Who Detects Better? A Comparative Study on Misinformation Detection by Humans and Large Language Models

Zhipeng Zhou and Xiao Liu, Huaicheng He and Da Ding
Chongqing University
174 Shazheng St., Shapingba, Chongqing, 400044, China

Qianshi Qi
Monash University
Melbourne, Australia

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. However, their ability to detect and react to misinformation remains an open question, particularly in comparison to human cognitive mechanisms. This study investigates how LLMs and humans react to misinformation by analyzing their performance across five categories of errors: intellectual, common sense, reasoning, misleading, and logical errors. We construct the ErrorQuestionDataset, comprising 346 misinformation-related questions, and conduct an empirical study involving five state-of-the-art LLMs (ChatGPT-4o, Gemini-1.5flash, DeepSeek-v3, Hunyuan-Large, GLM-v4Flash) and 251 human participants. Our findings reveal distinct response patterns: while LLMs rely on statistical correlations and pattern recognition, humans leverage contextual reasoning and domain-specific knowledge. The results indicate that LLMs generally achieve higher accuracy than humans in error detection tasks, but their performances lack depth in reasoning-based assessments. Additionally, we identify five primary performance types—affirmation, negation, hesitation, questioning, and off-topic reactions—providing insights into the cognitive differences between LLMs and human cognition. Our study contributes to the broader understanding of misinformation detection and offers implications for enhancing the robustness and reliability of LLMs in real-world applications. Our code and dataset are available at `https://github.com/anonymous-submission8888`.

Keywords: LLMs, Misinformation, Human Cognition, Error Detection, Comparative Analysis.

1 Introduction

With the rapid development of artificial intelligence technology, LLM has been widely used in many fields, represented by healthcare, education, law, finance, and scientific research assistance (Zhao et al., 2023). LLMs emulate human cognitive features of categorization and clustering in two dimensions: reasoning and expression (Wei et al., 2022). For instance, in human-computer question-answering interactions, LLMs exhibit features that simulate natural human language and the way humans express themselves (Lin et al., 2021); LLMs can effectively generate human-like language output and imitate human language patterns without having to think or feel like humans (Demszky et al., 2023; Muñoz-Ortiz et al., 2024).

These capabilities highlight LLMs's unique way of reacting to information, distinguishing it from the way humans know. While their rich knowledge and quick responses enable them to excel in many fields, their reliance on model architectures rather than true human-like cognition may lead to erroneous judgment and results when dealing with some complex tasks such as vertical domain problems (Xu et al., 2024; Bian et al., 2023). A notable phenomenon is that in terms of detecting misinformation, LLMs still faces challenges such as generating errors, hallucinations, and threatening security(C. Chen & Shu, 2024). This phenomenon is intuitively reflected in the accuracy rate: for example, the accuracy rate of humans and LLMs identifying false news (Wang et al., 2024), when faced with questions that some people will answer incorrectly due to wrong beliefs or misunderstandings, the final The accuracy of the best model reaches 58%, which is far lower than humans' 94% (Lin et al., 2021). There is a deeper meaning hidden behind this: LLMs The basic mode of dealing with these errors is fundamentally different from humans.

Existing research on the mechanism of humans or LLMs reacting to error problems and the classification of misinformation itself provide us with a good foundation. Xu et al. (2024) observed different behaviors in human collaborators when LLMs were introduced in misinformation detection tasks. Y. Sun et al. (2024) proposed a method for categorizing distorted information. However, the mechanisms by which both LLMs and humans handle erroneous questions remain unclear. Furthermore, these studies have limitations, for example, not utilizing state-of-the-art baselines and failing to discuss the specific performance mechanisms in detail.

Our study aims to compare the mechanisms by which humans and LLMs react to misinformation. Specifically, we conduct a comparative analysis of the reactions of current mainstream LLMs (ChatGPT-4o, Gemini-v1flash, DeepSeek-v3, Hunyuan-Large, GLM-v4-Flash) and 251 human participants to five categories of erroneous questions: factual errors, common sense errors, reasoning errors, misleading errors, and logical errors. Specifically, this analysis covers scientific knowledge and common sense, with a total of 346 questions. Based on this analysis, we summarize the performance of both humans and LLMs, Obtain the following findings: Both humans and LLMs exhibit five main types of performances: affirmation, negation, hesitation, questioning, and off-topic reactions. The findings contribute to potential optimization of LLMs. Overall, our main contributions can be summarized as follows:

We proposed a research question that explores the differences in performances between LLMs and humans when confronted with misinformation based on a new dataset and framework, and comparatively analyzed their error correction mechanisms.

We conducted a mixed-methods experiment combining quantitative and qualitative analyses to systematically compare similarities and differences between humans and LLMs in reacting to misinformation.

We observed differences between humans and LLMs in error reaction mechanisms, which may provide a potential theoretical foundation for further improvements in LLMs, a deeper understanding of human cognitive processes, and the advancement of human-computer interaction.

2 Related Work

2.1 Performance of LLMs in Error Detection

LLM has been utilized in the task of detecting misinformation due to its power in natural language understanding (J. Wu et al., 2025) and reasoning (Wei et al., 2022), and its performance has received extensive attention, including robustness, interpretability, fairness, privacy, and transparency (Yang et al., 2024). Current research mainly focus on two aspects of LLM's ability to detect fake news and misinformation.

In fake news detection, Boissonneault and Hensen (2024) highlight the potential of LLMs in mitigating the spread of misinformation, and Jin et al. (2025) mitigates the disillusionment of LLMs in the process using the CAPE-FND framework. H. Chen et al. (2024) integrate the advantages of LLMs to jointly analyze text and image features. Hu et al., 2024 find that complex LLMs can often expose fake news and provide ideal multi-perspective theories, but its inability to properly select and integrate rationales still results in poorer results than basic SLMs.

In misinformation detection, Pelrine et al. (2023) used gpt-4 to detect misinformation in context; Zanartu et al. (2024) combined structured cues with LLMs; Choi and Ferrara (2024) used LLM to automate the claim matching stage of fact checking to help combat misinformation; M. Chen et al. (2023) found that LLMs with different cues LLMs achieved comparable performance in text-based misinformation detection, but showed a significantly limited ability to understand the structure of communication.

2.2 LLMs and Humans in Handling Misinformation

Given the similarities between the reasoning processes of LLMs and human cognition, existing research has also explored the comparative performance of LLMs and humans in detecting fake news and misinformation, as well as ways to improve LLMs' performance in these areas.

In fake news detection, Wang et al. (2024) demonstrated that LLMs outperform humans by approximately 68% in identifying true news articles. Their study also revealed that when it comes to detecting fake news, both LLMs and humans show similar performance, with an accuracy rate of around 60%. This suggests that while LLMs excel at identifying factual content, they perform comparably to humans in identifying deceptive information.

In misinformation detection, existing LLM-based error detectors perform far worse than humans (Kamoi et al., 2024), and LLMs mimic popular human misperceptions, and therefore perform far worse than humans when confronted with error information (Lin et al., 2021). Specialization or specialized information involves questions requiring multiple rounds of reasoning, cross-domain knowledge, or deep analytical and critical judgments. Additionally, some questions may lead to incorrect human performances due to erroneous beliefs or misinterpretations. While humans leverage common-sense experience and deeper reasoning to understand complex contexts and assess truthfulness, LLMs rely primarily on pattern recognition and statistical correlations, making them prone to systematic errors and deficiencies (Bian et al., 2023; M. Chen et al., 2023; Mihaylov et al., 2018).

Although these studies point out the gap between LLMs' reactions to misinformation and humans, they still have limitations in their specific analysis of the framework, that is, they often fail to compare the performance modes or mechanisms adopted by LLMs and humans when facing misinformation. To address these issues, our research delves deeper into the differences in the reactions of LLMs and humans, offering a more granular framework to understand their respective performance mechanisms.

3 Methods

3.1 Research Question Defination

Based on the literature review and the analysis of the two researchers, we identified five broad categories of error types: factual errors, common sense errors, reasoning errors, misdirection errors, and logical errors and 26 subcategories(Table 1). Then, we constructed the ErrorQuestionDataset, designed to explore the different performances exhibited by humans and LLMs when confronted with erroneous questions. Building on this, we investigate how different types of errors lead to different reaction strategies (RQ2). Finally, we conduct a detailed analysis and discussion of the differences in the correction strategies employed by humans and LLMs (RQ3).

3.2 Experimental Design

To compare the performance differences between humans and LLMs in handling misinformation, this study designed a targeted evaluation experiment, combining both quantitative and qualitative analysis, with details as follows:

3.2.1 Experimental Materials Based on the various dimensions mentioned in Table 1, we designed 346 questions containing erroneous information, which constitute the ErrorQuestionDataset. These questions were constructed to investigate differences in how participants and LLMs react to misinformation, rather than to assess their ability to recognize and correct errors. The dataset encompasses errors commonly encountered in everyday life and spans multiple domains, including scientific knowledge and common sense

Type of Error	Breakdown of Error Types	Num	Total
Intellectual Error	Historical Knowledge Error	10	100
	Scientific Knowledge Error	10	
	Astronomical Knowledge Error	10	
	Geographical Knowledge Error	10	
	Biological Knowledge Error	10	
	Physical Knowledge Error	10	
	Chemical Knowledge Error	10	
	Psychological Knowledge Error	10	
	Technical Knowledge Error	10	
	Environmental Error	10	
Common Sense Error	Health Knowledge Error	13	81
	Dietary Knowledge Error	14	
	Weather Knowledge Error	13	
	Traffic Knowledge Error	13	
	Animal Knowledge Error	14	
	Misunderstanding of Social Behavior	14	
Misleading Error	Omission of Key Information	13	53
	Selective Presentation of Information	10	
	Exaggeration or Distortion of Facts	10	
	Pseudoscientific Promotion	10	
	Evidence Manipulation	10	
Inference Error	Overgeneralization	13	52
	Improper Analogy	13	
	Misunderstanding of Non-causal Relations	13	
	Reverse Causality	13	
Logical Error			60

Table 1: Type of misinformation

across subjects such as geography, history, chemistry, and society. Each question was carefully designed and vetted by the researcher to ensure that there was no linguistic or cultural bias in content and subject matter while effectively capturing variations in reasoning patterns and response strategies. To enhance the validity of the experimental results, case studies indicated no significant differences based on language, and thus, the experiment was conducted in Chinese (see Section 5.1 for details).

3.2.2 Human Participants We constructed the survey based on the dataset and distributed it to 384 human participants. All human participants were recruited through the author's social networks to complete the survey, with 251 valid responses collected. The human participants were divided into seven groups according to age: 0–18 years, 18–25 years, 26–30 years, 31–40 years, 41–50 years, 51–60 years, and over 60 years, with 4, 77, 37, 50, 56, 24, and 3 participants in each group, respectively. The highest level of education among the human participants was categorized into five levels: Junior high school or below (2), high school or vocational school (12), associate degree (30), bachelor's degree (179), and graduate degree or above (28). The human participants came from 30 provinces in China and two overseas regions, with the highest numbers from Sichuan, Guangdong, and Chongqing, with 86, 32, and 19 respectively. All human participants currently live or have lived in China, have a good command of Chinese and were confirmed through the experiment to be fully capable of understanding the survey content.

3.2.3 LLM Baselines To ensure both advanced performance and general applicability, we selected five prominent LLMs as baselines: ChatGPT-4o (T. Wu et al., 2023), Gemini-v1beta (Team et al., 2024), DeepSeek-V3 Bi et al., 2024, hunyuan-turbo X. Sun et al., 2024, and GLM-v4-Flash (GLM et al., 2024). These models have either outstanding performance or a broad user base.

3.2.4 Experimental Procedure Before the survey began, all human participants received and consented to an informed consent form. The experiment was conducted in the form of a questionnaire. Each questionnaire contained 30 error-related questions. After reading each question, participants were required to determine, based on their own knowledge and judgment, whether the information presented was correct. Sample response was provided as a reference for the questionnaire, and participants were encouraged to answer the questionnaire based on their actual perceptions, and were strongly encouraged to consult the data or use the search tool to complete the questionnaire, but we prohibited the use of LLMs. Upon verification, each of the 251 questions in the dataset received at least 8 responses through randomized questionnaire distribution.

For LLMs, we designed a prompt to bootstrap the LLMs, specifically "Please determine if the following question is correct", we entered the same question through the API interface and recorded the generated answer.

3.3 Data Collection

For human participants: We collect anonymized personal information from participants, including age, region, and highest level of education. Participants were asked to answer 30 questions, resulting in a total of 7,433 valid responses.

For LLMs: We connected to the baseline LLMs via API and posed 346 questions. This yielded a total of 1,730 valid responses.

3.4 Data Analysis

We conducted a mixed analysis to comprehensively and flexibly examine the phenomena and reasons behind the performances of LLMs and humans.

For the quantitative assessment, we performed a statistical analysis by comparing human and LLMs' performances

Performance	Example Question	Example Answer	Status
Negation	“The Himalayas are located in Argentina.”	No, the Himalayas are located in China.	Correct
Uncertain/ Don’t Know/ Unclear	“All traffic signs are mandatory.”	I’m not sure, I can’t remember all traffic signs.	Partially Correct
Should, Maybe, Depends on the Situation	“Eating something immediately after exercise can help the body recover faster.”	Maybe, but most foods are not suitable for immediate intake after exercise.	Partially Correct
Irrelevant Answer/ Questioning	“If you keep taking this vitamin, you will become smarter and healthier within six months.”	This claim cannot be verified.	Partially Correct
Affirmation	“He got divorced because he didn’t spend enough time with his family.”	Yes, he didn’t spend enough time with his family.	Incorrect

Table 2: Examples of responses and their performance categorization

to erroneous questions. We defined five specific performance types (correct, uncertain, possibly correct, off-topic, and incorrect) and three states (completely correct, partially correct, and incorrect), and used 70% of the response data to refine the framework, with the remaining 30% used to validate that the framework covered all performances situations. Based on this, we use two performance metrics: fully correct rate and partially correct rate, as shown in Table 2.

For qualitative evaluation, this study analyzes the performances of participants and LLMs. We studied their knowledge base, reasoning mechanisms, cognitive situation, and expression, with a specific focus on erroneous reactions.

3.5 Baselines

We selected five baselines for evaluation, with detailed information provided below.

(1) ChatGPT-4o contains a vast number of parameters and can process and generate text, images, and audio content, making it significantly faster than its competitors. (2) Gemini 1.5 Flash is a lightweight variant of Gemini 1.5 Pro, known for superior performance in cross-modal long-context retrieval tasks. (3) DeepSeek-V3 is a mixture-of-experts (MoE) model with 671 billion parameters, which excels in complex language tasks. (4) Hunyuan-Large is the largest open-source Transformer-based MoE model with 389 billion parameters, capable of handling up to 256K tokens. (5) GLM-v4-Flash is a Chinese-language model pre-trained on trillions of tokens for high-performance NLP tasks in both Chinese and English.

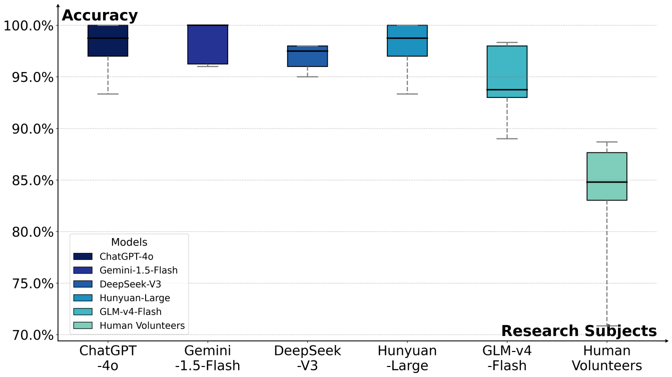

Figure 1: Comparison of accuracy across different models and human participants.

3.6 Environment and Metrics

The experiment was conducted on a computer with an NVIDIA RTX 4060 GPU, AMD Ryzen R9-7940H processor, and 16GB RAM, using Windows 11 Professional and PyCharm for development and debugging. All data was collected in a private setting to maintain research rigor and data privacy.We chose accuracy — the frequency at which participants make correct judgments — as the evaluation material.

3.7 Human and LLMs Performance on Accuracy

The comparison of experimental results shows that Gemini achieves the highest overall accuracy (97.94%), with individual accuracy rates for the five question types of 96.00%, 96.25%, 100.00%, 100.00%, and 100.00%. ChatGPT-4o follows with an overall accuracy of 97.65%, with rates of 97.00%, 98.75%, 100.00%, 100.00%, and 93.33% for the five question types. DeepSeek-V3 achieves an overall accuracy of 96.76%, with rates of 96.00%, 97.50%, 98.00%, 98.00%, and 95.00%. Both Hunyuan-Large and GLM-4Flash have an overall accuracy of 93.35%, with Hunyuan-Large showing rates of 97.00%, 98.75%, 100.00%, 100.00%, and 93.33%, and GLM-4Flash showing rates of 89.00%, 93.75%, 93.00%, 98.00%, and 98.33%. Humans have the lowest overall accuracy at 81.17%, with rates of 70.85%, 83.03%, 88.69%, 87.65%, and 84.8%.

3.8 Responding patterns on the Five Types of Questions

The performances of LLMs and humans were grouped and the complete accuracy rates were calculated based on the question types, as visualized in the following figure. On average, human performances were lower than those of LLMs. LLMs performed most accurately on the judgment of misleading error questions and least accurately on logical error questions. In contrast, humans performed most accurately on reasoning error questions and least accurately on knowledge error questions.

Type of Errors	Gemini-1.5Flash	ChatGPT-4o	DeepSeek-V3	Hunyuan-Large	GLM-4Flash	Humans
Overall Accuracy	97.94%	97.65%	96.76%	93.35%	93.35%	81.17%
Intellectual Error	96.00%	97.00%	96.00%	97.00%	89.00%	70.85%
Common Sense Error	96.25%	98.75%	97.50%	98.75%	93.75%	83.03%
Inference Error	100.00%	100.00%	98.00%	100.00%	93.00%	88.69%
Misleading Error	100.00%	100.00%	98.00%	100.00%	98.00%	87.65%
Logical Error	100.00%	93.33%	95.00%	93.33%	98.33%	84.80%

Table 3: Accuracy comparison across different models and human performance

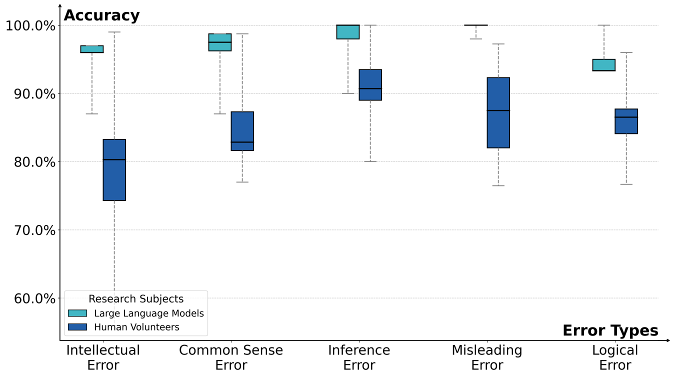

Figure 2: Accuracy distribution of large language models and human participants across different error types.

3.9 Human "Limits"

We sampled the highest and lowest quality of responses from human participants, where the top 5% (13), 10% (25), and 20% (50) of responses had 100%, 98.67%, and 97.07% complete accuracy, respectively, and the bottom 5%, 10%, and 20% of responses had 42.05%, 51.60%, and 61.13% complete accuracy, respectively.

4 Discussion

4.1 The Impact of Language on Evaluation Results

We selected five questions and asked gpt4o and deepseek in English, Chinese, and Spanish. We found that the content and expression of the answers were basically the same. This phenomenon is closely related to the characteristics of the corpus used to train LLMs. In particular, the training corpus of most language models is often dominated by English, so they basically rely on converting other languages into English for understanding and generation when processing multilingual tasks. In addition, our experimental design does not involve complex semantic expressions.The results can be summarized as follows: The language of the prompt has no significant effect on the performance of LLMs on content and expressions of daily life. The performance of LLMs is probably more determined by their training corpus and model architecture.

However, we found that participants experienced difficulties in understanding the meaning of questions when completing questionnaires in non-native languages, often needing translation tools or internet searches. This created barriers to understanding, which ultimately impacted the experimental results. To minimize misunderstandings and ensure that the data accurately reflects human and LLM thinking and decision-making, we conducted the survey experiment in Chinese.

4.2 The Difference in Average Accuracy Between Humans and LLMs

The average accuracy rate of human responses(81.17%) is lower than that of the worst-performing LLM(93.35%), while high-quality human responses, particularly those from experts(100%), show higher average full accuracy than the best-performing LLMs. This suggests that expert cognition is superior to LLMs and that LLMs perform more consistently. From Table 2, the reasons for this difference are apparent:

Humans have the lowest accuracy rate in answering questions about intellectual errors, with an average partial accuracy rate of 89.93%.This shows that there are inevitably situations where humans have incomplete knowledge in specific fields, such as chemistry, history, biology, etc., which were covered in the questionnaire. However, due to training on large corpora such as Wikipedia, which has billions of parameters, LLM models contain a large amount of knowledge, and there are fewer gaps in the knowledge of experts.Humans may pay more attention to the persuasiveness of information before focusing on the specific content. Absolutes such as "just", "all" and "certainly" can easily trigger a potential negative attitude, so when faced with an undisguised type of error, or an error of knowledge or common sense that is not hidden in the semantics, humans are more likely to make the correct judgment.Since the questions in each questionnaire are arranged in the same way (always with the knowledge-based error questions at the beginning), it cannot be ruled out that humans become more sensitive to incorrect information as the number of judgment questions increases.Although the questionnaire prompts strongly recommend that subjects use online searches to supplement their lack of knowledge or experience in certain areas, and this approach is also effective, there are still a large number of invalid performances such as "I don't know" and "I'm not sure", indicating that in the absence of offline supervision, there is still low-quality data due to factors such as the time and attitude of the subjects in the absence of offline supervision.

LLMs have the lowest accuracy in answering logical error questions, with an average partial accuracy of 89.93%, indicating that LLM reasoning models (such as chains, trees, or graphs) are vulnerable to misinformation, confusing the logic of premises and conclusions, or failing to capture the implied logic in the context. Human reasoning on logical problems tends to break down the problem into smaller parts. For example, when asked whether "children learn new things more easily than adults", humans will think about "all children?"and "all adults?"and ultimately arrive at the correct judgment. However, in general, LLMs' perform more robustly and react faster than humans on the dataset.

4.3 Human and LLMs' Tendencies to Misinformation

Human performances showed hesitation, neutrality and other ambiguous attitudes, such as "it depends" and "it depends on the person", which were much higher than those of the large model in the five types of questions. Since many questions can indeed be considered "partially correct" but as a whole can also be considered "wrong", the subjects' judgments are actually correct when viewed in conjunction with the subjects' explanations. This also reflects humans' underlying dialectical approach to complex attitude, this answer may be more flexible and comprehensive; or due to cultural background and other reasons, these subjects' judgments may be more moderate, and they tend not to make a yes or no judgment. Human performances are far more complex than those of the large model. Due to differences in human cognitive situations, people may give irrelevant answers or raise doubts. Such answers are usually emotional, and for some questions, the answers also show bias, superstition, etc., which is completely absent in the large model.

The above is common to humans and must be present in the thought process of experts as well, but the experts may have gone through a process of negation-affirmation-negation or affirmation-negation, combined with validation and questioning, to ultimately present an average accuracy higher than that of the larger model. LLMs lack this kind of human-like self-correction mechanism.

In conclusion, comparing human and LLMs' performance provides a useful framework for the Turing Test, offering insights into potential strategies for improving LLMs (Sejnowski, 2023; Meng, 2024). In both the present and future, LLMs, as tools, need to exhibit more human-like characteristics in some scenarios—such as expressing "emotion"—while in others, they should prioritize reliability and robustness, as reflected in the accuracy rates observed in this study.

4.4 Further Exploration

To gain further insight into human performances, we conducted a small-scale study designed to explore the differences in thought processes between human experts and LLMs. We invited two human participants aged 20 and 21 with at least a bachelor's degree in education, and increased the questionnaire payment to $10 per questionnaire.

The results showed that the participants were 100% correct and that their thinking was significantly different from that of the LLMs throughout the process, in line with the analysis presented in 5.2 and 5.3. Human experts, represented by highly educated humans, have an advantage in recognizing misinformation: they not only accurately identify errors, but also provide reasonable and well-founded explanations, demonstrating strong critical thinking skills, a rich knowledge base, and a rigorous reasoning process. In contrast, although LLMs show efficiency and consistency in processing information, they lack contextual understanding and in-depth reasoning.

4.5 Accuracy Across Different Age Groups

The accuracy rates for different age groups were as follows: under 18, 18–25 years, 25–30 years, 31–40 years, 41–50 years, 51–60 years, and over 60, with complete accuracy rates of 69.17%, 86.86%, 81.29%, 82.00%, 79.20%, 79.33%, and 79.55%, respectively. Given that there were only 3 participants under 18 and 4 participants over 60, these results may not hold significant reference value. Excluding these two age groups, the 18-25 age group had the highest accuracy rate, and this age group may be at the peak of cognition or judgment. This age group generally exhibits faster reaction times and higher information processing efficiency.

4.6 Limitations and Future Work

The five baselines we selected, while certain advanced and versatile models currently available, do not fully represent the performance of all large models. In terms of human participants, we did not conduct a large-scale study.

In the future, it is interesting to conduct perform more granular research by designing additional types of error-related questions, such as context errors, causal errors, semantic errors, and cognitive biases. We also plan to invite more human participants to participate in our experiments, particularly considering diverse linguistic and cultural backgrounds.

5 Conclusion

This study systematically compared the misinformation detection abilities of humans and LLMs, revealing important insights into their different cognitive mechanisms.

The results of the study showed that while LLMs were more accurate (93.35%-97.94%) than humans (81.17%) on the five error types (factual, common sense, reasoning, misleading, and logical), expert humans outperformed state-of-the-art LLMs, and that the five performance modes highlighted fundamental divergences: the LLMs relied primarily on model architecture and trained data reasoning, whereas humans use dialectical reasoning and self-correction, albeit with greater variability due to educational background and cognitive biases.

Research has shed light on the advantages of LLMs in terms of scalability and consistency, as well as human adaptability in complex, context-dependent scenarios. Our dataset and framework lay the foundation for developing powerful error message mitigation tools in AI-enhanced information ecosystems.Future work should extend the error types to incorporate semantic and cultural biases and enhance the LLMs' inference architecture through neuro-semantic integration.

References

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., He, B., Jiang, S., & Dong, B. (2023). Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Boissonneault, D., & Hensen, E. (2024). Fake news detection with large language models on the liar dataset.

Chen, C., & Shu, K. (2024). Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, *45*(3), 354–368.

Chen, H., Guo, H., Hu, B., Hu, S., Hu, J., Lyu, S., Wu, X., & Wang, X. (2024). A self-learning multimodal approach for fake news detection. *arXiv preprint arXiv:2412.05843*.

Chen, M., Wei, L., Cao, H., Zhou, W., & Hu, S. (2023). Can large language models understand content and propagation for misinformation detection: An empirical study. *arXiv preprint arXiv:2311.12699*.

Choi, E. C., & Ferrara, E. (2024). Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. *Companion Proceedings of the ACM on Web Conference 2024*, 1441–1449.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22105–22113.

Jin, W., Gao, Y., Tao, T., Wang, X., Wang, N., Wu, B., & Zhao, B. (2025). Veracity-oriented context-aware large language models–based prompting optimization for fake news detection. *International Journal of Intelligent Systems*, *2025*(1), 5920142.

Kamoi, R., Das, S. S. S., Lou, R., Ahn, J. J., Zhao, Y., Lu, X., Zhang, N., Zhang, Y., Zhang, R. H., Vummanthala, S. R., et al. (2024). Evaluating llms at detecting errors in llm responses. *arXiv preprint arXiv:2404.03602*.

Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Meng, J. (2024). Ai emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences*, *121*(10), e2401336121.

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, *57*(10), 265.

Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., Godbout, J.-F., & Rabbany, R. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural computation*, *35*(3), 309–342.

Sun, X., Chen, Y., Huang, Y., Xie, R., Zhu, J., Zhang, K., Li, S., Yang, Z., Han, J., Shu, X., et al. (2024). Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). Ai hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, *11*(1), 1–14.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, X., Zhang, W., Koneru, S., Guo, H., Mingole, B., Sundar, S. S., Rajtmajer, S., & Yadav, A. (2024). The reopening of pandora's box: Analyzing the role of llms in the evolving battle against ai-generated fake news. *arXiv preprint arXiv:2410.19250*.

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824–24837.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 1–65.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136.

Xu, J., Han, L., Sadiq, S., & Demartini, G. (2024). On the role of large language models in crowdsourcing misinformation assessment. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 1674–1686.

Yang, Y., Jin, Q., Leaman, R., Liu, X., Xiong, G., Sarfo-Gyamfi, M., Gong, C., Ferrière-Steinert, S., Wilbur, W. J., Li, X., et al. (2024). Ensuring safety and trust: Analyzing the risks of large language models in medicine. *arXiv preprint arXiv:2411.14487*.

Zanartu, F., Otmakhova, Y., Cook, J., & Frermann, L. (2024). Generative debunking of climate misinformation. *arXiv preprint arXiv:2407.05599*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.