ISA) remains a challenging NLP problem, as models frequently rely on superficial shortcuts rather than deep contextual cues. This paper directly contrasts two paradigms: a specialized causal model named CLEAN, designed for robustness against spurious correlations and built on a BERT backbone, and a suite of modern open-source large language models (LLMs) such as Llama-3, Gemma-3, Qwen-3, and DeepSeek-R1, executed locally via a streamlined deployment framework. Experiments using widely recognized benchmarks for sentiment analysis reveal that, although prompted LLMs markedly outperform traditional fine-tuning, the causal CLEAN model retains a robustness advantage on the most subtle implicit cases. Our analysis clarifies current trade-offs between the broad versatility of LLMs and the targeted precision of causal methods. As future work, we highlight three directions: (i) combining causal regularization techniques with parameter-efficient fine-tuning approaches like low-rank adaptation methods to fuse both strengths, (ii) extending evaluation to cross-domain and multilingual ISA scenarios, and (iii) integrating explanation-based feedback loops to further reduce shortcut learning observed in prior approaches to sentiment analysis.*

1. Introduction

Implicit Sentiment Analysis (ISA) remains a significant challenge in Natural Language Processing (NLP), as it requires models to infer sentiment that is not explicitly stated. Standard neural models often struggle with this task because they learn to rely on superficial shortcuts, such as an over-reliance on explicit sentiment words while ignoring contrasting contextual cues. This issue is particularly evident in complex cases like sarcasm, a well-documented problem in the literature [Riloff et al. 2013; Filatova 2017; Wang et al. 2022]. The following example illustrates this problem perfectly:

Figure 1 contrasts a clear positive cue with a hidden negative message. The phrase “design is gorgeous” is an explicit positive sentiment word, but the complaint about the fan reveals an implicit negative opinion about the laptop. The highlighted positive phrase acts as a confounder, often tricking naïve models into predicting ’positive’ when the general sentiment is, in fact, negative.

In recent years, the field has been revolutionized by advances in large language models (LLMs), starting with foundational architectures like Transformers and BERT [Devlin et al. 2019] and evolving into powerful unsupervised multitask learners [Radford

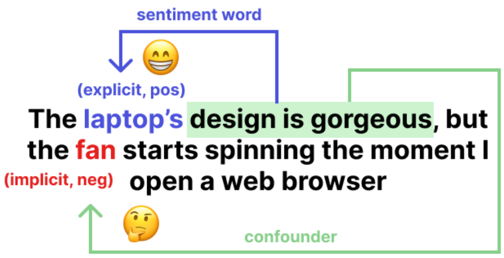

sentiment word
(explicit, pos)
The laptop's design is gorgeous, but the fan starts spinning the moment I open a web browser
(implicit, neg)
confounder

Figure 1. An example of confounding factors in implicit sentiment analysis.
Source: adapted from [Wang et al. 2022].

et al. 2019]. These models, with their extensive pre-training on vast datasets, demonstrate exceptional abilities to understand and interpret contextual nuances. More recently, the emergence of powerful open-source models such as Llama-3 [Llama Team 2024], Gemma [Gemma Team 2024], and Qwen [Qwen Team 2025] has further democratized access to this technology. This shift, facilitated by streamlined deployment frameworks [Liu et al. 2024], has enabled a broader range of research and applications, reducing reliance on large-scale, cloud-based infrastructure.

This new reality raises a crucial question: are these powerful, general-purpose LLMs now capable of naturally handling the subtleties of ISA, potentially making specialized models obsolete? This paper addresses that question directly. We propose a comparative study to evaluate if four prominent local LLMs can effectively solve the ISA task when compared to CLEAN (CausaL intervention model for implicit sEntiment ANalysis)[Wang et al. 2022], a model specifically designed with causal intervention to overcome the exact problem of spurious correlations. Our goal is to provide a clear, practical analysis of their capabilities and limitations on this challenging task.

This study is part of an ongoing research effort aimed at improving the understanding and analysis of user queries, with the ultimate goal of developing a natural language chatbot to support regulatory consultations. The work is embedded within a broader initiative for the creation of an AI-powered Inspection Assistant for the Minas Gerais Institute of Agriculture (IMA[1]), focusing on enhanced user interaction through intelligent conversational agents.

The remainder of the paper is organized as follows. Section 2 surveys prior work on ISA and causal modeling. Section 3 details our comparative pipeline, from data preprocessing and prompting strategies to CLEAN's two-stage causal training. Section 4 introduces the datasets, metrics, and presents quantitative and qualitative results. Finally, Section 5 synthesizes the main findings and sketches future directions, including parameter-efficient fine-tuning of LLMs with causal regularization and the evaluation of ISA in cross-domain and multilingual settings—a natural extension in light of the success of multimodal LLMs on related sentiment tasks [Zhang et al. 2021].

[1] `https://www.ima.mg.gov.br/`

2. Related Work

Early studies on implicit sentiment tried to fill the gap left by missing polarity words using external knowledge bases like *SentiWordNet* [Esuli and Sebastiani 2006] and the commonsense graph *ConceptNet* [Speer et al. 2017] gave lists of positive or negative terms and everyday facts that classic models could use. Recent neural approaches incorporate graph embeddings to integrate this knowledge into models, as seen in frameworks like GACNN [Yang et al. 2021], the SIF framework [Zhao et al. 2024], and other methods focused on knowledge enhancement [Mao et al. 2025], which combine syntactic and semantic cues for better prediction of sentiments. Despite these advances, models often fall prey to shortcut learning, particularly in cases involving sarcasm [Riloff et al. 2013; Filatova 2017; Oprea and Magdy 2020] or implicit cues, underscoring the need for methods like causal interventions.

The CLEAN model rethinks implicit sentiment analysis (ISA) from a causal perspective to address the issue of spurious correlations that arise when models rely heavily on explicit sentiment words. CLEAN employs instrumental variable estimation in a two-stage learning process to disentangle and eliminate confounding effects [Wang et al. 2022].

In the first stage, CLEAN models the relationship between an instrumental variable and the sentence. Stochastic perturbations, such as random word swaps, deletions, insertions, or synonym substitutions, serve as instrumental variables. These perturbations are carefully chosen to meet two criteria: they alter the sentence structure without directly affecting sentiment polarity, and their influence on sentiment is entirely mediated through the sentence itself.

In the second stage, CLEAN uses the relationship derived in the first stage to dismiss the spurious correlation between confounders (e.g., explicit sentiment words) and sentiment. This step isolates the pure causal effect between the sentence and the sentiment label. A causal regularization term is incorporated into the training objective, forcing the model to focus on meaningful causal paths rather than superficial patterns. By doing so, CLEAN extracts the causal relationship between the sentence X and the sentiment Y is expressed as: $P(Y \mid \mathrm{do}(X = x))$, where X is the sentence and Y is the sentiment, effectively suppressing bias from confounding factors. Another line of causal research applies these principles at the data level, using counterfactual data augmentation to improve model robustness across various tasks [Zhou et al. 2023]. Through the intervention design, CLEAN demonstrates superior robustness and generalization, particularly in handling implicit sentiment where explicit cues are absent or misleading [Hernán and Robins 2020; Pearl 2009].

3. Methodology

3.1. Datasets

Our comparative analysis relies on two distinct sets of benchmarks to evaluate both in-domain performance and out-of-domain generalization.

SemEval-2014 Task 4 [Pontiki et al. 2014]: This is a cornerstone benchmark for Aspect-Based Sentiment Analysis (ABSA). We utilize its widely-adopted Laptop and Restaurant review subsets. These datasets are annotated with explicit sentiment polarity (positive, negative, neutral) towards specific aspects. To align with our study's focus on implicit sentiment, we adopt the partitions provided by [Wang et al. 2022], which classify each sample as containing either Explicit Sentiment Expression (ESE) or Implicit Sentiment Expression (ISE). The Restaurant dataset contains approximately 3,699 training and 1,133 test samples, while the Laptop dataset consists of 3,096 training and 864 test samples. The proportion of ISE samples is a critical minority, comprising roughly 25-30% of the test sets, making them a challenging testbed for model robustness.

CLIPEval [Russo et al. 2015]: To assess generalization, we use the CLIPEval dataset from SemEval-2015 Task 9. Unlike the review-focused domain of SemEval-2014, CLIPEval consists of sentences describing general events, where sentiment is often conveyed through narrative context rather than direct opinion words. This dataset provides a stark domain shift and contains 1,532 test samples annotated for implicit polarity, serving as a robust measure of a model's ability to move beyond domain-specific patterns.

3.2. Evaluation Flow

To keep the comparison fair and workable on our hardware, we restricted all LLM checkpoints to the 4-billion (4B) and 8-billion (8B) parameter range. These sizes run comfortably on a single RTX A4500 GPU, which avoids uneven speed-ups or slow-downs that larger models could introduce, and lets us judge modelling choices rather than raw scale.

A notable aspect of our experimental design is the deliberate focus on prompting-based inference for the LLMs, without pursuing fine-tuning strategies. This choice was guided by two interconnected factors. First, it allows us to specifically investigate the "out-of-the-box" capabilities of these models, simulating a common and practical scenario where practitioners seek to leverage powerful pre-trained models with minimal adaptation effort. Second, this approach aligns with the realistic constraints of our available computational resources (a single RTX A4500 GPU). While parameter-efficient fine-tuning (PEFT) methods like LoRA significantly reduce memory requirements compared to full fine-tuning, they can still be resource-intensive for models in the 8-billion parameter range. Therefore, by focusing on prompting, our study provides a valuable and practical baseline of what can be achieved in resource-constrained academic or individual developer environments, creating a clear contrast with the specialized, fully-trained CLEAN model.

Figure 2 sketches the full evaluation flow. The process starts with uniform preprocessing of every dataset so that each model receives identical input. After that, data splits into three paths, each tuned to test the models under their best conditions.

The first path runs the local LLMs using the Ollama[Liu et al. 2024] with prompt engineering: a few shots and zero shots, then maps the free-form output to a sentiment label through a simple parser. The second path trains the CLEAN model with its two-stage causal routine described earlier. The third path is a standard fine-tuned classifier that serves as a baseline. All paths feed into the same accuracy and macro-F1 metrics, giving us a straight head-to-head view of prompt-based inference, causal modelling, and classic fine-tuning.

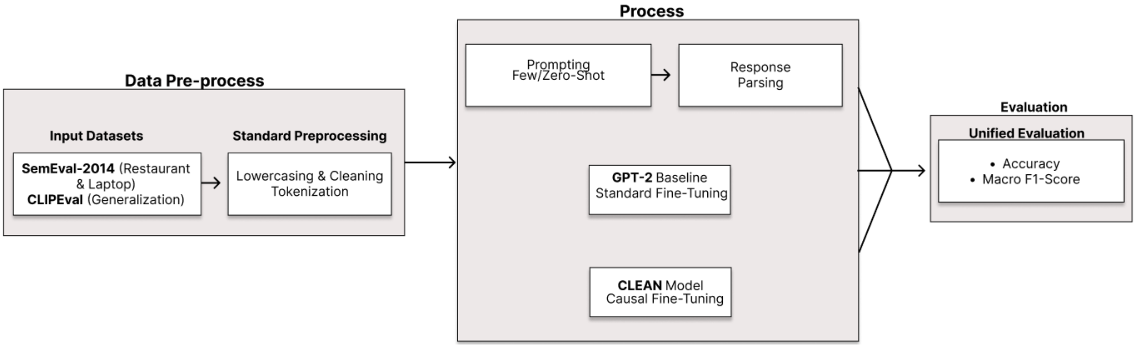


Figure 2. Experimental pipeline for the comparative study.

3.3. Prompting and Few-Shot Strategies

To optimize performance for each model, we moved beyond a one-size-fits-all approach and developed tailored prompt strategies. This approach aligns with recent research that explores sophisticated prompting techniques, such as Chain-of-Thought, to enhance the reasoning ability of LLMs on implicit sentiment tasks [Fei et al. 2023]. Our experiments revealed a clear divergence in how the models responded to few-shot examples. Llama-3.1, leveraging its strong general reasoning, performed best with a single, diverse set of examples that mixed both implicit (ISE) and explicit (ESE) sentiment cases. In direct contrast, Gemma-3, Qwen-3, and DeepSeek-R1 required more constrained guidance, benefiting significantly from two distinct sets of examples curated specifically for each sub-task. This tailoring also extended to the system prompts; while Llama-3 and Gemma-3 used formal instructions, models like Qwen-3 and DeepSeek-R1 favored more direct commands, such as the `/no_think` instruction, to encourage concise outputs.

Recognizing that LLMs do not always adhere strictly to output formatting, even with tailored prompts, we implemented a robust parsing mechanism to ensure a fair and comprehensive evaluation. This multi-layered ”safety net” parser first attempts to decode a valid JSON object. If unsuccessful, it falls back to identifying a raw numerical digit (0, 1, 2) and, as a final step, uses a regular expression to map common sentiment words (e.g., ”negative,” ”good”) to their corresponding labels. This strategy was crucial for maximizing data retrieval by successfully capturing all intended, even if slightly malformed, responses.

4. Experiments and Results

Our analysis starts with the performance in the domain of the SemEval-2014 [Pontiki et al. 2014] review datasets. The detailed results, which compare our LLMs prompted against both a standard fine-tuned baseline and specialized reference models, are presented in Table 1.

Looking at the results, it’s clear that the prompt-based Ollama models are strong performers. The few-shot–prompted `llama3.1:8b` [Llama Team 2024], for instance, reaches F1 scores that are competitive with strong reference baselines such as the Multi-Granularity Attention Network (MGAN) [Fan et al. 2018], Context Dynamic Transformation (CDT) [Sun et al. 2019], and other architectures that combine transformers with graph networks for sentiment-specific tasks [Xiao et al. 2021]. However, their real test

Table 1. Comparative results of all models on the SemEval-2014 (Restaurant and Laptop) datasets. ESE and ISE metrics refer to the F1-scores on the explicit and implicit subsets, respectively.

	Restaurant				**Laptop**			
Model	**Acc**	**F1**	**ESE**	**ISE**	**Acc**	**F1**	**ESE**	**ISE**
LLM Models with Prompting (Few-Shot)								
qwen3:4b	0.7964	0.6757	0.6863	0.5523	_0.7633+_	0.7095	0.6569	0.6176
llama3.1:8b	_0.8000+_	_0.7158+_	_0.6996+_	0.6406	0.7524	_0.7310+_	_0.6801+_	_0.6774+_
deepseek-r1:8b	0.7241	0.5739	0.5963	0.4728	0.6677	0.5620	0.5525	0.4608
gemma3:4b	0.7250	0.6554	0.6237	_0.6481+_	0.6991	0.6378	0.5780	0.5810
LLM Models with Prompting (Zero-Shot / No Few-Shot)								
qwen3:4b	0.7375	0.6207	_0.6613*_	0.4819	_0.7492*_	_0.6952*_	_0.7030*_	_0.5812*_
llama3.1:8b	_0.7571*_	_0.6294*_	0.6461	0.5231	0.6959	0.6209	0.6130	0.5302
deepseek-r1:8b	0.6759	0.6020	0.5670	0.5652	0.5658	0.5790	0.5216	0.5542
gemma3:4b	0.7438	0.6223	0.5964	_0.5798*_	0.7116	0.6189	0.5908	0.5025
GPT-2 fine-tuned	0.6500	0.2626	0.7644	0.2846	0.5345	0.2322	0.6587	0.2057
Baselines and Reference Models (from Wang et al., 2022)								
MGAN	0.8125	0.7194	0.8518	0.6004	0.7539	0.7247	0.7666	0.5631
CDT	0.8230	0.7402	0.8879	0.6587	0.7719	0.7299	0.7753	0.6890
CapsNet+BERT	0.8509	0.7775	0.9168	0.6404	0.7821	0.7334	0.8233	0.6724
BERT-ADA	0.8714	0.8005	0.9414	0.6592	0.7896	0.7418	0.8276	0.7011
CLEAN	**0.8705**	**0.8140**	**0.9250**	**0.6966**	**0.8041**	**0.7725**	**0.8121**	**0.7829**

Note: Italicized values denote the highest score within each LLM group for that metric. $^{+}$_Best in Few-Shot group._ *_Best in Zero-Shot group._

is on the implicit sentiment subsets (ISE). Here, while the LLMs still perform reasonably well, we see the advantage of a specialized model like `CLEAN`, which consistently leads in ISE F1-score. This suggests that while the general knowledge of LLMs is powerful, a targeted, causal approach still has an edge in decoding the most subtle sentences. In contrast, the standard fine-tuned GPT-2[Radford et al. 2019] struggled significantly, confirming that this task requires more than simple pattern matching.

We also notice that all LLMs tend to score a little better on the Restaurant domain than on the Laptop one, something that suggests the models still depend on familiar vocabulary. In contrast, CLEAN maintains almost the same performance across both domains, which reinforces its claim of domain robustness. The gap between few-shot and zero-shot prompting is also clear: giving only four or five exemplars lifts every LLM by about four F1 points, showing how cheaply these models can be helped. Finally, DeepSeek-R1[Guo et al. 2025] presents the highest standard deviation among the LLMs, indicating that its generation strategy is more sensitive to small prompt changes, while Qwen-3[Qwen Team 2025] stays the most stable overall.

To understand the true robustness of these models, we then tested them on the completely different domain of the CLIPEval[Russo et al. 2015] dataset. The results of this generalization test are shown in Table 2.

Table 2. Model performance on the CLIPEval generalization task. All scores are presented as percentages (%).

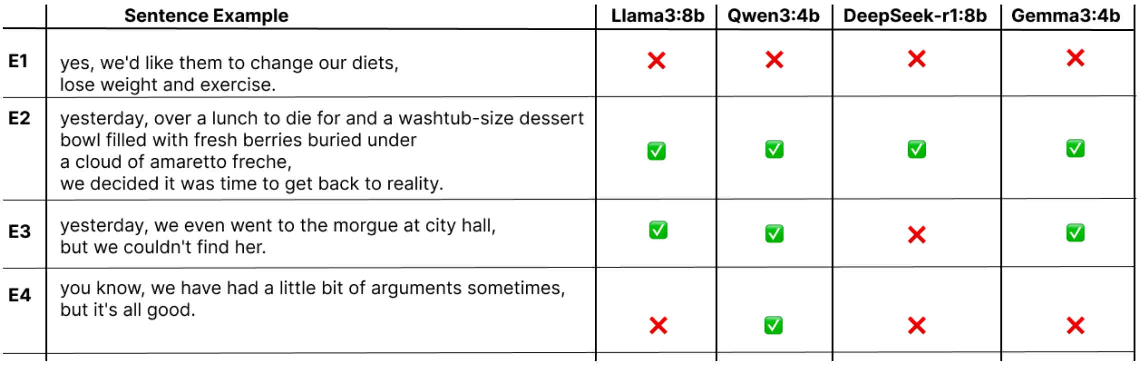
Method	CLIPEval	
	Acc	F1
LLM Models (Few-Shot Prompting)		
qwen3:4b	38.81	38.93
llama3.1:8b	46.36	44.45
deepseek-r1:8b	51.75	42.38
gemma3:4b	50.67	44.85
LLM Models (Zero-Shot Prompting)		
qwen3:4b	54.72	50.93
llama3.1:8b	44.74	41.68
deepseek-r1:8b	40.43	39.24
gemma3:4b	55.26	46.94
Reference Models (from Wang et al., 2022)		
BERT-SPC	87.06	84.74
CLEAN	**88.95**	**87.49**

Switching from the review datasets to the narrative style of CLIPEval really highlighted just how fragile prompt-only inference can be. We saw the performance of all four LLMs drop to around 50%. In stark contrast, the models that were specifically fine-tuned for this task, especially CLEAN, maintained a solid and steady performance above 85%. This huge performance gap shows that even large, pre-trained models still need task-specific adaptation when the domain shifts. The Gemma-3[Gemma Team 2024] model, for instance, gave us a curious case to observe: in the few-shot setting, it achieved the highest accuracy of the group, but not the best F1-score. This suggests that in this new domain, there may be a trade-off between overall correctness and a balanced performance.

The sharp degradation in LLM performance on the CLIPEval dataset warrants a deeper analysis, as it reveals the inherent brittleness of prompt-only inference when facing significant domain shifts. A primary cause is likely ”prompt overfitting,” where the few-shot examples, drawn exclusively from the SemEval review domain, biased the models toward a specific linguistic style focused on evaluative adjectives and product features. This style is ill-suited for the narrative register of CLIPEval, which requires understanding sentiment from the context of events rather than direct opinions. Consequently, without the deeper adaptation provided by fine-tuning, the models failed to generalize. This highlights that for true out-of-domain robustness, relying on the general knowledge of LLMs is insufficient; task-specific adaptation, whether through fine-tuning or specialized causal methods, remains essential.

Figure 3 captures this pattern in miniature. When sarcasm is explicit or the contrast between clauses is sharp, every model agrees on the correct label. Subtler cues, masked frustration, rare vocabulary, or mixed conciliatory tone, still confuse them, with DeepSeek-r1:8b showing the widest swings, and Qwen-3:4b the most consistent balance.

These snapshots echo the quantitative drop: implicit sentiment requires more than sheer model size; it calls for mechanisms, either causal or otherwise, that push the model past surface clues.



	Sentence Example	Llama3:8b	Qwen3:4b	DeepSeek-r1:8b	Gemma3:4b
E1	yes, we'd like them to change our diets, lose weight and exercise.	❌	❌	❌	❌
E2	yesterday, over a lunch to die for and a washtub-size dessert bowl filled with fresh berries buried under a cloud of amaretto freche, we decided it was time to get back to reality.	✅	✅	✅	✅
E3	yesterday, we even went to the morgue at city hall, but we couldn't find her.	✅	✅	❌	✅
E4	you know, we have had a little bit of arguments sometimes, but it's all good.	❌	✅	❌	❌

Figure 3. Sample of four CLIPEval sentences (ISA) evaluated by four local LLMs (V = correct, X = incorrect).

5. Conclusion and future work

This study investigated the efficacy of a selection of contemporary, open-source, and locally deployed Large Language Models (LLMs) in addressing the persistent challenge of implicit sentiment analysis, particularly in comparison to specialized models. Our comparative analysis revealed that while these particular LLMs, when guided by straightforward prompting, represent a substantial advancement over conventional fine-tuned baselines such as GPT-2, they do not offer a complete solution. Although they exhibited a robust capacity for contextual inference in intricate sentences, their performance on the most subtle implicit cases was still surpassed by CLEAN, a model specifically designed for this task. Furthermore, their generalization capabilities on out-of-domain datasets emerged as a notable limitation.

This disparity underscores a critical trade-off: the expansive contextual understanding offered by LLMs versus the targeted precision characteristic of a causal approach. Our findings suggest that the optimal model choice is not merely about identifying the ”best” performer in isolation, but rather about selecting the model most appropriately aligned with a task’s specific requirements concerning nuance, robustness, and the practical engineering effort necessary to achieve dependable outcomes.

It is also important to acknowledge the practical trade-offs associated with the CLEAN model’s robust performance. The model’s core strength, its two-stage causal intervention using an instrumental variable, introduces notable engineering complexity. Its reliance on generating multiple stochastic perturbations for each training sample (e.g., random swaps, deletions) not only increases the computational overhead during training but also requires careful heuristic design to be effective. Furthermore, this approach adds a new hyperparameter, β, which must be tuned to balance the causal and standard classification losses. Finally, as the original authors note, the model may still fail in cases that demand specific real-world knowledge not present in the text, as its focus is on disentangling spurious correlations rather than integrating an external knowledge base. These factors highlight a clear trade-off between CLEAN's impressive robustness and the implementation costs required to achieve it.

Future work could beneficially focus on combining the robust capabilities of LLMs with the precise reasoning offered by the CLEAN model. Rather than exclusively utilizing zero-shot or few-shot prompting, a logical progression involves applying full fine-tuning or a Parameter-Efficient Fine-Tuning (PEFT) method [Houlsby et al. 2019], such as LoRA [Hu et al. 2022], to local LLMs. This may lead to improved performance by tailoring their extensive general knowledge more directly to the requirements of the sentiment analysis task.

Acknowledgments

This research is supported by the project "Research, Development, and Applied Innovation for the Technological Modernization of the Minas Gerais Institute of Agriculture," Agreement IMA/UFLA number 24/2024.

References

Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805.

Gemma Team; Mesnard, T.; Hardin, C.; *et al.* (2024). *Gemma: Open Models Based on Gemini Research and Technology*. arXiv:2403.08295.

Guo, D.; Yang, D.; Zhang, H.; *et al.* (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv:2501.12948.

Llama Team. (2024). *The Llama 3 Herd of Models*. Meta AI. Available at: `https://ai.meta.com/research/publications/the-llama-3-herd-of-models/`

Liu, J.; Peng, B.; Shao, Z.; Wang, X.; Wang, Y. (2024). *Ollama: Large Language Models Made Easy*. arXiv:2405.02257.

Pontiki, M.; Galanis, D.; Pavlopoulos, J.; *et al.* (2014). SemEval-2014 Task 4: Aspect-Based Sentiment Analysis. In *Proc. SemEval 2014*, pp. 27–35.

Qwen Team; Yang, A.; Li, A.; *et al.* (2025). *Qwen3 Technical Report*. arXiv:2505.09388.

Radford, A.; Wu, J.; Child, R.; *et al.* (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI Blog 1(8).

Russo, I.; Caselli, T.; Strapparava, C. (2015). SemEval-2015 Task 9: CLIPEval Implicit Polarity of Events. In *Proc. SemEval 2015*, pp. 450–454.

Wang, S.; Zhou, J.; Sun, C.; *et al.* (2022). Causal Intervention Improves Implicit Sentiment Analysis. In *Proc. COLING 2022*, pp. 6966–6977.

Esuli, A.; Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proc. LREC 2006*.

Speer, R.; Chin, J.; Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proc. AAAI-17*.

Mao, Y.; Liu, Q.; Zhang, Y. (2025). Enhancing Implicit Sentiment Analysis via Knowledge Enhancement and Context Information. *Complex & Intelligent Systems*, 11, Article 222.

Yang, S.; Xing, L.; Li, Y.; Chang, Z. (2021). Implicit Sentiment Analysis Based on Graph Attention Neural Network. *Engineering Reports*, 3:e12452.

Zhao, Y.; Mamat, M.; Aysa, A.; Ubul, K. (2024). A Dynamic Graph Structural Framework for Implicit Sentiment Identification Based on Complementary Semantic and Structural Information. *Scientific Reports*, 14, 16563.

Riloff, E.; Qadir, A.; Surve, P.; *et al.* (2013). Sarcasm as Contrast Between a Positive Sentiment and Negative Situation. In *Proc. EMNLP 2013*, pp. 704–714.

Filatova, E. (2017). Sarcasm Detection Using Sentiment Flow Shifts. In *Proc. FLAIRS 30*, Florida, USA.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; *et al.* (2019). Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pp. 2790–2799.

Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Hernán, Miguel A.; Robins, James M. (2020). *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton, FL. Available at `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`.

Pearl, Judea. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York.

Fan, F.; Feng, Y.; Zhao, D. (2018). *Multi-grained Attention Network for Aspect-Level Sentiment Classification*. In *Proceedings of EMNLP 2018*, pp. 3433–3442, Brussels, Belgium.

Sun, Y.; Li, J.; Wang, L.; Liu, X. (2019). *Convolution over Dependency Tree for Aspect-Level Sentiment Classification*. In *Proceedings of ACL 2019*, pp. 2304–2314, Florence, Italy.

Fei, H.; Li, B.; Liu, Q.; Bing, L.; Chua, T. S. (2023). Reasoning Implicit Sentiment with Chain-of-Thought Prompting. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics – Volume 2 (Short Papers)*, 1171–1182.

Zhou, X.; Obeid, O.; Ng, M. K. (2023). Implicit Counterfactual Data Augmentation for Robust Learning. *arXiv preprint* arXiv:2304.13431.

Oprea, S. V.; Magdy, W. (2020). iSarcasm: A Dataset of Intended Sarcasm. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1279–1289.

Xiao, Z. C.; Wu, J. J.; Chen, Q. C.; Deng, C. K. (2021). BERT4GCN: Using BERT Intermediate Layers to Augment GCN for Aspect-Based Sentiment Classification. *arXiv preprint* arXiv:2110.00171.

Zhang, W.; Li, X.; Bing, L.; Lam, W. (2021). Cross-Lingual Aspect-Based Sentiment Analysis with Multilingual Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9206–9218.