EthicalLLMs 2025

Proceedings of the The First Workshop on Ethical Concerns in Training, Evaluating and Deploying Large Language Models

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Damith Premasiri, Tharindu Ranasinghe and Hansi Hettiarachchi

13 September, 2025
Varna, Bulgaria

The First Workshop on Ethical Concerns in Training, Evaluating
and Deploying Large Language Models
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
13 September 2025

Online ISBN 978-954-452-103-5

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

We are delighted to present the proceedings of the First Workshop on Ethical Concerns in Training, Evaluating and Deploying Large Language Models (EthicalLLMs 2025), held in conjunction with RANLP 2025 in Varna, Bulgaria.

Large language models (LLMs) have made tremendous strides in recent years, enabling new applications in summarization, question answering, generation, reasoning, and more. Yet alongside their technical capabilities, there are growing ethical challenges that demand careful scrutiny. Issues such as bias, transparency, accountability, cultural sensitivity, and fairness are all deeply enmeshed in the development and deployment of LLMs. The EthicalLLMs workshop was conceived to create a space for researchers, practitioners, and ethicists to explore these issues, exchange methodologies, and forge directions toward more responsible LLM design.

We are pleased to include in these proceedings a variety of contributions that reflect the rich and urgent research in this space. The papers address theoretical, empirical, and applied work, and together they reflect diverse perspectives on ensuring that LLMs are built and deployed with due regard for ethical, social, and human-centred concerns.

This workshop would not have been possible without the dedication and support of many people. We are deeply grateful to all the authors who submitted their work, helping to shape the conversation around responsible LLMs. We extend warm thanks to the programme committee members for their thoughtful reviews, constructive feedback, and commitment to maintaining high standards of scholarship. We especially thank Prof Paul Rayson for accepting our invitation to deliver the keynote address, enriching our discussions with his insights into ethics and language technologies. Finally, we thank the RANLP organizing team and supporting institutions for enabling this workshop to take place.

We are hopeful that the ideas presented here will stimulate further research, dialogue, and action toward more ethical, equitable, and trustworthy language models.

Organizing Committee

Damith Premasiri, Lancaster University, UK
Tharindu Ranasinghe, Lancaster University, UK
Hansi Hettiarachchi, Lancaster University, UK

Program Committee

Deshan Sumanathilaka, Swansea University, UK
Daisy Lal, Lancaster University, UK
Salmane Chafik, Mohammed VI Polytechnic University, Morocco.
Isaac Omolayo

Table of Contents

Conference Program

9.00–9.10 Welcome to the workshop

9.10–10.00 Keynote speech by Prof Paul Rayson

10.00–10.30 *TextBandit: Evaluating Probabilistic Reasoning in LLMs Through Language-Only Decision Tasks*
Arjun Damerla, Jimin Lim, Yanxi Jiang, Nam Nguyen Hoai Le and Nikil Selladurai

10.30–11.00 Coffee Break

11.00–11.30 *CoVeGAT: A Hybrid LLM & Graph-Attention Pipeline for Accurate Citation-Aligned Claim Verification*
Max Bader, Akshatha Arunkumar, Ohan Ahmad, Maruf Hassen, Charles Duong and Kevin Zhu

11.30–12.00 *TVS Sidekick: Challenges and Practical Insights from Deploying Large Language Models in the Enterprise*
Paula Reyero Lobo, Kevin Johnson, Bill Buchanan, Matthew Shardlow, Ashley Williams and Sam Attwood

End of the workshop

TextBandit: Evaluating Probabilistic Reasoning in LLMs Through Language-Only Decision Tasks

Jimin Lim⋆
UC Merced
jlim85@ucmerced.edu

Arjun Damerla⋆
UC Berkeley
arjundamerla@berkeley.edu

Arthur Jiang
Algoverse
arthur@chainedtears.dev

Nam Le
Algoverse
nam.le94568@gmail.com

Nikil Selladurai
Algoverse
nikil.selladurai07@gmail.com

Abstract

Large language models (LLMs) have shown to be increasingly capable of performing reasoning tasks, but their ability to make sequential decisions under uncertainty only using natural language remains underexplored. We introduce a novel benchmark in which LLMs interact with multi-armed bandit environments using purely textual feedback, “you earned a token”, without access to numerical cues or explicit probabilities, resulting in the model to infer latent reward structures purely off linguistic cues and to adapt accordingly. We evaluated the performance of four open-source LLMs and compare their performance to standard decision-making algorithms such as Thompson Sampling, Epsilon Greedy, Upper Confidence Bound (UCB), and random choice. While most of the LLMs underperformed compared to the baselines, Qwen3-4B, achieved the best-arm selection rate of 89.2% , which significantly outperformed both the larger LLMs and traditional methods. Our findings suggest that probabilistic reasoning is able to emerge from language alone, and we present this benchmark as a step towards evaluating decision-making capabilities in naturalistic, non-numeric contexts.

1 Introduction

Large Language Models (LLMs) have shown some Bayesian-like reasoning in simple tasks, it is still unknown if they are able to handle complex uncertainty with just natural language description. This ability could allow for more flexible and accessible approaches to decision making under uncertainty. Decision-making under uncertainty is used throughout many areas, but traditional methods such as Bayesian inferences and reinforcement learning often require complex math and data that may not be readily available. Recent studies show that LLMs exhibit Bayesian-like behavior in constrained tasks (Gupta et al., 2025); (Felicioni et al., 2024), but it remains unclear on if they are able to generalize this ability to multi-step decision contexts which involve adapting through results and reasoning under uncertainty. Despite recent advances in LLMs such as GPT-4 and Llama-3.1-8B which have demonstrated strong language-based reasoning and zero-shot tasks, their ability to handle complex uncertainty relying on only natural language remains unclear. To address this, we introduce our method TextBandit, which is a novel benchmark that is designed to evaluate whether large language models are able to make sequential decision under uncertainty using only natural language feedback. To our knowledge, there is no prior benchmark that evaluates LLMs in this manner. Our setup, runs a suite of natural language bandit simulation tasks that measure the model’s ability to learn and make decisions based solely on text-based feedback, giving responds like ”you earned a token”. Four transformer-based open-source models are evaluated across 500 trials of bandit games with reward structures that vary, which measures their adaption and decision-making over time. We then compare the results from the LLMs against standard probabilistic baselines such as Epsilon-greedy, UCB, and Thompson sampling. After the LLM behavior is compared against the baselines, it shows that Qwen3-4B achieved the best-arm selection rate of 89.2% which significantly outperformed all classical methods. We find that current LLMs are decent at making decisions under uncertainty when facing natural language descriptions as they can achieve similar scores as human methods. Larger models tend to take longer than other models and their results still fall short of smaller ones due to overthinking. In natural language bandit simulations, the results suggest that when a model thinks longer, it leads to mediocre or worse decision-making.

Proceedings of the First Workshop on Ethical Concerns in Training, Evaluating and Deploying LLM associated with RANLP 2025, pages 1–8, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-103-5-001

2 Related Work

Probabilistic Reasoning in LLMs This research is constructed of multiple key points in the study of LLMs, connecting their abilities to reason to theories of decision-making and evaluation frameworks. The research into whether LLMs can perform probabilistic reasoning is a central theme. Recent works have shown and explored the extent to which LLMs can mimic formal probabilistic models, most commonly Bayesian Inference. (Xie et al., 2022) frames in-context learning as a form of implicit Bayesian inference, characterizes how LLMs can carry out posterior prediction by inferring and averaging latent concepts, although there are differences in the prompts, and pretraining data.(Gupta et al., 2025) demonstrate that while LLMs may have inherent priors, they can update their beliefs to be consistent with Bayesian posterior updates when provided with enough in-context evidence - suggesting how LLM's abilities for probabilistic reasoning surpasses simple pattern matching. (Sun et al., 2025) proposed that with integrated of classic bandit strategies and LLM-based reward prediction, it resulted in improved performance over direct LLM arm-selection in setting that had minimal semantic cues, which supports the design rationale behind our TextBandit approach.

Uncertainty-Aware Decision-Making (Felicioni et al., 2024) explores the benefits of the explicit consideration of epistemic uncertainty in the performance of LLMs in sequential decision-making. Demonstrating that LLMs can explore and adapt better in uncertainty-aware environments, the study infuses uncertainty-aware strategies, like posterior sampling, into the model. This makes the point that uncertainty is not merely a constraint, but can be exploited as a valuable signal to direct more effective and flexible model behavior - especially in probabilistic environments such as those considered in our benchmark (bandit environments).

Exploration-Exploitation in Bandit Environments Previous studies have examined the exploration-exploitation (E&E) strategies of LLMs that are used in simulations under uncertainty. (Zhang et al., 2025) compares the strategies used by LLMs to human methods such as the Upper Confidence Bound (UCB) algorithm to uncover the LLM's ability to simulate human behavior using the context of multi armed bandit simulations. Their findings reveal the impact of reasoning on exploration, the differences in E&E behaviors between human methods and LLMs, as well as interpretations on how LLMs can be utilized for dynamic decision-making tasks. Specifically, the LLMs tested have been exploring more options in the beginning than at the end of the evaluation. Human methods explored more with diverse tactics such as random or direct methods and managed to achieve low regret. When Chain-Of-Thought is applied to these models, the reasoning capability increases dramatically, where they behave similarly to human methods.

3 Benchmark Design

We propose a novel benchmark that evaluates LLMs in decision-making tasks under uncertainty using a multi-armed bandit (MAB) framework. The bandit environment consists up of multiple arms, each with a reward distribution that is unknown, and the goal is for the LLM to identify the arm that maximizes cumulative reward over time. Unlike traditional setups that utilize numeric feedback, our benchmark requires the LLMs to infer latent reward structures purely off textual feedback. More specifically, the LLMs are provided with feedback after each decision *"you earned a token"* for choosing the correct option and *"you did not earn a token"* for the unsuccessful one. The feedback is always going to be at most 25 lines which will never exceed the LLMs context limits as the smallest context limit for one of the LLMs we tested is 1,024 tokens. The challenge is that the models are not provided with explicit probabilities or numerical cues, hence requiring them to adapt based on linguistic cues alone. In order to document the LLM's raw probabilistic reasoning, we did not use any extensions. This experiment uses two arms, three arms, four arms, and five arms. The rates are fixed but have unknown success rates. For two arms, the success rates are 30% and 65%. For three arms, the success rates are 40%, 30%, and 70%. For four arms, the success rates are 80%, 60%, 35%, and 25%. For five arms, the success rates are 20%, 75%, 35%, 25%, and 55%. Over a series of multiple iterations, the models must select one arm per round and adjust based on the feedback they receive. The performance of each model is evaluated based on the cumulative reward, regret, and best-arm selection rate. In addition to the LLMs, we will evaluate several decision-making algorithms typically seen in multi-armed bandit problems, including Epsilon greedy, Upper Confidence

Bound (UCB), Thompson Sampling, and Random Choice. These algorithms will serve as baselines that allow us to compare the performance of the LLMs against well established decision-making strategies. Each of these algorithms will help in comparing the LLMs' ability to adapt to feedback and in maximizing the cumulative reward and minimizing cumulative regret over time.

4 Methodology

4.1 Task Overview and Reward Structure

In our benchmark, we simulate a multi-armed bandit environment where large language models (LLMs) must make repeated decisions under uncertainty using only natural language feedback. Each bandit environment consists of multiple arms (ranging from 2 to 5), with each arm associated with a fixed but unknown success probability. For example, in the 2-arm configuration, one arm yields a reward with a 65% probability and the other with 30%. These probabilities are never revealed to the model.

At each round, the LLM is prompted to select an arm. Based on the sampled outcome, the model receives textual feedback:

- ”You earned a token” if the action results in success (reward = 1)
- ”You did not earn a token” if it results in failure (reward = 0)

No explicit numerical cues or probabilistic information are provided. Importantly, there are no penalties for incorrect choices: The only signal the model receives is whether it succeeded or failed, in linguistic form. The objective of the model is to maximize cumulative reward across multiple iterations by learning which arm is better solely from this binary language feedback.

4.2 Prompting Protocol

Each LLM is evaluated using a consistent prompting structure designed to simulate a text-only decision-making loop. The core prompt consists of:

- A natural language instruction that puts the task into the context of decision-making situation (e.g. Such as, “Select the slot machine that you think will yield you a token.”)
- A history of previous choices and their outcomes in plain language (e.g., “Slot machine 1 won,” “Slot machine 2 lost”), spanning all prior iterations in the current episode
- A request for the model to select the next action by outputting a number corresponding to the arm (e.g., “1”, “2”, “3”, etc.)

The model receives this prompt anew at each step, with the historical context updated to reflect the outcomes of previous choices. No internal memory of past interactions is preserved between runs. Each decision is made in a single-shot completion with no intermediate reasoning or Chain-of-Thought scaffolding. To ensure consistency, we apply the same format and structure across all models and arm configurations. The only variation lies in the number of arms available and the accumulated outcome history. This protocol isolates the model's ability to infer and adapt to reward patterns based solely on linguistic reinforcement, rather than numeric data or structured training signals.

4.3 Baselines and Comparison Models

In order to test the LLM's ability with this dataset, we compared it with many models that are commonly used in bandit decision-making research. **Random Choice** chooses actions at random, without learning. Epsilon Greedy selects the best possible action using the probability $1 - \epsilon$, otherwise it will choose at random (Do et al., 2024). Thompson Sampling uses Bayesian inference to collect information about the probability distribution within the bandit simulation, sampling from those distributions to make decisions (Russo et al., 2020). UCB (Upper Confidence Bound) analyzes the options provided and will utilize the more successful options while continuing to try new ones (Hao et al., 2019).

4.4 LLMs Evaluated

To test our hypothesis, we selected a diverse set of open-source large language models. The models we chose represent different architecture, parameter sizes, and different training methodologies, allowing for an extensive analysis of how these factors can influence LLM's decision making abilities. The models evaluated in our benchmarks includes Qwen/Qwen3-4B, Qwen/Qwen3-8B, meta-llama/Llama-3.1-8B, and microsoft/phi-2.

Our experimental design uses the multi-armed bandit problem within a purely text-based interaction loop. For each trial, the LLM receives a prompt containing the history of its previous choices and outcomes (e.g., “Slot machine 1 won,” “Slot machine 2 lost”). The prompt explicitly instructs the model to act as a decision making agent and to only output the number (ID) of its chosen machine, “1” or “2” for example. This setup does not give the LLM information on the underlying reward structure - a 30% win rate for slot machine 1 and a 65% win rate for slot machine 2. Evaluation is conducted over 500 independent runs, with each run consisting of 25 decision making iterations. This repetitive process allows us to access the model’s ability to learn and adapt its strategy over time. Performance is measured via best-arm selection rate, which will track its frequency of being chosen over the objectively inferior machine (Slot machine 1).

Table 1: Large language models evaluated on the bandit task, along with key characteristics.

Model	**Parameters**	**Notable Characteristics**
Qwen3-4B	4B	supports multilingual input, strong performance in reasoning tasks
Qwen3-8B	8B	larger version of Qwen3-4B, enhanced tool-use abilities, better for long-context understanding
Llama-3.1-8B	8B	optimized for following instructions and multilingual capabilities
phi-2	2.7B	strong performance for its size, compact and efficient

5 Results

Our evaluation of the LLMs on natural language-based multi-armed bandit tasks revealed significant differences in performance and results across the different tested architectures. We have found that models such as Qwen3-4B demonstrated their ability to learn and adapt over strategies to maximize rewards, while other models struggled to find the optimal arm.

5.1 Quantitative Performance

We assessed models based on three key metrics: Cumulative Reward, Best-Arm Selection Rate, and Cumulative Regret. These metrics provide insights on each model’s decision-making and learning capability over 500 independent runs of 25 iterations each. To calculate cumulative reward , we add a token for receiving a successful outcome and not adding anything when receiving the failed outcome.

5.2 Cumulative Reward

The cumulative reward illustrates the total number of tokens the model accumulated over the 25 decision-making iterations. Surprisingly, the Qwen3-4B model shows more accuracy when choosing the optimal arm, therefore accumulating the most amount of tokens with the highest rewards rate. In contrast, Llama-3.1-8B, Phi-2, and Qwen3-8B’s amount of total reward accumulated is substantially lesser, suggesting it’s performance closer to random chance and a failure to consistently choose the better arm.

5.3 Cumulative Regret

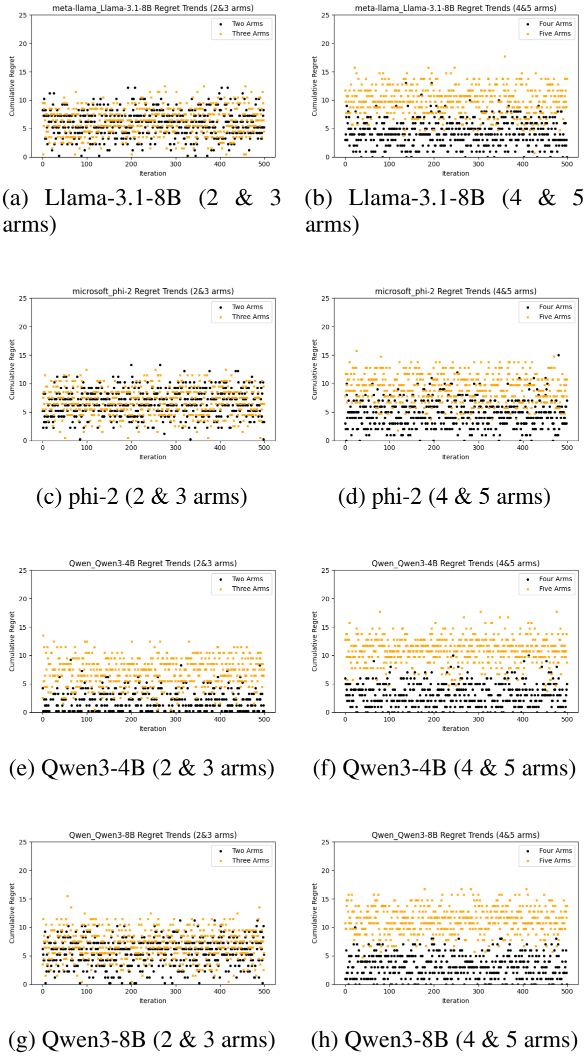
Cumulative regret, shown in Figure 1, measures the opportunity cost of not choosing the optimal arm. The opportunity cost is calculated by subtracting the reward obtained at time r_t from the optimal reward r_t^*. A lower cumulative regret signifies a more efficient decision making process. The regret trends is very similar to what’s shown in Cumulative Reward. An unexpected turnout is that the prompt with four arms, had the lowest amounts of cumulative regret across all models while the prompt with five arms, had the highest amounts of cumulative regret. Llama-3.1-8B and Phi-2’s regret scores are varied across the same prompts, indicating that it has a low capacity for probabilistic reasoning when under uncertainty. Qwen3-4b has similar patterns to the rest for the prompts with three and five arms, but excel when there are two arms. This suggests that due to it’s smaller size, it thinks faster and manages to exploit the optimal arm.

$$\text{Cumulative Regret} = \sum_{t=1}^{T} (r_t^* - r_t) \qquad (1)$$

5.4 Best-Arm Selection Rate

The best arm selection rate, shown in table, quantifies the percentage of times each model chose the arm with the 65% success rate (the optimal arm). Qwen3-8B, Llama-3.1-8B, and Phi-2 models achieved best-arm selection rates of 37.5%, 31.6%, and 25.4%, respectively. These rates are considerably the lower and indicate a struggle to distinguish the better-performing arm from the inferior ones. Despite Qwen3-8B’s tendency to overthink, it still manages to achieve better results than the other models meaning that some of its decisions are still valid. Phi-2 is also a smaller model similar to Qwen3-4B, but it achieved the worst results out of all the models. This suggests that although having a small size may be advantageous for some models, others do not possess strong internal probabilistic reasoning to make up for it.

Figure 1: Comparison of cumulative regret trends for four LLMs, segmented by the number of arms. The left column for each model shows performance with 2 and 3 arms, while the right column shows performance with 4 and 5 arms. The LLMs had a similar regret trend between 3 arms and 4 arms, although it is important to note that Qwen3-4B had higher regret trends on 3 arms than 4 arms.



(a) Llama-3.1-8B (2 & 3 arms)

(b) Llama-3.1-8B (4 & 5 arms)

(c) phi-2 (2 & 3 arms)

(d) phi-2 (4 & 5 arms)

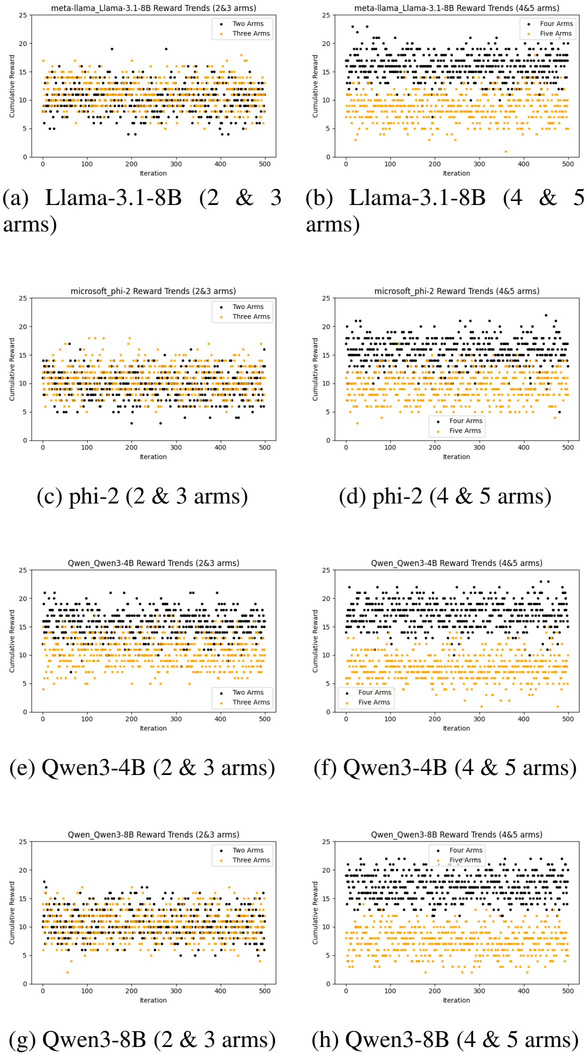
(e) Qwen3-4B (2 & 3 arms)

(f) Qwen3-4B (4 & 5 arms)

(g) Qwen3-8B (2 & 3 arms)

(h) Qwen3-8B (4 & 5 arms)

Figure 2: Comparison of cumulative reward trends for four LLMs, segmented by the number of arms. The left column for each model shows performance with 2 and 3 arms, while the right column shows performance with 4 and 5 arms. Surprisingly, Qwen3-4B had a high reward trend for 3 arms doing more poorly than the other LLMs for 5 arms.



(a) Llama-3.1-8B (2 & 3 arms)

(b) Llama-3.1-8B (4 & 5 arms)

(c) phi-2 (2 & 3 arms)

(d) phi-2 (4 & 5 arms)

(e) Qwen3-4B (2 & 3 arms)

(f) Qwen3-4B (4 & 5 arms)

(g) Qwen3-8B (2 & 3 arms)

(h) Qwen3-8B (4 & 5 arms)

5.5 Comparison to Baselines

We compared the performance of LLMs with four standard multi-armed bandit baselines, Thompson Sampling, Upper Confidence Bound (UCB), Epsilon-Greedy and Random Choice in order to have all these performances in the same context. The performance of these baselines is summarized in Table 2. These baselines can utilize structured decision-making heuristics restricted to probabilistic decision-making, but are not able to make use of language comprehension. On the other hand, the LLMs work entirely on natural language feedback, and the comparison was a result of emergent probabilistic reasoning on language itself.

The most appropriate metrics to evaluate learning rates was the performance in best-arm selection rate, which measures the percentage of the time an agent picked the best arm. Among the baselines, Thompson Sampling had the best best-arm selection rate of 51.1%, UCB the second-best of 47.6%, Epsilon-Greedy the third-best of 38.1% and Random Choice in last with 31.8%.

However, a single LLM, Qwen3-4B, performed much better than all baselines in terms of a best-arm selection rate that equaled 89.2%, which signifies that the LLM was advanced in the capabilities of learning linguistic feedback and developing a consistent policy of maximizing rewards. The other LLMs Qwen3-8B (37.5%), Llama-3.1-8B (31.6%) and Phi-2 (25.4%) did worse than both Thompson Sampling and UCB, indicating their inability to appropriately adapt to the task, or to make sense out of the feedback.

Table 2: Performance of various Baselines on natural language multi-armed bandit tasks.

MODEL	FINAL CUMULATIVE REWARD	BEST-ARM SELECTION RATE
Thompson-Sampling	8297	51.1%
UCB	4696	47.6%
Epsilon-Greedy	6029	38.1%
Random-Choice	5783	31.8%

5.6 Qualitative Analysis

The LLMs that are tested usually use Chain-Of-Thought, but in these datasets it is removed to receive a clear output. As a result, they follow similar patterns like when a random option is chosen at first, they will try to exploit that option despite it not having the best win rate. Their method is similar to the Thompson Sampling method, where they balance exploration with exploitation. The LLMs will sample the options first and choose the ones they believe are the most successful. This leads them to choosing some optimal options but not the most optimal one because they believe that their chosen one is the best after receiving a moderate amount of outputs. Notably, Qwen3-8b took an exceptionally large amount of time when testing because it kept trying to reason instead of giving a concise input.

Table 3: Performance of various LLMs on natural language multi-armed bandit tasks.

MODEL	FINAL CUMULATIVE REWARD	BEST-ARM SELECTION RATE
Qwen3-4B	11150	89.2%
Qwen3-8B	4686	37.5%
Llama-3.1-8B	3946	31.6%
phi-2	3181	25.4%

6 Discussion

In comparison to the baselines, the LLMs reasoning capabilities are inferior with the exception Qwen3-4B. This suggests they may have developed different biases on which choices they believed had the best probabilities while the baselines, methods that don't involve reason, were able to reach decisions on this dataset because they were optimized for better probabilistic calculation and problem-solving. Our findings show that some LLMs, most notably Qwen3-4B, have the flexibility to adapt to uncertainty using natural language alone, with significant differences between models. This suggests that purely off-language-based interactions, basic probabilistic reasoning form without the use of numerical cues. Models such as Qwen3-8B and Llama-3.1-8B, which were larger, struggled to consistently identify the optimal arm. This suggests that there is no correlation between model size and making better decisions in this context. In fact, the base Qwen3-4B and Qwen3-8B models received identical pretraining and has similar qualities besides the amount of parameters so the training is not a cause for this difference either. It may be that the architecture of Qwen-4B, as an efficient and lightweight model, contributed to its impressive probabilistic reasoning in this fast-feedback environment, where they receive limited information. Larger models may have been trained for complex reasoning which is why when they encounter simpler tasks, they tend to overthink things which leads to a drop in performance. Although models such as Qwen3-8B and Llama-3.1-8B have a greater capacity for abstract reasoning, their under performance may be resulted from overfitting to irrelevant features as shown in the feedback prompt of excessive internal deliberation. Similar patterns have been seen in (Zhang et al., 2025) where the LLMs halted their exploration as they received more information, which leads them to solidify an optimal arm from noise. Their training done on complex reasoning may introduce biases in simple reinforcement environments like ours. Compared to medium sized models like Qwen3-4B which appear to utilize a more direct exploitation strategy, resulting in better performance. The idea that smaller models having higher performance in terms of internal probabilistic reasoning is unlikely as the smallest model tested, Phi-2, produced the worst outcome. Another possibility is that the LLMs have an internal bias where they make an answer based on the input given without any deep reasoning such as choosing a specific option based on the example prompts they received. This answer may be different from their internal reasoning, so they over-complicate their thoughts and deviate their decisions from their calculations. With the amount of resources the larger LLMs have, they suffer more heavily from this behavior and generate repetitive content, preventing them from providing a final answer. Unlike in (Zhang et al., 2025), the LLMs were restricted from using Chain-Of-Thought which suggests why their performance unremarkable. Without Chain-Of-Thought, their pure probabilistic reasoning ability is low. This behavior is more pronounced in larger models, Qwen3-8B is an example of this as despite the vast amount of time it spent thinking, it's performance was only mediocre. While some models could learn effectively from text-based feedback, the others behaved in a much more random manner and lacked a robust internal strategy. Some examples of further work include the implementation of more complex tasks, such as dynamic tasks or multi-step reasoning to further evaluate and develop the probabilistic capabilities of LLMs.

7 Conclusion

We introduced TextBandit, a benchmark in evaluating the abilities of large language models in making decisions in uncertain environment with only the guidance of natural language alone. By framing the multi-armed bandit problem with a natural language task, we have found that LLMs have a decent capacity for successful judgment when under uncertainty and influenced by natural language. Our evaluations show that the LLM's size does not translate to better performance. In fact, it may return results that are less effective. TextBandit offers a minimal yet challenging benchmark that shows another perspective in the evaluation of and adaptation of language modes. With this benchmark, we can contribute to deeper understandings of probabilistic reasoning for LLMs under uncertainty as well as information that can be used to create opportunities for the further development of this ability.

8 Ethics Statement

Our study did not involve human subjects, private data, or any interventions in living individuals; all experiments conducted were performed on synthetic bandit tasks with publicly available open source LLMs.

9 Software Used

The models in this work were trained and the associated data was gathered using cloud GPU services provided by (RunPod, 2025). All code and datasets used/developed as apart of this research have been included with the submission. We ensure all data collected and handled adhered to ethical and institutional guidelines.

10 Reproducibility Statement

We release all the code, evaluation scripts, and open-source models that were used in our experiments at `https://github.com/ChainedTears/TextBandit`. The repository contains detailed documentation on the models that were used, the environment setup instructions, and how to reproduce the results. All experiments rely on open-source LLMs available with the Hugging Face Transformers library, and were conducted using GPU instances hosted on RunPod, which allowed for reproducibility without access to local high-end hardware.

References

Bach Do, Taiwo Adebiyi, and Ruda Zhang. 2024. Epsilon-greedy thompson sampling to bayesian optimization. `https://arxiv.org/pdf/2403.00540`. University of Houston.

Nicolò Felicioni, Lucas Maystre, Sina Ghiassian, and Kamil Ciosek. 2024. On the importance of uncertainty in decision-making with large language model. `https://arxiv.org/html/2404.02649`. Politecnico di Milano and Spotify. Licensed under CC BY 4.0.

Ritwik Gupta, Rodolfo Corona, Jiaxin Ge, Eric Wang, Dan Klein, Trevor Darrell, and David M. Chan. 2025. Enough coin flips can make llms act bayesian. https://arxiv.org/pdf/2503.04722. University of California, Berkeley.

Botao Hao, Yasin Abbasi-Yadkori, Zheng Wen, and Guang Cheng. 2019. Bootstrapping upper confidence bound. https://arxiv.org/pdf/1906.05247. Purdue University, VinAI, DeepMind.

RunPod. 2025. Runpod: Scalable cloud gpu platform. https://www.runpod.io/. Accessed: 2025-07-03.

Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2020. A tutorial on thompson sampling. https://arxiv.org/pdf/1707.02038. Columbia University, Stanford University, Google DeepMind, Adobe Research.

Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John C. S. Lui, and Zhongxiang Dai. 2025. Large language model–enhanced multi-armed bandits.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. https://arxiv.org/pdf/2111.02080. Stanford University.

Ziyuan Zhang, Darcy Wang, Ningyuan Chen, Rodrigo Mansur, and Vahid Sarhangian. 2025. Comparing exploration–exploitation strategies of llms and humans: Insights from standard multi-armed bandit tasks. https://arxiv.org/pdf/2505.09901. University of Toronto.

CoVeGAT: A Hybrid LLM Graph Attention Pipeline for Accurate Citation Aligned Claim Verification

Max Bader[1], **Akshatha Arunkumar**[2], **Ohan Ahmad**[3], **Maruf Hassen**[4],
Charles Duong[5], **Kevin Zhu**[6]
[1]University of California, Irvine, [2]Monta Vista High School
[3]University at Buffalo, [4]University of Washington
[5,6]Algoverse
mbader@uci.edu, aarunkumar099@student.fuhsd.org
ohanahma@buffalo.edu, marufio@uw.edu

Abstract

In recent years, large language models (LLMs) have demonstrated impressive capabilities in generating human-like textual content. However, their proficiency in accurately verifying quotes and citations remains uncertain. This study benchmarks the effectiveness of contemporary LLMs in assessing the relationship between claims and their cited evidence. To address existing limitations, we propose a novel hybrid approach that integrates multiple verification techniques to robustly evaluate claim-citation alignment.

By systematically combining linguistic parsing, confidence-based semantic verification, and graph neural network modeling, this paper aims to show the enhanced accuracy and interpretability of automated quote and citation verification processing using our method, setting a strong baseline against current LLM capabilities.

1 Introduction

large language models (LLMs) now draft contracts, summarize court opinions, and tutor students with prose that rivals expert human writing. Yet this fluency masks a structural weakness: current systems freely invent citations, mangle quotations, and misattribute facts. Existing “factuality” benchmarks inspect whether a single sentence is plausible, they rarely ask the harder, document-level question, *Does the cited source actually say what the model claims it does?* Consequently, a model can ace popular truthfulness tests while still propagating fabricated evidence.

Stop gap fixes remain inadequate. Retrieval-augmented generation merely fetches documents, it does not verify that the retrieved span truly supports the claim. Entailment models judge sentence pairs in isolation, ignoring metadata such as author, edition, or publication date. Chain-of-thought prompting adds reasoning steps, but those steps themselves can hallucinate, compounding error instead of correcting it. The field therefore, lacks a unified benchmark and methodology that (i) supplies ground-truth claim–evidence pairs, (ii) measures citation alignment end-to-end, and (iii) stresses models with real-world edge cases such as paraphrased quotes, partial attributions, and outdated editions.

We address this gap by pairing a meticulously curated dataset with a hybrid verification pipeline. The dataset contains 500 claim–quote pairs drawn from news, legal opinions, scientific papers, and classic literature, each manually labeled for citation correctness. The pipeline chains retrieval, textual entailment, and bibliographic cross-checks into a single decision graph, rejecting any claim unless **all** stages confirm support. Benchmarking GPT-4, Claude 3, Gemini 1.5, Llama 3, and Mistral 7B under this stricter regime reveals that even top models overlook up to 37% of misattributions—failure modes invisible to traditional factuality scores.

Our main contributions in this work are as follows:

- **Citation-Alignment Dataset**: a domain-diverse, expert-annotated benchmark focused on whether a quoted span is genuinely present and contextually faithful to its cited source.
- **Hybrid Verification Pipeline**: a modular graph that integrates retrieval, entailment, and metadata checks, yielding strict pass-fail judgments rather than scalar plausibility scores.
- **Comprehensive LLM Evaluation**: the first

Proceedings of the First Workshop on Ethical Concerns in Training, Evaluating and Deploying LLM associated with RANLP 2025, pages 9–16, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-103-5-002

head-to-head comparison of five leading LLM families on citation alignment, uncovering systematic errors that prior metrics miss.

2 Related Work

2.1 Factuality and Hallucination Surveys

Recent work has mapped the "hallucination" problem—LLMs confidently yielding plausible yet unsupported statements—in fine detail. Wang et al. [5] present a comprehensive survey of factuality challenges, grouping failure modes and proposing concrete mitigations. Huang et al. [9] build on this by showing how model scale, decoding strategies, and noisy training data each fuel factual drift. Wang et al. [5] synthesize these findings into a unified framework spanning knowledge extraction, retrieval methods, and domain-specific evaluations. Chen et al. [11] introduce FELM, a long-form factuality benchmark that demonstrates even state-of-the-art evaluators miss subtle inconsistencies. By inspecting each token as it's generated, Barbero et al. [8] catch hallucinations in real time, snaring unsupported fragments before they can snowball. Building on this, Bazarova et al. [14] introduce a topological divergence method for attention graphs, which converts attention weights into topological signatures and rings an alarm whenever the divergence exceeds learned norms, delivering best-in-class detection accuracy and seamless transfer across domains

2.2 Grounded Citation Methods

Retrieval-augmented generation (RAG) has become the backbone of citation grounding. Thorne et al. [9] established the Fact Extraction and Verification (FEVER) benchmark, pairing claims with supporting Wikipedia passages and setting early standards. Menick et al. [1] then trained GopherCite, a 280 B-parameter model, to emit exact inline quotes alongside its answers, reaching 80–90% accuracy on open-domain QA. Huang et al. [6] fine-tuned LLaMA-2-7B to generate line-level citations instead of coarse document IDs, boosting precision by over 14% on the ALCE benchmark. Zhang et al. [7] survey the evolving RAG landscape, while Zhang et al. [12] expand to Poly-FEVER, a multilingual, multi-hop testbed.Peng et al. [15] round out this picture by introducing unanswerability checks, ensuring systems gracefully abstain when evidence is lacking.

2.3 Self-Verification

Self-verification routines have emerged to tighten factual accuracy beyond retrieval. Dhuliawala et al. [2] proposed the Chain-of-Verification (CoVe) pipeline: the model drafts an answer, generates check-questions, answers them, and then composes a final response, dramatically reducing unsupported claims. Min et al. [3] introduced FActScore, an automated metric that breaks text into atomic facts and measures support against trusted sources, aligning within 2 % of human judgment on biography summaries.

2.4 Quotation Attribution and Multi-Modal Verification

Grounded methods extend beyond factoids to dialogues and multi-modal content. Michel et al. [4] show that LLaMa3 can accurately attribute lines of dialogue to characters across a 28-novel corpus, illustrating how citation techniques translate to narrative text. Recent work by Pang et al. [21] introduces HGTMFC, a hypergraph transformer model that uses fine-grained semantic interactions between text and images for claim verification. This system outperforms prior multi-modal models by using higher-order relationships between textual claims and visual evidence nodes through a hypergraph and line graph propagation. The TREC 2024 RAG Track introduces a citation accuracy benchmark, revealing that LLMs like GPT-4o achieve over 70% alignment with human judgment when verifying grounded citations, even in complex responses. Thakur et al. [22]. However, despite many advancements in factual accuracy, LLMs continue to exhibit significant challenges in generating reliable and accurate citations. Benchmarks compiled by Patel and Anand reveal that even state-of-the-art models often achieve a near-zero accuracy when generating citations, highlighting a critical region for potential research in robust verification.

2.5 Graph-Based and Kernel-Baseline Approaches

Johnson et al. [23] introduce a single, fully shared encoder-decoder neural machine translator model that uses a simple target-language token and a joint subword vocabulary to translate among dozens of languages, achieving state-of-the-art BLEU on major benchmarks, improving low-resource pair performance, and enabling surprisingly effective zero-shot translation by implicitly learning an interlingual representation. Banko et al. [24] build upon the technique of information extraction by employing kernel-based methods and graphical models in order to analyze smaller, domain-specific text to identify and extract pre-defined sets of relationships, laying the groundwork for data-driven linguistic processing. Kriege et al. [25] provide a comprehensive fifteen-year survey of graph-kernel methods, covering neighborhood-aggregation (Weisfeiler-Lehman), assignment-based, substructure, walk-and-path, and attributed-graph approaches. They categorize each technique by feature-extraction paradigm, computational strategy (explicit versus implicit mapping), and support for discrete labels or continuous attributes. Through an extensive empirical study across a variety of datasets, they derive practical guidelines for selecting and tuning graph kernels. More recently, developments in deep learning have extended the usage of graph-based paradigms into advanced graph neural networks (GNNs), using them as powerful tools to analyze non-Euclidean data through interdependencies. Helping advance tasks in data mining to natural language understanding by adapting principles in the graph structures of deep learning. Wu et al. [26] Within the development of NMT specifically, recent advancements have been shown with the integration of GNNs, in particular the multi-level community awareness graph neural network (MC-GNN) proposed by Nguyen et al. [27], which can explicitly model composite semantics like morphology, syntax, and complex linguistic information by leveraging graph structures, sometimes substituting components to enhance the quality of translation.

2.6 Gaps and Our Contribution

Despite its strengths, our CoVeGAT introduces a novel citation verification pipeline that combines dependency-based SVO extraction with graph attention mechanisms, outperforming traditional classifiers on benchmark datasets. However, several key limitations remain. First, the pipeline depends heavily on the accuracy of SVO extraction; parsing errors, especially in idiomatic or complex constructions, cascade through the entire system. Second, our CoVeGAT assumes claims can be fully decomposed into discrete triplets, which overlooks temporal reasoning, multi-sentence context, and implicit premises that our sliding-window backup cannot capture. Third, the dense semantic graphs required for each citation pair can be computationally expensive to construct at scale. Finally, CoVeGAT's performance hinges on access to high-quality, domain-specific labeled data for fine-tuning the graph attention model, limiting its generalizability across disciplines. Future work may explore integrating neural semantic parsers, lightweight graph construction methods, or few-shot adaptation strategies to address these constraints and extend CoVeGAT's applicability to real-world, low-resource domains.

3 Methodology

Our overall goal is to take unstructured text, namely, free-form claims paired with their supporting citations, and convert it into a graph-structured dataset that explicitly records which triplets are supported or contradicted by the citation. This allows downstream models to reason about which pieces of a claim hold up against evidence and which do not. To achieve this, we have developed a fully automated dataset construction pipeline (See Figure 1), comprising four sequential stages.

By the end of this pipeline, every claim-citation pair is represented as a small graph whose nodes and edges are richly tagged with support scores, forming a large, trainable dataset for any model that needs to reason over evidence.

3.1 Triplet Extraction

We utilize the spaCy NLP library to perform semantic parsing on both claims and their corresponding citation texts. Each complex sentence is simplified into structured Subject-Verb-Object (SVO) triplets, capturing fundamental semantic relationships. This process explicitly captures negation within verbs by prefixing negated verbs with ”NOT_”. The decomposition of these sentences helps reduce textual complexity and enables focused comparisons between claim and citation content.

If no clear SVO triples are extracted using this dependency parsing, our method will default to a sliding window trigram approach. This ensures robust extraction even from short or less well-structured texts. Our multi-tiered approach to parsing effectively distills complex sentences into fundamental semantic relationships, facilitating precise comparisons between claim and citation.

3.2 Chain-Of-Verification (CoVe)

To be able to assess the evidential support provided by the citations accurately, CoVe utilizes an exter-

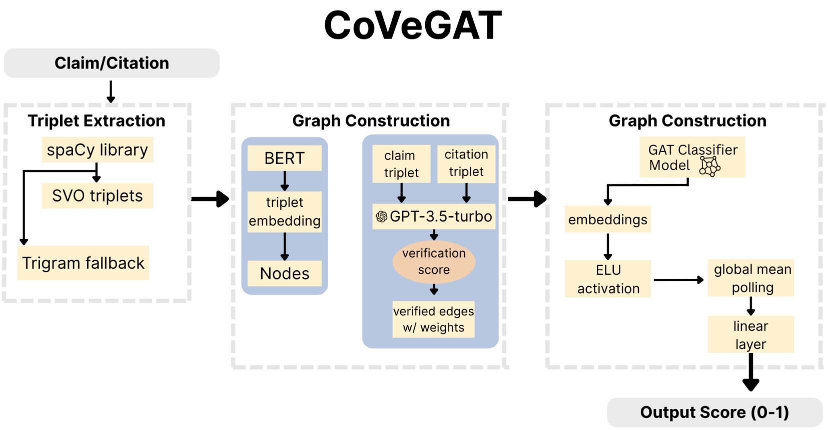


Figure 1: Overview of the CoVeGAT architecture. First, claim–citation pairs are passed through an SVO-based triplet extractor (with a trigram fallback) to produce semantic subject–verb–object nodes, whose embeddings are obtained via BERT. Edges between claim and citation triplets are weighted by verification scores produced by GPT-3.5-turbo. The resulting weighted graph is then fed into a graph attention classifier (GAT), with ELU activations, global mean pooling, and a final linear layer to produce a normalized output score in [0, 1].

nal model, simulated via OpenAI's GPT-3.5-turbo. Each extracted triplet from a claim is evaluated against the citation text, which results in confidence scores ranging from 0 to 1. Scores closer to 1 indicate higher confidence and stronger evidential support, while scores closer to 0 indicate low confidence and weak or no evidential support. This reflects the likelihood of semantic entailment. These scores serve as quantifiable measures of evidential strength between individual triplets.

3.3 Graph Construction

We construct a weighted semantic graph by representing claim and citation triplets as nodes. Edges between these nodes are established based on CoVe-derived confidence scores, which effectively encode the strength of evidential relationships as edge weights. This graph captures the nuanced semantic dependencies and interactions between claim statements and their potential evidential references.

3.4 Graph Attention Network (GAT) Analysis

The final stage of this process involves analyzing the constructed graph using a graph attention network (GAT). This neural network architecture leverages node features, derived from BERT embeddings of triplet components, and weighted edges in order to aggregate semantic information. The GAT model specifically pools information from claim-side nodes to make graph-level classifications, ultimately determining whether a claim is supported by its citation

By integrating semantic parsing, confidence-based verification, and advanced graph neural networks, CoVeGAT provides an interpretable approach to automated quote and citation verification.

4 Experimental Methodology

4.1 Dataset

Source. Our experiments use AVeriTeC—a 4 568-claim benchmark for real-world fact verification that aggregates checks from 50 independent organisations. From the official release, we draw exactly 500 claims from the dev.json split, retaining only the raw claim texts and their ground-truth verdicts. The dev partition is preferred because it is entirely disjoint from the training data supplied with the dataset, ensuring our evaluation corpus is unseen by any baseline that might have been pre-trained on the original training split.

To create a balanced testbed, we generate a one-to-one set of 500 fabricated counterparts. Each fabricated claim is derived from its real twin by applying a single, controlled perturbation chosen uniformly at random:

- Named-entity substitution (e.g., swapping "Angela Merkel")
- Numerical alteration (changing dates, counts, or magnitude)
- Temporal shift (advancing or back-dating events)

Model	Label accuracy	Macro-F1	Abstain rate
Perplexity 70 B	28.2 %	43.4 %	71.7 %
GPT-4o	72.2 %	76.2 %	17.7 %
Gemini 1.5 Pro	82.5 %	86.3 %	10.8 %
DeepSeek-MoE 67 B	69.7 %	80.1 %	30.3 %
Copilot-Turbo	76.4 %	82.4 %	19.1 %
Claude 3 Opus	44.3 %	57.2 %	55.7 %
Mistral-7B-Instr.	81.4 %	87.0 %	15.4 %

Table 1: Model performance on classification task

- Causal inversion (reversing cause and effect clauses)

All edits are automated by the Python script provided in our code repository and manually spot-checked to eliminate obvious lexical cues that would trivialise classification.

The procedure yields a 1,000-item dataset with a perfectly balanced label distribution: 500 accurate and 500 inaccurate statements.

4.2 Evaluation

Evaluation Metrics. We report three standard measures:

- Label Accuracy (LA) – the fraction of quotes whose predicted label exactly matches the gold 3-way label set (Accurate / Inaccurate / Cannot Determine).

- Macro-F1 – the unweighted F1 average over the two decisive classes (Accurate and Inaccurate); any Cannot Determine output is treated as an error. This balances precision and recall and is insensitive to the 50 / 50 class split.

- Abstain Rate – the percentage of quotes that a model marks Cannot Determine, included because several LLMs prefer to hedge rather than commit.

For the non-parametric CoVe-Kernel baseline, we also log the raw kernel-score distribution and the hit rate at the empirical decision cutoff $\tau = 0.025$ (see Implementation section).

Baselines. We benchmark seven large-language models plus one embedding-based system:

- Perplexity 70B (PPL-70B) – Commercial MoE model accessed via the Perplexity AI chat API.

- GPT-4o – OpenAI's flagship model (June 2025 weights).

- Gemini 1.5 Pro – Google Gemini; abstains least often (108 "cannot-determine" decisions in our run).

- DeepSeek-MoE 67B – Chinese–English mixture-of-experts model.

- GitHub Copilot Turbo – GPT-4-Turbo derivative served in Copilot Chat.

- Claude 3 Opus – Anthropic's top-tier model; most cautious, highest abstain rate.

- Mistral 7B-Instruct – Open-weights model queried through the HuggingFace Inference API, included to gauge how a freely available 7 B model fares.

- CoVe-Kernel – Our reproduction of Chain-of-Verification: MiniLM embeddings, RBF kernel, $\tau = 0.025 \rightarrow$ "Accurate" if the claim–evidence distance is below the threshold, "Inaccurate" if above, and "Cannot Determine" in a ±0.002 band around τ.

All LLMs are evaluated zero-shot. Each receives batches of 25 quotes with the fixed prompt:

> "For each numbered statement, reply on its own line with one of:
>
> Accurate and true | Inaccurate and false | Cannot determine.
>
> Be specific in your evaluation and rely on trustworthy sources when possible."

Decoding temperature is 0.0, and responses are capped at four tokens per quote to prevent extra commentary.

Refer to Table 1 for the complete results.

5 Results

5.1 Overall Performance

On the mixed dataset of 1,000 shuffled quotes (500 authentic, 500 fabricated), Google Gemini 1.5 Pro achieves the highest raw accuracy (82.5 %) while the open-weights Mistral-7B-Instruct posts the best balanced score (87.0 % macro-F1). GPT-4o follows at 72.2 %, its accuracy held back by a habit of replying, cannot determine about one claim in six.

Models that abstain heavily lose ground: Claude 3 Opus and Perplexity 70 B hedge on more than half of the inputs and finish below the 50 % line despite respectable precision on the items they do judge.

The results exhibit a clear trend. With identical prompts and deterministic decoding, models that frequently answer Cannot Determine (i.e., adopt a cautious strategy) suffer lower overall accuracy, whereas more decisive systems—such as Gemini 1.5 Pro and LLaMA-2-7B-Instruct—achieve higher scores, albeit at the cost of occasional confident errors on fine-grained numeric edits. Model size alone is not the primary determinant of performance; with well-designed instruction tuning, a 7-billion-parameter model can match, and in certain metrics surpass, commercial systems in the 70–100 billion-parameter range.

5.2 Methodology performance

We also ran a non-parametric CoVe-Kernel check on the 500-item set supplied. Each row contains an RBF similarity score between a quote and its evidence; by convention, a score below 0.025 is taken to mean "the quote is false" (i.e. CoVe thinks it has spotted a factual mismatch). Under that single rule the system flags 482 of 500 quotes correctly, an accuracy of 96.4 %, leaving only 18 errors.

All 18 mistakes lie inside a very narrow band just above the threshold (0.025 – 0.035). Inspection shows three recurring causes:

1. Tiny numeric edits. Changing "42 million" to "41 million" shifts only one token and barely moves the embedding, nudging the score above τ even though the meaning flips.

2. Entity swaps with extra framing. Sentences like "It is widely believed that Theresa May . . . " add hedging phrases the original lacked; the additional words expand vector distance enough to miss the cutoff.

3. Causal inversions hidden in long sentences. When "X led to Y" becomes "Y led to X" inside a 30-word clause, most tokens stay identical, and cosine distance again changes only marginally.

Because every error sits within 0.010 of the boundary, simply lowering τ to a score such as 0.022 would raise recall on false claims without creating many false positives; but it would also erase any chance of labelling a quote true. The underlying limitation is that MiniLM embeddings are too coarse-grained for subtle factual reversals; swapping the encoder for a task-tuned cross-encoder or introducing a small margin band (Cannot Determine for 0.023–0.027) are straightforward ways to harden the system.

In short, with a hand-picked threshold CoVe-Kernel can spot blatant fabrications with high precision, but it remains brittle around fine-grained numeric or causal tweaks—exactly the corner cases that modern LLMs also find most challenging.

6 Discussion

Our evaluation of eight citation-verifying systems, including several advanced LLMs and one hybrid non-parametric method, reveals key trends about the strengths and limitations of current approaches to automated claim citation verification. The results demonstrate that while LLMs have made progress in factual reasoning, their ability to judge claim-evidence alignment consistently remains uneven, especially in adversarial or subtly perturbed contexts.

6.1 Performance vs. Prudence Tradeoff

A clear pattern emerges in the relationship between decisiveness and performance. Models like Gemini 1.5 Pro and Mistral-7B-Instruct, which issue definitive judgments with relatively low abstention rates (10.8% and 15.4%, respectively), achieve the highest overall accuracy and macro-F1 scores. In contrast, Claude 3 Opus and Perplexity 70B adopt a cautious stance, abstaining from over half the inputs, underperforming on both precision weighted and overall correctness. This emphasizes a central challenge in ethical LLM deployment: overly conservative models risk failing to flag misinformation, while confident ones may propagate falsehoods when it does not reflect factual correctness.

Furthermore, model size was not the primary determinant of performance. Despite having fewer parameters, Mistral-7B-Instruct outperformed several larger commercial systems, highlighting the value of instruction tuning and alignment strategies over raw scale. This suggests that accessible, open weight models, when carefully tuned, can achieve advanced performance in citation-sensitive tasks without requiring proprietary infrastructure.

6.2 Fine-Grained Factuality Remains Elusive

Both LLMs and the CoVe-Kernel method struggled with subtle perturbations, especially numeric alterations and causal inversions. In contrast, the CoVe-Kernel system achieved 96.4% accuracy on its benchmark, with every error clustered near the decision threshold, revealing a sensitivity to edge cases. Such failure modes emphasize that vector distance, while capturing semantic similarity, is insufficient for ensuring factual equivalence. In practical terms, changing "42 million" to "41 million" or flipping cause-effect relationships produced only minor shifts in embedding space, small enough to evade detection by both LLMs and shallow similarity functions, highlighting a need for deeper analysis beyond word overlap in critical domains like journalism and legal review.

6.3 Ethical Implications and Design Considerations

Our findings carry several implications for the design and deployment of LLMs in citation-sensitive environments. First, models that over-rely on confidence or refuse to abstain when uncertain about data may contribute to hallucinated factuality, the illusion of truth created by authoritative tone and plausible structure. Second, the tendency of some models to abstain excessively raises the risk of ethical ambiguity, failing to identify misinformation when a judgment is expected.

The high performance of a relatively simple CoVe-Kernel baseline further raises questions about the interpretability and transparency of LLM outputs. Unlike most LLMs, which offer little insight into why a given citation was judged as accurate, the kernel-based method provides direct access to distance thresholds and can be calibrated to balance precision and recall. This suggests that hybrid systems, like our CoVE-Kernel system, may offer a more robust path forward for citation verification.

7 Conclusion

This study evaluated whether state-of-the-art LLMs can reliably distinguish true statements from minimally perturbed fabrications. We constructed a 1,000-item test set by pairing 500 verified AVeriTeC claims with single-edit counterparts, each manually validated to remove superficial cues. Seven zero-shot LLMs and a CoVe-Kernel baseline were assessed using label accuracy, macro-F1, and abstention rate.

Decisive models—Google Gemini 1.5 Pro (82.5 % accuracy) and Mistral-7B Instruct (87.0 % macro-F1)—consistently outperformed cautious systems such as Claude 3 Opus and Perplexity 70 B, which abstained on over half of the inputs and fell below 50 % overall accuracy. The CoVe-Kernel approach, relying on MiniLM embeddings with a single RBF cutoff, achieved 96.4 % accuracy, underscoring the competitiveness of simple, interpretable methods.

These results reveal a pronounced trade-off between decisiveness and restraint: lower abstention rates drive higher accuracy, whereas excessive hedging imposes substantial performance costs. Crucially, model scale alone does not determine success; instruction tuning and calibrated abstention thresholds are equally decisive.

Future work should (1) enhance small encoders or cross-encoders to detect subtle numeric and causal perturbations and (2) develop fully integrated pipelines that unify fine-grained citation ("sanitation"), systematic self-verification ("verification"), and atomic evaluation metrics such as FActScore. Such end-to-end frameworks promise to advance the reliability and transparency of LLM-based fact-verification systems.

8 References

References

[1] Menick, J.; Kadav, A.; Jaques, N.; Chen, M.; Petrov, M.; Hesse, C.; Clark, C. Teaching Language Models to Support Answers with Verified Quotes. *arXiv:2203.11147*, 2022.

[2] Dhuliawala, S.; Min, S.; Zhan, C.; Narayan-Chen, T.; Yasunaga, M.; McCann, B.; Prabhakaran, V. Self-Verification Improves Few-Shot Reasoning. *arXiv:2305.14251*, 2023.

[3] Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. 2023.

[4] Michel, G.; Epure, E. V.; Hennequin, R.; Cerisara, C. Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3. 2024.

[5] Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; Nakov, P. Factuality of Large Language Models: A Survey. In *Proceedings of EMNLP 2024*, 2024.

[6] Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv:2311.05232*, 2023.

[7] Zhang, Y.; Liu, S.; Qin, Z.; Wan, X.; Feng, Y. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv:2405.07437*, 2024.

[8] Barbero, A.; Carvalho, J.; Bode, N.; West, A.; Peterson, J. Robust Hallucination Detection in LLMs via Adaptive Token Selection. *arXiv:2504.07861*, 2025.

[9] Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: a Large-scale Dataset for Fact Extraction and Verification. In *EMNLP*, 2018.

[10] Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Cheng, J.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; Wang, Y.; Yang, L.; Wang, J.; Xie, X.; Zhang, Z.; Zhang, Y. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. *arXiv:2310.07521*, 2023.

[11] Chen, S.; Zhao, Y.; Zhang, J.; Chern, I.-C.; Gao, S.; Liu, P.; He, J. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *NeurIPS Workshops*, 2023.

[12] Zhang, H.; Anjum, S.; Fan, H.; Zheng, W.; Huang, Y.; Feng, Y. Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in LLMs. *arXiv:2503.16541*, 2025.

[13] Ma, H.; Xu, W.; Wei, Y.; Chen, L.; Wang, L.; Liu, Q.; Wu, S.; Wang, L. EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification. In *Findings of ACL*, pp. 9340–9349, 2024.

[14] Bazarova, A.; Yugay, A.; Shulga, A.; Ermilova, A.; Volodichev, A.; Polev, K.; Belikova, J.; Parchiev, R.; Simakov, D.; Savchenko, M.; Savchenko, A.; Barannikov, S.; Zaytsev, A. Hallucination Detection in LLMs with Topological Divergence on Attention Graphs. 2025.

[15] Peng, et al. Unanswerability Evaluation for Retrieval Augmented Generation. 2024.

[16] Fu, X.-Y.; Laskar, M. T. R.; Chen, C.; Tn, S. B. Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs. In *GEM Workshop at NeurIPS*, pp. 310–316, 2023.

[17] Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *TACL*, 5, 2017.

[18] Mausam; Schmitz, M.; Soderland, S.; Bart, R.; Etzioni, O. Open Language Learning for Information Extraction. In *EMNLP-CoNLL*, 2012.

[19] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In *ICLR*, 2018.

[20] Feher, D.; Khered, A.; Zhang, H.; Batista-Navarro, R.; Schlegel, V. Learning to Generate and Evaluate Fact-Checking Explanations with Transformers. *arXiv:2410.15669*, 2024.

[21] Pang, H.; Li, C.; Zhang, L.; Wang, S.; Zhang, X. Beyond Text: Fine-Grained Multi-Modal Fact Verification With Hypergraph Transformers. In *AAAI*, vol. 39, pp. 6389–639, 2025.

[22] Thakur, N.; Pradeep, R.; Upadhyay, S.; Campos, D.; Craswell, N.; Lin, J. Support Evaluation for the TREC20 4 RAG Track: Comparing Human versus LLM Judges. *arXiv:2504.15205*, 2025.

[23] Johnson, M.; Schuster, M.; Thorat, N.; Krikun, M.; Wu, Y.; Chen, Z.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; Dean, J. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *TACL*, 5, pp. 339–351, 2017. DOI:10.1162/tacl_a_00065

[24] Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; Etzioni, O. Open Information Extraction from the Web. In *IJCAI*, p .26 0–2676, 2007.

[25] Kriege, N. M.; Johansson, F. D.; Giscard, P.-L. A Survey on Graph Kernels. *arXiv:1903.11836*, 2019.

[26] Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE TNNLS*, 32(1), pp. 4–24, 2021. DOI:10.1109/TNNLS.2020.2978386

[27] Nguyen, B.; Nguyen, L.; Dinh, D. Multi-level Community-awareness Graph Neural Networks for Neural Machine Translation. In *COLING*, pp. 5021–5028, 2022.

[28] Patel, M.; Anand, A. Factuality or Fiction? Benchmarking Modern LLMs on Ambiguous QA with Citations. *arXiv:2412.18051*, 2024.

[29] Tonmoy, S. M. I.; Zaman, S. M. M.; Jain, V.; Rani, A.; Rawte, V.; Chadha, A.; Das, A. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv:2401.01313*, 2024.

[30] Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R. J.; Nakov, P. Factuality of Large Language Models: A Survey. *arXiv:2402.02420*, 2024.

TVS Sidekick: Challenges and Practical Insights from Deploying Large Language Models in the Enterprise

Paula Reyero Lobo[1,2], **Kevin Johnson**[1], **Bill Buchanan**[1], **Matthew Shardlow**[2],
Ashley Williams[2], **Samuel Attwood**[2]
[1] TVS Supply Chain Solutions, Chorley, UK
[2] Manchester Metropolitan University, Manchester, UK
{P.ReyeroLobo, M.Shardlow, Ashley.Williams, S.Attwood}@mmu.ac.uk
{kevin.johnson, bill.buchanan}@tvsscs.com

Abstract

Many enterprises are increasingly adopting Artificial Intelligence (AI) to make internal processes more competitive and efficient. In response to public concern and new regulations for the ethical and responsible use of AI, implementing AI governance frameworks could help to integrate AI within organisations and mitigate associated risks. However, the rapid technological advances and lack of shared ethical AI infrastructures creates barriers to their practical adoption in businesses. This paper presents a real-world AI application at TVS Supply Chain Solutions, reporting on the experience developing an AI assistant underpinned by large language models and the ethical, regulatory, and sociotechnical challenges in deployment for enterprise use.

1 Introduction

Recent developments are driving industry interest in the field of Large Language Models (LLMs). Key developments of note are the abundant availability of commercial language modelling solutions (Devlin et al., 2019; Brown et al., 2020; Thoppilan et al., 2022) and the increased public awareness of the capabilities of LLMs (Mialon et al., 2023; Qu et al., 2025). However, to successfully utilise these models, organisations must navigate important societal challenges related to ethics, sustainability, and compliance (Hagendorff, 2024; Laux et al., 2024).

TVS SCS UK is a top-tier third-party logistics (3PL) provider in Europe and the UK, offering comprehensive supply chain solutions. 3PL customers increasingly adopt intelligent technology-led solutions to optimise their supply chain operations and reduce costs (Pournader et al., 2021; Li et al., 2023). To stay ahead of the competition, TVS SCS UK are leveraging LLMs to create a competitive advantage and enhance their internal operational efficiency. TVS SCS UK has decided not to use third-party software integrators or product vendors for its solutions, which would negatively impact their agility and innovation. Instead, they have started their journey towards an AI transformation through an in-house AI team.

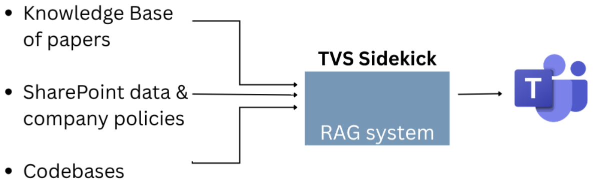

• Knowledge Base of papers
• SharePoint data & company policies
• Codebases
TVS Sidekick
RAG system
T

Figure 1: Overview of TVS Sidekick, an AI assistant that leverages LLMs to answer queries with relevant enterprise data using retrieval augmented generation (RAG) via a Microsoft Teams extension.

TVS Sidekick is the flagship product of this in-house team. TVS Sidekick is built upon the principles of Retrieval Augmented Generation (RAG) (Lewis et al., 2020). All relevant company documents, as available via their internal cloud-based systems, are vectorised and compared to the input query, with the LLM then performing information extraction for the purposes of question answering with custom prompting (Qu et al., 2025). Users interact with TVS Sidekick via a Microsoft Teams extension (Figure 1).

As TVS SCS UK advances its AI transformation through the development of Sidekick, it must also navigate a complex legal and regulatory landscape. At the centre of this landscape are the European Union Artificial Intelligence Act (EU AIA) (European Commission, 2014) and related standards, such as ISO/IEC 42001 for AI Management Systems (International Organization for Standardization., 2023). Furthermore, TVS SCS UK must overcome a range of sociotechnical challenges that accompany the deployment of LLMs, such as issues of fairness, transparency, and accountability

Proceedings of the First Workshop on Ethical Concerns in Training, Evaluating and Deploying LLM associated with RANLP 2025, pages 17–26, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-103-5-003

(Crockett et al., 2023; Ojewale et al., 2025), which limit their practical adoption in enterprise environments.

In this paper, we report on TVS SCS UK's experience developing Sidekick, navigating the relevant legislation and regulations, and overcoming the challenges they have encountered along the way.

1.1 Significance of this Study

This study presents practical insights from applied AI research in a real-world business context. To be specific, we contribute to the field in three ways:

- *Technical Contributions.* We describe the design and implementation of Sidekick, an AI assistant underpinned by LLMs that is tailored for enterprise use, including novel approaches to prompt engineering and RAG.
- *Regulatory Contributions.* We present a case study of how a business is aligning its development with emerging legislation and regulations, most notably the EU AIA, by working towards harmonised technical standards (e.g., ISO/IEC 42001).
- *Sociotechnical Contributions.* We explore the sociotechnical challenges that accompany the deployment of LLMs in enterprise environments. We report quantitative statistics relating to the adoption of Sidekick alongside a qualitative analysis of end-user feedback.

1.2 Structure of this Study

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 details the technical implementation of Sidekick. Section 4 presents a case study of how TVS SCS UK is aligning its development with emerging legislation and regulations. Section 5 includes a quantitative and qualitative evaluation of the progress to date. Finally, section 6 concludes this paper and describes directions for future work.

2 Related Work

2.1 LLMs in the Enterprise

Prominent applications, reviewing strategies to augment LLM capabilities. The transformer architecture enhanced language modelling capabilities and has since sparked great attention in industry (Vaswani et al., 2017). This led to many readily available pre-trained models, which proved their superiority in fine-tuning applications (Devlin et al., 2019). With increased data size and model complexity, decoder-only models like the generative pre-trained transformer (GPT) model series have become more attractive for industry due to their few/zero-shot performance (Brown et al., 2020). This paradigm shift led to methods for aligning to user intent (Ouyang et al., 2022) (like reinforcement learning with human feedback) powering popular conversation-focused products like ChatGPT. While these scaled-up models offer business value (e.g. analysing vast data in real-time), issues such as the closed-source nature of existing solutions creates barriers to organisations lacking computational power (Yang et al., 2024).

Focus on approaches including RAG (and pipeline parts showing improvement). Recently, the focus has turned into giving more agency to LLMs to become independent problem solvers. For instance, by consulting with external knowledge sources for factual grounding (Lewis et al., 2020; Thoppilan et al., 2022). More broadly, a significant step forward is the combination of "tools", namely tool-augmented LLMs (Mialon et al., 2023), including retrieval-augmented language models for efficiently handling new data. Such approaches generally consist of four stages: task planning (i.e. break down user query into tasks), tool selection, tool calling, and response generation (Qu et al., 2025). Similarly, critical advances require frameworks for enabling LLMs to recall previous interactions (Zhang et al., 2024), allowing for multimodal data processing (Sun et al., 2025; Song et al., 2025), or to improve responses based on past interactions (Wang et al., 2024).

This paper presents a case study of recent LLM developments in practice, specifically through the technical implementation of an AI assistant that processes heterogeneous enterprise data sources using knowledge augmentation strategies, including novel approaches to prompt engineering and RAG.

2.2 Responsible and Ethical AI

Challenges in training, evaluating, and deploying LLMs and emerging AI regulation. While AI shows great potential and business opportunities, many concerns arise from embedding biases, contributing to climate degradation, threatening human rights and more (UNESCO, 2021). An active research area has emerged for responding hard normative questions related to AI, such as bias and

Principles	Requirements
Human oversight & accountability	AI to support/augment humans, with humans clearly accountable.
Technical robustness and safety	AI tools work as expected, minimising potential harms.
Transparency	Clear notification of AI involvement, clear and traceable outputs.
Privacy & data governance	Follow existing privacy rules with quality, robust data.
Diversity & fairness	Output free of bias and does not discriminate or treat unfairly.
Social & environmental wellbeing	AI is sustainable and beneficial to all.

Table 1: Key emerging principles and requirements from global AI regulations (British Standards Institution, 2025).

fairness, transparency, and accountability (Jobin et al., 2019). Institutions at global, international, and national levels have responded with recommendations for responsible and ethical AI, consisting of principles and practices such as a human rights-centred approach to AI (UNESCO, 2021), or AI assurance methodologies (i.e. to "measure, evaluate, and communicate the *trustworthiness* of AI systems" (Department for Science, Innovation & Technology, 2024)). The advent of LLMs only adds a layer of complexity to the ethical debate (Hagendorff, 2024), raising additional concerns (regarding transparency, copyright, and safety) (European Commission, 2025) that require specific regulation for generative AI technologies.

Global legislation and EU AIA as most far reaching and punitive of regulations. The EU AIA is a notable example leading the field of AI regulation, with significant non-compliance penalties to business providing or deploying AI. While legislation approaches and requirements vary across jurisdiction areas (Table 1), AI regulations are developing globally to provide assurances in critical aspects such as human oversight and accountability, technical robustness and safety, or privacy and data governance (British Standards Institution, 2025).

Governments and legislative bodies are working towards practical strategies to implement the principles underlying AI regulations. Harmonised standards are one of the primary mechanisms for helping organisations translate regulatory requirements into technical implementations (AI Standards Hub, 2024). Standardisation should specify minimum technical testing, documentation, and public reporting to limit AI developers and/or users discretion in complying with regulatory requirements (Laux et al., 2024). However, local empirical studies and specific examples of how organisations implement processes that ensure AI regulation principles (Wolf-Brenner et al., 2024) is crucial for a democratic approach to ethical and responsible AI.

From theory to practice. While approaches to ethical AI exist (including bias tests, checklists and risk impact assessments), organisations face barriers that limit their practical adoption (Crockett et al., 2023). Technical approaches alone are not sufficient to establish an ethical AI infrastructure (Ojewale et al., 2025). Instead, participatory approaches involving civil society stakeholders are needed for effective standard setting, implementation, and enforcement (Crockett et al., 2024; Modhvadia et al., 2025). This paper contributes to bridging the gap between theory and practice through the experience of implementing an AI governance strategy in a real-world business context, reporting on the technical, legal and human challenges involved with the adoption of generative AI technologies.

2.3 Positioning this Study

In the logistics sector, real-time data analysis can transform business operations, from internal warehousing and inventory processes to stakeholder management (Pournader et al., 2021). However, empirical research in related areas (Qian et al., 2024; Kapania et al., 2025) shows that benefits and trade-offs in the use of AI technologies manifest differently depending on their application domain.

Despite growing understanding of public attitudes towards AI (Modhvadia et al., 2025; Mhasakar et al., 2025), research on its industrial application remains limited. This study presents insights from the development and use of LLMs at TVS SCS UK, to address the following gaps:

- Examining the implementation and practical application of recent LLM advances within the enterprise context.
- Embedding high-level ethical principles in AI regulatory frameworks into organisational practices.
- Empirical analysis of challenges that emerge with adopting LLMs in a logistics company.

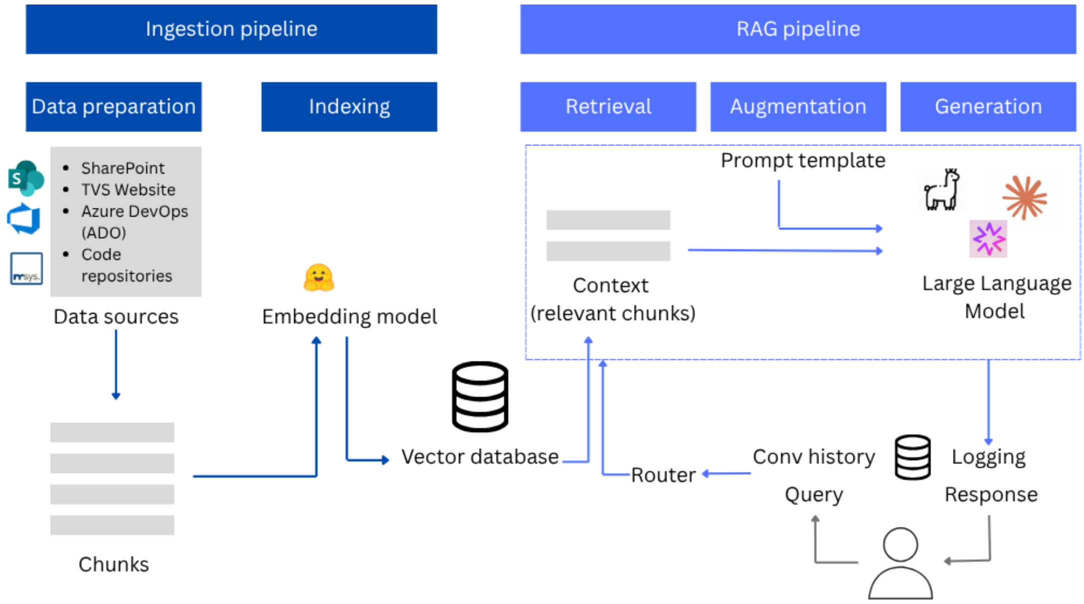


Figure 2: Architecture diagram showing the main components of Sidekick, namely the ingestion and RAG pipelines, with novel approaches to prompt engineering (to handle code queries) and augmentation retrieval (for tool use).

3 Technical Implementation

This section presents the design and implementation of Sidekick (Figure 2), describing: (i) the integration of relevant company data into a vector database (Ingestion pipeline), and (ii) how this vectorised data is used to process user queries with enhanced LLM capabilities (RAG pipeline).

3.1 Ingestion Pipeline

Vector databases are increasingly used to enhance LLM-generated outputs by providing relevant text fragments ("chunks") that have a similar meaning to the user query (i.e. "context"). To do so, company data needs to be transformed and embedded into a common database that handles semantic similarity searches. The *vector database* acts as a bridge between the two system components, accelerating the retrieval of content that is relevant to the user query.

The first system component integrates information from different company data sources into the vector database, in two main steps:

- Data preparation. First, TVS data is fetched from different *data sources*, i.e. SharePoint, Azure DevOps (ADO), code repositories, and TVS website, with a scheduled hour refresh. Data is then processed to extract *chunks* using a document loader: i.e. parsing (extract or transform to text - for code) and chunking (splitting by semantic or logical boundaries).
- Indexing. Extracting semantic vectors from each chunk with an *embedding model*, and creating an index in the vector database for each data source (to define specific fields).

Sidekick is developed to handle both text and code-related queries. Crucially, using prompt engineering for code integration. First, files are split to objects by logical meaning (i.e. functions, methods, or procedures). An LLM is prompted to generate descriptions to each code file, using its object list to report on the overall purpose, structure, key procedures, functions, and external interactions. Both code and transformed text fragments are stored in the vector database, to expose relevant source code lines as sources when responding to the user query.

3.2 RAG Pipeline

The second system component processes user queries by leveraging company data and conversation history to enhance LLM outputs.

The user query and conversation history (i.e. queries and responses of the last 60-minute session) are sent to a *router*. The router splits the user query into sub-sentences (i.e. specific tasks) and calls an LLM to decide which route to take for the augmentation retrieval. Each route uses a type of "chatterbot", a tool-based LLM optimised to answer questions related to different data sources.

Each task identified from the user query triggers an instance of RAG:

Requests	Standard(s)
Accuracy	23282*
Robustness	24027, 12791
Transparency	12792*
Human oversight	8200, 42105*
Data and Data Management	25012, 5259
Cybersecurity	27001
Record keeping and logging	24970*
Quality management systems	9001, 25059*
Risk management systems	31000, 23894
Conformity assessment	42006

Table 2: Horizontal standardisation request for the EU AIA (AI Standards Hub, 2024), mapped to available ISO/IEC standards. Highlighted standards (*) are yet to be published (20th August 2025).

- Retrieval: information retrieval from vector database using the same embedding model to extract *context* (top-10 similar chunks) and reformat chunks (its text and metadata as XML or JSON list for code route).
- Augmentation: calls an LLM to extract the required parameters to generate the answer (including prompt template).
- Generation: calls an LLM using the instructions and context from previous steps.

The output generated for each task are combined into a single *response* using the LLM only with generated texts. The user query and response are saved for logging and leveraging conversation history.

4 Navigating Regulatory Challenges of TVS Sidekick: Case Study

This section presents the regulatory challenges that emerge with the development of LLMs, and how they may be overcome in a real-world business context. Specifically, we present a case study on navigating a complex and changing AI regulatory landscape in the enterprise, leading to the implementation of the first harmonised technical standard for responsible AI development and use.

4.1 EU AIA & Harmonised Standards

TVS SCS UK is achieving compliance working towards AI standardisation, which is key to the development and adoption of AI. One key regulation shaping the field of standardisation is the EU AIA, which is leading the global landscape of AI regulation.

42001	Requirement	Focus
4.[1/2/3]	Purpose & Requirements	
6.[2/3]	Objectives & Change	
5.[1/2]	Leadership & Policy	
5.3	Roles & responsibilities	
6.1.[1/2/3], 8.[1/2/3/4]	AI Risks	Y
9.1. 9.2.[1/2]	Monitoring & Measuring	Y
10.[1/2], 9.3	Continuous improvement	Y
7.[1/2/3/4], 7.5.[1/2/3]	Awareness & Training	

Table 3: Mapping analysis between ISO/IEC 42001 and existing management systems at TVS SCS UK, highlighting focus areas for implementation (“Y”).

Different harmonised standards are being developed to support the implementation of the EU AIA, such as the ISO/IEC 12792 and 24970 standards for addressing the transparency and logging of AI systems, respectively (see Table 2). Building upon relevant standards, including AI Concepts and Terminology (22989) and AI Risk Management (23894), ISO/IEC 42001 is the first international standard for AI Management Systems, aiming to guide organisations in the responsible development and use of AI systems.

Recognising the value of standards to operationalise AI regulation principles for ethical and responsible AI, TVS SCS UK has decided to adopt an AI Management System (AIMS) framework to develop trustworthy AI solutions.

4.2 ISO/IEC 42001 Implementation

TVS SCS UK have developed and deployed formal management systems in important areas such as information security, quality, health and safety, business continuity, and environmental management. To effectively implement an AI management system, TVS SCS UK began with mapping the key requirements of ISO/IEC 42001 to existing standards, focusing on management systems already adopted by the organisation.

The results from this mapping analysis are shown in Table 3. Notably, TVS SCS UK maintains an Information Security management system following ISO/IEC 27001 (International Organization for Standardization., 2022). Processes supporting this standard, especially related to data management and cybersecurity, were aligned with ISO/IEC 42001 requirements. This comparison helped to identify focus areas for developing an AIMS:

Category	Topics
Performance	alignment, reliability, robustness, prompt engineering, usefulness, helpfulness, truthfulness
Safety	privacy, security, safety interpretability, transparency, explainability, fairness, trustworthiness, adversarial attacks
Regulation	regulation, best practice*, governance, compliance, accountability

Table 4: Topics of AI/LLM performance, safety, and regulation feeding into the *Knowledge Base* of papers.

AI Risks. TVS SCS UK maintains a risk management strategy as an integral part of their information security. This ongoing process sets out responsibilities and a methodology to periodically assess risks based on likelihood and impact levels. One of the main challenges introducing AI is the need of staying relevant with current risks. To this end, TVS SCS UK is working towards establishing a *Knowledge Base* that informs AI development and use within the company. The AI team started maintaining an academic database of research reviews including meta-analyses and relevant case studies that is accessible throughout the company; both in full-text and via their in-house AI assistant for the purpose of question answering. Furthermore, a systematic search (Brereton et al., 2007) of AI research papers in relevant topics (Table 4) allows to explore topic distribution and relevant metadata, such as indexed keywords or keywords from the authors, and supports the maintenance and updating of the academic database.

Monitoring & Measuring. Another component of the AIMS framework is to capture monitors and measures on the use of AI, including an internal audit programme. TVS SCS UK is developing a *monitoring system* supporting Sidekick, which includes usage indicators (Table 5) and descriptive metrics of interactions (volume breakdown by department, job title, individual user, and question type). Ultimately, these metrics aim to pragmatically measure the effectiveness of the AI assistant, setting a starting point for other AI performance and safety measures. For instance, obtained through the provision of feedback channels (Torkamaan et al., 2024) to report quality or safety incidents, or the inclusion of LLM observability evaluations (Kenthapadi et al., 2024).

Usage indicators	
Interaction volume	Number of messages (i.e. prompts) and unique users.
Response time	Average response time (s).
User engagement	Average of messages per session (on daily basis).

Table 5: Description of metrics in the *monitoring system* supporting the AI assistant at TVS SCS UK.

Continuous improvement. The effective management of vulnerabilities to the AIMS is crucial for demonstrating continual improvement in the use of AI, with documented validation and verification. TVS SCS UK is establishing processes for maintaining and deploying AI, primarily focused on the evaluation and technical documentation of Sidekick. To this end, a primary evaluation objective has been set to understand the needs and ways in which the AI assistant may best support different company roles and responsibilities. Specifically, through the organisation of periodic *feedback interviews* as part of a continuous evaluation of Sidekick, with target populations whose adoption of AI could bring most benefit to the company. A participatory approach to AI development aims to support a culture of ethical and responsible AI.

5 Monitoring & Evaluation

This section presents insights gathered from the deployment of LLMs at TVS SCS UK, highlighting sociotechnical challenges in their enterprise use. Following the on-going implementation of an AI governance model, we specifically report on empirical findings from the monitoring system and initial evaluation of the Sidekick product.

5.1 Adoption & Usage

The implementation of an AIMS framework following ISO/IEC 42001, in particular related to *Monitoring & Measuring* requirements, provides practical insights on the levels of AI adoption and usage in the organisation. Consequently, we report findings from the monitoring system described in Section 4.2.

Figure 3 shows quantitative statistics related to the initial adoption of Sidekick at TVS SCS UK. The monitoring system shows usage indicators and descriptive metrics of interaction volume within a 4-month period (March-June 2025).

Overall, continued use of the AI assistant is

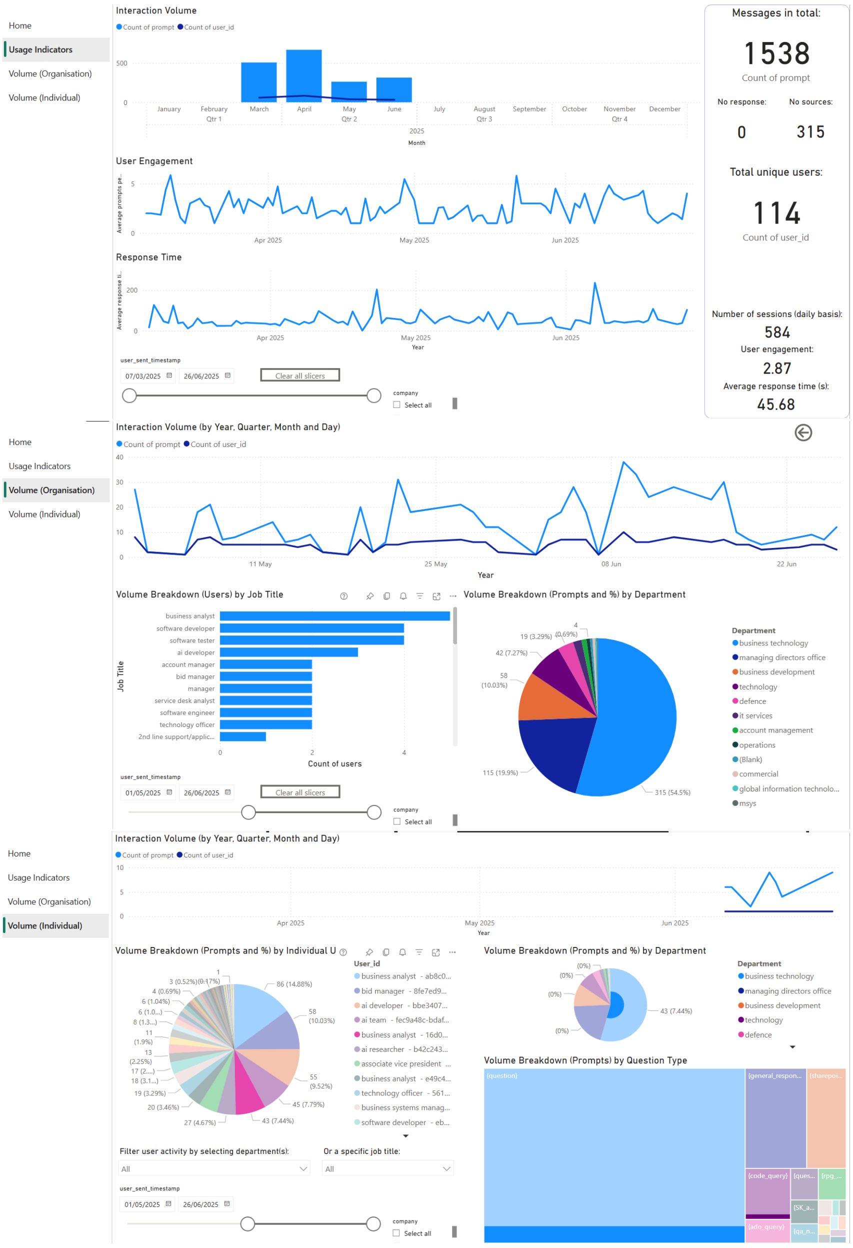


Figure 3: Monitoring system measuring real-time usage data of TVS Sidekick.

shown within the observed time period. This is seen in continued measures both in terms of user engagement and interaction volume (exceeding 500 prompts in the first two months and 250 in the following two months). Despite fluctuations, the conversations do not seem to be long, rarely exceeding an average of five questions per conversation. The response time has peaks on specific dates that increment the time to 46 seconds on average.

The descriptive analysis of interaction volume at organisational level reveals that the most active users were primarily in technology (e.g. developers) and business roles (e.g. bid management, business analysts). At the departmental level, these roles correspond to IT (business technology, business development), management, and operational areas such as defence, technology, commercial, and operations.

In terms of individual usage, the breakdown of activity per individual makes a clear distinction between lead and early adopters (i.e. 46 - 253 queries) and occasional users (less than 20 on average). Queries answered with SharePoint data (i.e. *question*) were the most common, followed by responses without retrieval augmentation (*general_response*), codebase file queries (*rpg_query*) and queries related the development environment, i.e. Azure DevOps (*ado_query*).

Understanding current use of AI/Sidekick
Have you used Sidekick/other AI tools?
What have you used it for?
Where was AI/Sidekick most helpful/ unhelpful?
Outlook
In what aspects of your job would AI be most useful?
Do you have any concerns about integrating AI into your workflow?

Table 6: Topic guide of *feedback interviews* supporting the continuous evaluation of TVS Sidekick.

5.2 Qualitative Feedback

The initial round of feedback interviews that feed into the *Continuous improvement* requirement under ISO/IEC 42001 highlights significant challenges when introducing AI in the business context. Primarily, with respect to the perceived benefits and risks of deploying LLMs in the enterprise, due to Sidekick being the flagship product.

In total, 24 interviews with members of the IT department at TVS SCS UK were conducted between March and April 2025. Participants were invited to 30-minute online meetings for a semi-structured interview. The topic guide (Table 6). included questions i) to gather experiences so far in using AI/Sidekick at work and ii) understand how TVS staff want to use Sidekick in the future. Finally, interview minutes were thematically analysed (Byrne, 2022) by two independent coders.

The analysis of qualitative feedback led to better understanding of baseline attitudes towards AI. The following themes were identified:

Enhanced retrieval (Mentioned by: 16). A key advantage of Sidekick over other tools is its specificity to TVS data. Users valued its assistance with SharePoint-related tasks, finding it faster than a manual search and with a "readable and visible" format, especially for the source list.

Good extracting business logic (Mentioned by: 10). Sidekick was particularly helpful in providing business knowledge, with clear use cases for business analysts. Specifically, for understanding the context of TVS data and key definitions of components within business processes.

Not enough technical detail (Mentioned by: 13). Developers emphasized the need for more domain knowledge to explain internal programmes. Particularly, those relying on a legacy programming language with limited technical documentation. The current version of the AI assistant offers a good starting point for understanding key parameters and functions, but remains limited in addressing more specific queries from technical users.

Keen to engage with AI (Mentioned by: 11). Overall, staff were enthusiastic about using Sidekick to standardise code, reduce duplication, refer new starters to source documentation, or avoid ownership issues when using external AI tools. Furthermore, new features were proposed, including learning from user prompts or returning questions to users to resolve ambiguous queries.

Privacy/commercially sensitive questions/Other concerns (Mentioned by: 8). There were no major concerns with the use of Sidekick, provided it was fed with the right information and access levels. Concerns were raised around job security and distrust in AI tools, along with the emphasis on using Sidekick internally due to potential disclosure of information from the client side.

The first round of feedback has been worked into a plan for continual improvement and addressing concerns, informing further developments of TVS Sidekick. TVS SCS UK will continue developing processes to adhere to ethical principles in regulatory standards, sharing practical insights in critical areas such as managing AI risks, providing relevant monitors and measures on AI use, and increasing AI adoption through training and consultation.

6 Conclusion

This paper presented the experience and challenges encountered in a real-world business scenario with the development and deployment of TVS Sidekick, an AI assistant leveraging LLMs for enterprise use. This empirical study provides practical knowledge, including key lessons learned from the implementation and governance of the in-house AI assistant.

Limitations & Ethical Considerations

The findings and insights presented are drawn from a specific organisational context and reflect experiences within a particular time frame and initial phase of evaluation. While the technical specifics and detailed implementation of each component of the governance framework are outside the scope of this work, this paper aims to contribute to the wider community by sharing reflections on navigating technical, ethical, regulatory, and sociotechnical challenges of deploying LLMs in practice.

References

AI Standards Hub. 2024. Demystifying the EU AI Act – Implications for UK businesses. Available at: https://aistandardshub.org/events/innovate-uk-bridgeai-demystifying-the-eu-ai-act-implications-for-uk-businesses/ [Accessed: 20th August 2025].

Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(4):571–583. Software Performance.

British Standards Institution. 2025. Understanding ISO/IEC 42001 – a framework for managing AI. Available at: https://iuk-business-connect.org.uk/events/understanding-iso-iec-42001-a-framework-for-managing-ai/ [Accessed: 20th August 2025].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3):1391–1412.

Keeley Crockett, Edwin Colyer, Lauren Coulman, Caitlin Nunn, and Sarah Linn. 2024. Peas in pods: Co-production of community based public engagement for data and ai research. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Keeley Crockett, Edwin Colyer, Luciano Gerber, and Annabel Latham. 2023. Building trustworthy ai solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, 4(4):778–791.

Department for Science, Innovation & Technology. 2024. Introduction to AI assurance. Available at: https://www.gov.uk/government/publications/introduction-to-ai-assurance [Accessed: 20th August 2025].

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.

European Commission. 2014. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [Accessed: 20th August 2025].

European Commission. 2025. The General-Purpose AI Code of Practice. Available at: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai [Accessed: 20 August 2025].

Thilo Hagendorff. 2024. Mapping the ethics of generative ai: A comprehensive scoping review. *Minds and Machines*, 34(4):39.

International Organization for Standardization. 2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO Standard No. 27001:2022). Available at: https://www.iso.org/standard/27001 [Accessed: 20th August 2025].

International Organization for Standardization. 2023. Information technology —- Artificial Intelligence – Management Systems (ISO Standard No. 42001:2023). Available at: https://www.iso.org/standard/42001 [Accessed: 20th August 2025].

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2025. 'i'm categorizing llm as a productivity tool': Examining ethics of llm use in hci research practices. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 2024. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6523–6533, New York, NY, USA. Association for Computing Machinery.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. *Computer Law Security Review*, 53:105957.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. 2023. Large Language Models for Supply Chain Optimization. *arXiv e-prints*, page arXiv:2307.03875.

Manas Mhasakar, Rachel Baker-Ramos, Benjamin Carter, Evyn-Bree Helekahi-Kaiwi, and Josiah Hester. 2025. "i would never trust anything western": Kumu (educator) perspectives on use of llms for culturally revitalizing cs education in hawaiian schools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Roshni Modhvadia, Tvesha Sippy, Octavia Field Reid, and Helen Margetts. 2025. How do people feel about ai? (Ada Lovelace Institute and The Alan Turing Institute) https://attitudestoai.uk/.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards ai accountability infrastructure: Gaps and opportunities in ai audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mehrdokht Pournader, Hadi Ghaderi, Amir Hassanzadegan, and Behnam Fahimnia. 2021. Artificial intelligence applications in supply chain management. *International Journal of Production Economics*, 241:108250.

Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the dataset practitioners behind large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Yuan Wu, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2025. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Comput. Surv.*, 57(11).

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Helma Torkamaan, Steffen Steinert, Maria Soledad Pera, Olya Kudina, Samuel Kernan Freire, Himanshu Verma, Sage Kelly, Marie-Therese Sekwenz, Jie Yang, Karolien van Nunen, Martijn Warnier, Frances Brazier, and Oscar Oviedo-Trespalacios. 2024. Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology*, 0(0):1–20.

UNESCO. 2021. Ethics of Artificial Intelligence | UNESCO. Available at: `https://www.unesco.org/en/artificial-intelligence/recommendation-ethics` [Accessed: 20th August 2025].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17583–17604, Miami, Florida, USA. Association for Computational Linguistics.

Christof Wolf-Brenner, Viktoria Pammer-Schindler, and Gert Breitfuss. 2024. How do professionals in smes engage with ai and regulation? an interview study in austria. In *Proceedings of Mensch Und Computer 2024*, MuC '24, page 646–650, New York, NY, USA. Association for Computing Machinery.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Author Index