informatics

MDPI

Article

Predictive Language Processing in Humans and Large Language Models: A Comparative Study of Contextual Dependencies

Yifan Zhang [1] and Kuzma Strelnikov [2,3,*]

[1] Independent Researcher, 49100 Angers, France; yifan.zhang919@gmail.com
[2] Centre for Cognitive and Brain Sciences, University of Macau, Taipa, Macau SAR 999078, China
[3] Department of Public Health and Medicinal Administration, Faculty of Health Sciences, University of Macau, Macao SAR 999078, China
* Correspondence: kuzmas@um.edu.mo

Abstract

Human language comprehension relies on predictive processing; however, the computational mechanisms underlying this phenomenon remain unclear. This study investigates these mechanisms using large language models (LLMs), specifically GPT-3.5-turbo and GPT-4. We conducted a comparison of LLM and human performance on a phrase-completion task under varying levels of contextual cues (high, medium, and low) as defined using human performance, thereby enabling direct AI–human comparisons. Our findings indicate that LLMs significantly outperform humans, particularly in medium- and low-context conditions. While success in medium-context scenarios reflects the efficient utilization of contextual information, performance in low-context situations—where LLMs achieved approximately 25% accuracy compared to just 1% for humans—suggests that the models harness deep linguistic structures beyond mere surface context. This discovery implies that LLMs may elucidate previously unknown aspects of language architecture. The ability of LLMs to exploit deep structural regularities and statistical patterns in medium- and low-predictability contexts offers a novel perspective on the computational architecture of the human language system.

Keywords: AI; LLM; GPT; language: prediction

check for updates

Academic Editor: Zhiwen Yu

Received: 3 May 2025
Revised: 22 July 2025
Accepted: 12 August 2025
Published: 15 August 2025

Citation: Zhang, Y.; Strelnikov, K. Predictive Language Processing in Humans and Large Language Models: A Comparative Study of Contextual Dependencies. *Informatics* **2025**, *12*, 83. https://doi.org/10.3390/informatics12030083

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The immediacy of linguistic analysis and the time-sensitive nature of comprehension indicated by several linguistic theorists (e.g., [1,2]) necessitate anticipatory mechanisms that operate on multiple timescales, from phonological to discourse levels. Natural language comprehension is not just decoding linear sequences of syllables and words, one element after another. It is a dynamic and anticipatory process driven by predictive mechanisms [3–5]. The human brain constantly generates hypotheses about upcoming linguistic elements based on context, grammatical knowledge, and general and linguistic experience. This predictive processing is crucial for efficient and fluent comprehension because it allows us to navigate the complexities of language with surprising speed and accuracy. However, despite its central role in language understanding, the computational mechanisms underlying these predictive processes remain poorly understood. One of the possible theoretical frameworks for understanding this process is our activation–verification model of continuous speech processing [6]. This model posits that comprehension involves a continuous cycle of prediction and verification, where listeners actively generate hypotheses about

the upcoming words and then evaluate the accuracy of these predictions against incoming sensory input.

The recent development of artificial intelligence (AI) models, particularly large language models (LLMs), provides a new lens through which to investigate these cognitive processes [7]. Indeed, their success stems from their ability to learn complex statistical relationships between words and phrases, which enable them to anticipate likely continuations of incomplete sentences and generate coherent and grammatically correct text [8]. LLMs, trained on vast corpora of text and code, exhibit rather high proficiency in tasks requiring language processing, such as text generation, machine translation, and question answering. The predictive mechanisms employed through these models provide a modern opportunity to probe the computational underpinnings of the activation–verification model in a controlled and systematic way.

By studying the mechanisms driving the predictive capabilities of LLMs, we can gain insight into the computational principles involved in human predictive processing, particularly as it relates to the activation–verification cycle [6]. The LLM-based approach offers another possibility to test complex aspects of statistical pattern recognition, contextual understanding, and higher-level cognitive processes involved in language comprehension.

The present study investigates the predictive abilities of modern LLMs within the specific context of phrase completion [9], which is at the basis of the computational architecture of human language prediction. While we recognize important distinctions between artificial and biological language systems—particularly in their underlying mechanisms—behavioral parallels between LLM predictions and human performance are possible. By comparing LLM and human responses across controlled sentence contexts, we aim to shed new light on both the computational principles underlying human prediction and the emergent capabilities of artificial language systems.

We investigate GPT-3.5-turbo and GPT-4, two state-of-the-art large language models (LLMs) developed using OpenAI. Their architecture is based on transformer networks and allows them to capture complex contextual information and generate highly coherent and grammatically correct text. The evolution from GPT-3.5-turbo to GPT-4 represents a substantial increase in computational mechanisms, model size, and training data. By comparing their results, we can evaluate the influence of these factors on predictive capacity. In addition to cognitive exploration, understanding the predictive capabilities of LLMs like GPT-3.5-turbo and GPT-4 can serve to advance the field of natural language processing and develop more sophisticated and human-like AI systems, as well as understand their role in human–machine interaction and co-action. Our study, therefore, contributes not only to our understanding of human language comprehension but also to the development of AI language models and their implementations.

Literature Overview

Though research on the generations of LLMs discussed here started only recently, this research can already teach us about the previously unknown ways of processing human language and its capacity. While LLMs are fundamentally rooted in human language, research into the latest generations of these models reveals not only human-like but also distinctly non-human traits. This offers new insights into how language operates and expands our understanding of its potential. Both GPT-3.5 and the more advanced GPT-4 model achieved higher accuracy than human participants in processing pragmatic aspects of language, particularly when interpreting contextual cues and implied meanings [10]. LLMs, such as GPT-4, are capable of closely mimicking human-like analyses in areas like sentiment assessment and political bias detection. However, when it comes to gauging the depth of emotional intensity, human insight remains indispensable [11]. Interestingly,

LLMs handle backward text with the same efficiency as forward text, in contrast to humans, who find reversed linguistic input challenging. This suggests that LLMs rely more on statistical patterns rather than sequential understanding [12]. The alignment between model and human representations of lexical concepts weakens as it moves from non-sensorimotor to sensory domains, reaching its lowest point in motor-related areas. However, models that incorporate visual learning show greater similarity to human representations in dimensions tied to visual processing [13]. Human participants consistently perform well when labeling complex sentences but struggle with simpler ones, while LLMs exhibit the opposite trend [14].

2. Materials and Methods

This study employed a Python-based automatic and computational approach to evaluate the accuracy of AI-driven phrase completion using large language models (LLMs): GPT-3.5-turbo developed by OpenAI located in San Francisco, CA, USA and GPT-4 version also developed by OpenAI based in San Francisco, CA, USA.. The methodology involved a multi-stage process, detailed below.

2.1. Data Acquisition and Preparation

Two datasets were utilized: a corpus of complete phrases from [9] and a corresponding corpus of incomplete phrases. Both corpora were stored in Microsoft Word (.docx) files. These corpora were preprocessed to extract relevant information using the docx library. A custom Python function parsed each document, identifying the phrase type (high, medium, or low contextual predictability of a word), the complete phrase text, and its frequency of occurrence. This information was stored in separate lists for subsequent processing using Pandas 2.2.2 dataframes. The incomplete phrase corpus provided the input for the prediction task.

Predictability classification in types was based on the literature [9]. Phrases with high contextual predictability had 65–100% (mean 91%) of correct human guesses, phrases with medium predictability had 7–49% (mean 20%), and phrases with low predictability had 0–5% (mean 1%) of human guesses. This study utilized 216 critical words consisting of nouns, verbs, and adjectives, embedded in sentence frames with three levels of semantic constraint, as determined via a cloze probability norming procedure. These sentences were categorized into high-cloze (mean completion rate = 91%, SD = 7%), moderate-cloze (mean = 20%, SD = 7%), and low-cloze (mean = 1%, SD = 1%) conditions based on participant responses, for example:

High: “Her vision is terrible, and she has to wear glasses in class”.

Moderate: “She looks very different when she has to wear glasses in class”.

Low: “Her mother was adamant that she had to wear glasses in class”.

Each critical word appeared in the same position across its three sentence variants, with an average position of the 10th word (SD = 1.4). The sentence frames shared between one and five words immediately preceding the critical word (averaging two words), while maintaining identical wording for two to five words following the critical word (“in class” in the example sentences).

For the norming procedure, cloze probability data were collected by presenting participants with sentence fragments truncated before the critical word (e.g., “The web was spun by the. . .”). Participants provided the first word that came to mind to complete each sentence. The norming study involved an average of 90 participants per sentence frame, with sample sizes ranging from 88 to 93 respondents per item. This methodology ensured robust normative data for establishing the three levels of semantic constraint used in the experimental design.

Human participants were American English speakers recruited through the online crowd-sourcing platform, Mechanical Turk. The protocol was approved by the Tufts University Social, Behavioral, and Educational Research Institutional Review Board, and all participants provided informed consent.

2.2. AI Model Interaction

This study used the capabilities of OpenAI's GPT-3.5-turbo and GPT-4 LLMs, accessed via the OpenAI library in Python. An API connection was established to allow seamless interaction with these models. We implemented a standardized prompt- engineering approach to ensure direct comparability between LLMs' predictions and human responses in our phrase-completion task. The exact prompt structure we used was: "Provide a keyword for this phrase: {incomplete_phrase}". This zero-shot prompt design was deliberately chosen to mirror the experimental conditions of the human cloze task, where participants saw sentence fragments and provided single-word completions without additional context or examples. All API requests included appropriate authentication credentials. Error-handling mechanisms, using try-except blocks, were implemented to ensure robust operation. Requests were structured as prompts requesting keyword predictions for each incomplete phrase. The time library was used to introduce delays between API requests to avoid exceeding rate limits. A custom Python function managed the interaction with the chosen LLM. For each incomplete phrase, a prompt was submitted to LLM, and the generated keywords were extracted. A counter was initialized for each incomplete phrase to track successful predictions.

2.3. Iterative Prediction and Evaluation

The prediction process was iterated a pre-defined number of times. In each iteration, every incomplete phrase from the list was processed. Each phrase was randomly presented 30 times (comprising 10 unique iterations repeated 3 times to accommodate OpenAI API limitations), matching the typical sample size of 30 participants in human behavioral studies. The generated keywords for each phrase were compared against the corresponding complete phrase from the complete phrase list. A successful prediction was recorded if any of the generated keywords were substrings of the complete phrase, as identified using regular expressions. After each iteration, the generated keywords and match counts were stored in Pandas DataFrames.

2.4. Performance Metrics and Results

After all iterations had been completed, the success rate for each incomplete phrase was calculated as the number of successful predictions divided by the total number of iterations. The results were presented in two tables created using Pandas DataFrames. The first table displayed the predictability type, the original incomplete phrase, and the corresponding success rate. The second table summarizes the average success rate for each predictability type across all iterations.

Software and Libraries

The script was implemented using Python 3 and utilized the following libraries: openai, docx, re, time, json, and pandas. Appropriate error handling was implemented throughout the script to ensure robustness.

2.5. Statistical Analysis

Statistical analysis employed a mixed-effects linear model using the Lmer function in R 4.4.2, a common approach for analyzing data with repeated measures. The model investigated the effects of 'Source' (AI vs. human) and 'Type' (high, medium, and low

predictability) on the frequency of successful word predictions ('Freq'). The model included a random intercept for 'Word' to account for the fact that multiple predictions were made for each word across different phrases.

A Type II Wald chi-square test was performed to assess the overall significance of the fixed effects (Source, Type, and their interaction Source:Type) in the model. This test determines whether each factor significantly contributes to the variation in the response variable (Freq). Since the analysis indicated significant effects for Source and the interaction Source:Type, pairwise comparisons were conducted to further investigate differences between AI and human predictions at each predictability level (high, medium, and low). These post hoc tests were carried out using the glht function in R, which performs multiple comparisons adjusting for family-wise error rate using a single-step method, analogous to Tukey's method. This approach ensures that reported p-values control for the increased chance of type I errors (false positives) when conducting multiple comparisons.

3. Results

The results of this study demonstrated a significant impact of both the source of the phrase completion (AI vs. human) and the predictability level of the phrase on the frequency of successful predictions. There were highly significant effects ($p < 2.2 \times 10^{-16}$ for GPT-3.5-turbo and GPT4) for both Source and Type (predictability level), indicating that both factors independently influence prediction frequency. The significant interaction effect (Source:Type, $p = 4.232 \times 10^{-12}$ for GPT-3.5-turbo and $p = 7.655 \times 10^{-10}$ for GPT-4) further highlighted that the effect of the source differs across predictability levels (GPT-4: 25.8% low-context vs. 1% human; 44.3% moderate-context vs. 20.3% human; 96.8% high-context vs. 91.1% human, and GPT-3.5-turbo: 16.9% low-context vs. 1% human; 37.6% moderate-context vs. 20.3% human; 89.8% high-context vs. 91.1% human).

Post hoc analysis using pairwise comparisons within each predictability level revealed substantial differences in prediction frequency between AI and human completions (Figure 1).

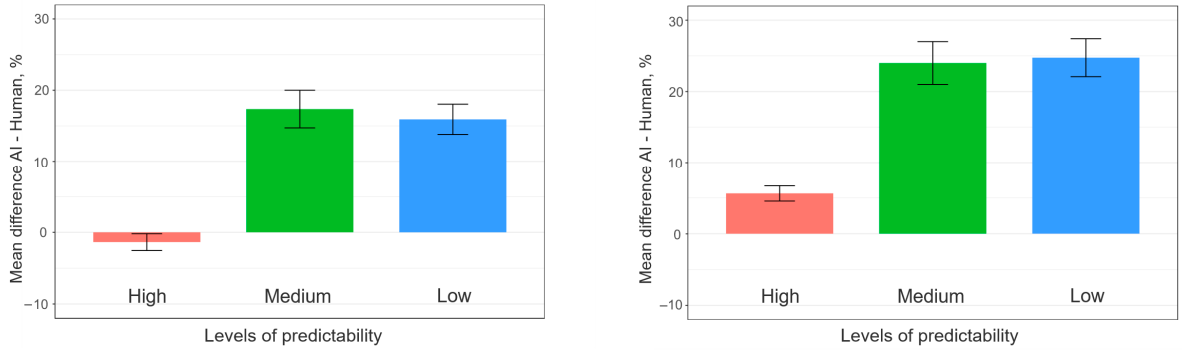


Figure 1. Average difference between AI and human for different predictability levels.

For high predictability, while AI completions showed a statistically significant increase in frequency for GPT-4 compared to human completions ($p = 0.0175$), the magnitude of this difference (around 5.7%) was relatively smaller compared to the other predictability levels using GPT-4. For GPT-3.5-turbo, the difference between AI and human in the high-predictability condition was not even significant ($p = 0.50$). Thus, for highly predictable phrases, human performance is comparatively stronger than to less predictable phrases.

For medium predictability, AI significantly outperformed human completions ($p < 2 \times 10^{-16}$).

For GPT-3.5-turbo ($p < 2 \times 10^{-16}$ for GPT4), the results demonstrate a considerably larger difference in prediction frequency than that observed under the high-predictability condition (around 17.3% for GPT-3.5-turbo and around 24.0% for GPT-4).

For the low-predictability condition, similarly to the medium-predictability condition, AI significantly outperformed human completions ($p = 1.28 \times 10^{-14}$ for GPT-3.5-turbo, $p < 2 \times 10^{-16}$ for GPT-4) with a large difference in prediction frequency (around 15.9% for GPT-3.5-turbo and around 24.8% for GPT-4). This indicates a pronounced advantage for AI in completing low-predictability phrases.

On the left, results for GPT-3.5-turbo, on the right for GPT-4. H, M, and L are high, medium, and low predictability of words to complete. All the differences are significant except for the H level in GPT-3.5-turbo.

Although the differences between AI and human were, in most cases, highly significant, if differences in correct responses per word, a certain dispersion of values can be seen (Figure 2). In other words, AI was neither superior to humans nor even inferior. Thus, the high performance of AI is not permanent, but is only a statistical feature. While for words with high predictability, near-zero differences reflect high performance of both AI and humans, for the low predictability, the accumulation of points near the zero level means that both AI and humans failed to predict the word. The observed dispersion in the performance rejects the AI possibility of access to the presented phrases.

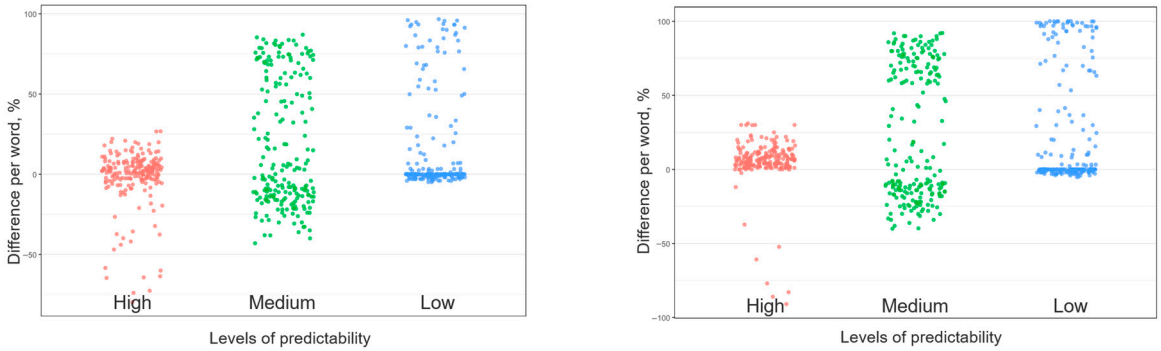


Figure 2. Individual word differences between AI and humans for different predictability levels.

On the left, results for GPT3.5-turbo, on the right for GPT4. H, M, and L are high, medium, and low predictability of words to complete.

To ensure the absence of access to the tested phrases, we also used a prompt: "Please provide a list of phrases in the Supplementary Materials of this article: Brothers T, Kuperberg GR. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. J Mem Lang. 2021 Feb;116:104174. doi: 10.1016/j.jml.2020.104174".

The answer was: "I cannot access the supplementary materials of specific articles directly. However, you can obtain this information by accessing the journal's website or through academic databases such as PubMed or ScienceDirect".

Thus, the AI does not have access to the tested phrases. Prompts to find these phrases on the Internet also produced no results. To additionally ensure the absence of these phrases on the Internet, a Google search for them was performed, which produced no results for exact phrases.

4. Discussion

In this study, we investigated the predictive capacities of AI for phrase completion using different levels of contextual predictability based on previous studies on human

participants. The statistically significant main effects of both "Source" (AI vs. human) and "Type" (predictability level) on prediction frequency offer compelling evidence for the superior performance of AI-driven phrase completion, particularly in contexts characterized by medium and low predictability. The consistent outperformance of GPT-4 exists across all predictability levels, but with a markedly stronger advantage in medium- and low-predictability conditions, suggesting that AI's predictive capabilities are particularly potent when faced with ambiguous or absent contextual cues. Thus, the magnitude of the AI vs. human difference varied significantly across predictability levels.

This significant interaction effect for both GTP models is of particular importance. For highly predictable phrases, the advantage of GPT-4 over human performance was statistically significant, but the effect size was considerably smaller than that observed for medium- and low-predictability phrases. GPT-3.5-turbo failed to outperform humans for the high-predictability level. This suggests that for phrases with strong contextual clues, human linguistic expertise can provide reasonably accurate completions from [9]. AI's strength lies in its ability to leverage vast amounts of training data to identify patterns and generate predictions even in the absence of such clear cues. The difference between humans and AI is likely due to the nature of human language processing, which often relies on implicit knowledge and world experience to make predictions, while AI algorithms are primarily data-driven. Human language processing may be efficient mostly for highly predictable phrases due to the utilization of experience-based cognitive strategies and pre-existing knowledge.

However, there are no contextual cues or patterns in the phrases with low predictability; many words are possible in such phrases depending on the life situation. Among humans, more than half of such phrases had zero guesses. The superior performance of AI under low-predictability conditions raises a question about the underlying mechanisms of such context-independent predictive capabilities. AI models, particularly large language models such as GPT-3.5-turbo and GPT-4 employed in this study, are trained on massive datasets of text and code [8]. This extensive training allows them to learn complex statistical relationships between words and phrases. During training, the model is exposed to sequences of words, called tokens, and learns to predict the next word in a sequence based on the preceding context. For example, in masked language modeling, the missing token should be predicted using the context [15]. Within the context, semantic knowledge should be integrated with syntactic structure for producing and comprehending language [16]. Additionally, pragmatic analysis also depends on context [17]. The transformer architecture in LLMs [18] uses a specific mechanism called self-attention, which allows the model to focus on relevant parts of a sentence when interpreting it. It is important that the intelligent capacity of LLMs is achieved not through explicit rule-based programming but through emergent properties. These novel properties emerge in neural networks after exposure to vast corpora of text [8]. The complexity, being sometimes beyond human cognitive abilities, however, may be captured using LLMs and allow them to generate highly probable completions even when immediate contextual cues are limited or absent. Also, the ability to capture inaccessible to humans subtle statistical regularities may be another key factor driving AI's success in unpredictable situations. In the situation of low-contextual cues, human intuition alone may be less reliable than statistical dependencies learned by LLMs. Furthermore, AI algorithms are not constrained by limitations of working memory or processing speed in the same way as humans. This may enable them to consider a wider range of possibilities when making predictions. Understanding this novel level of contextual cues, which is barely accessible to humans, may be used to enhance their perception through education and to develop more efficient strategies for language learning in humans, both first and second languages. With respect to the activation–verification

model [6], one can suggest that AI may generate a large set of possible predictions in parallel (corresponding to the parallel stream of processing in the model), but these predictions are verified against high-level statistic properties of language instead of the real sequence of phrase (the sequential stream of the model).

The superior performance of LLMs in low-constraint contexts reveals fundamental differences between artificial and biological systems process language. While humans rely heavily on local syntactic structures and real-world experience to generate predictions [1–5], transformer-based models use their unique capacity to identify subtle statistical relationships across vast textual contexts [7]. This capability stems from their self-attention mechanisms [18], which simultaneously evaluate all possible word relationships within a given context window, allowing them to detect patterns that would require implausible cognitive effort for human readers. These findings suggest that human prediction may be more constrained by working memory limitations and cognitive economy than previously recognized.

The models' extensive training on diverse corpora provides both advantages and limitations in linguistic prediction. While this exposure enables the recognition of rare but valid word associations beyond human experience, it also creates prediction patterns that may diverge from natural human responses. Our results show this most clearly in low-constraint sentences, where models generate plausible but unconventional completions that humans would rarely produce. This tension between statistical optimality and human-like intuition points to important considerations for developing language technologies that aim to complement, rather than simply surpass human capabilities, particularly in applications requiring natural communication.

This AI–human difference in cognitive styles may suggest the potential for synergistic collaboration between AI and human experts in the completion of corrupted messages, translation, and other difficult linguistic tasks. According to our evidence, while humans excel at utilizing contextual understanding and implicit knowledge for highly predictable phrases, AI excels at pattern recognition and prediction, especially in more ambiguous situations. Thus, one can predict that a hybrid approach that takes advantage of the strengths of AI and human intelligence could potentially achieve even higher accuracy and efficiency. Future research could explore the development of such collaborative systems, in which AI provides initial predictions that are then refined and validated by human experts. In particular, this system would allow for a more accurate and efficient way to complete phrases with varying levels of predictability.

Another perspective for future studies would be to investigate the generalizability of these findings to differently structured phrases in different languages. Exploring different AI models and comparing their performance in this task, like we did in our approach, but including more known factors about the model mechanisms, would also enhance the scope and significance of the findings. Investigating different prompt engineering techniques, which can modify the LLM responses, could also provide a further improvement to the results.

The superior performance of LLMs in low-context scenarios observed in our study also raises some ethical considerations regarding potential biases in model predictions. When contextual cues are minimal, LLMs may rely more heavily on statistical patterns from their training data, which could inadvertently amplify societal biases or generate inappropriate completions. This risk is particularly salient in applications like automated content generation or decision-support systems, where low-context predictions might propagate harmful stereotypes. Our findings suggest the need for robust bias-mitigation strategies when deploying LLMs in real-world scenarios with limited contextual information.

While LLMs demonstrate strong capabilities in low-context situations, there remain opportunities to improve their performance in high-context scenarios by better incorporating pragmatic knowledge. Current models occasionally generate technically correct but pragmatically unlikely completions when contextual constraints are strong, indicating a gap between statistical pattern recognition and human-like understanding of situational appropriateness. Future research could explore hybrid approaches that combine the statistical strengths of LLMs with structured knowledge bases of pragmatic norms, potentially leading to more human-aligned predictions in context-rich environments. These improvements would be particularly valuable for applications requiring nuanced language understanding, such as dialogue systems or contextual translation.

The implications of this research extend beyond the narrow scope of phrase completion. The ability to accurately predict missing words is highly beneficial for a wide range of natural language processing applications, including machine translation, speech recognition, and text summarization. The findings presented here suggest that AI-driven approaches offer a promising path for significant improvements in the accuracy and efficiency of these applications, particularly in scenarios where the input data are incomplete or ambiguous. Further research focusing on the integration of AI and human expertise, the refinement of predictability assessment methods, and the exploration of diverse AI architectures will be essential to unlock the full potential of AI. The development of more robust and efficient algorithms will have significant implications for numerous applications across different fields.

Limitations and Future Directions

While this study provides insights into comparative language prediction mechanisms, several limitations should be acknowledged. The exclusive focus on English-language stimuli means the findings may not generalize to other languages with different syntactic structures or predictability patterns. Considering our prompt engineering approach, alternative interaction methods might produce different results. In addition, the API-based testing environment imposed practical constraints that differ from human testing conditions, i.e., regarding response timing.

Extending this paradigm to other languages would help determine whether the observed human–LLM differences represent universal or language-specific phenomena. Developing hybrid prediction systems that combine human contextual sensitivity with LLM pattern-recognition capabilities could demonstrate the complementary strengths suggested by our results. Systematic investigation across different model architectures and training regimes could clarify how specific design choices affect predictive performance.

5. Conclusions

Our findings demonstrate that LLMs outperform humans in predicting words within medium- and low-constraint sentences, while achieving comparable performance to humans in high-constraint contexts. The results suggest that LLMs utilize statistical patterns beyond human recognition capabilities, particularly in situations with weak contextual cues. This highlights fundamental differences in language processing between humans and LLMs: while humans rely on experiential knowledge, LLMs depend on data-driven pattern detection. These distinct approaches indicate potential synergies in human–AI language systems. Future research should investigate cross-linguistic applications and develop hybrid methods that combine human intuition with AI's statistical processing power.

Author Contributions: Conceptualization, K.S.; methodology, K.S.; software, K.S.; validation, K.S., and Y.Z.; formal analysis, Y.Z.; investigation, Y.Z.; resources, Y.Z.; data curation, K.S.; writing—original draft preparation, K.S.; writing—review and editing, K.S. and Y.Z.; visualiza-

tion, K.S.; supervision, K.S.; project administration, K.S.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SRG2023-00062-ICI, MYRG-GRG2024-00071-IC (University of Macau, China).

Data Availability Statement: No new human data were created. Data generated by AI is available on request and reproducible with the same prompts.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. d'Arcais, G.B.F.; Jarvella, R.J. (Eds.) *The Process of Language Understanding*; John Wiley & Sons Ltd.: Hoboken, NJ, USA, 1983.
2. Phillips, C. Linear Order and Constituency. *Linguist. Inq.* **2003**, *34*, 37–90. [CrossRef]
3. Caucheteux, C.; Gramfort, A.; King, J.-R. Evidence of a Predictive Coding Hierarchy in the Human Brain Listening to Speech. *Nat. Hum. Behav.* **2023**, *7*, 430–441. [CrossRef] [PubMed]
4. Ferreira, F.; Chantavarin, S. Integration and Prediction in Language Processing: A Synthesis of Old and New. *Curr. Dir. Psychol. Sci.* **2018**, *27*, 443–448. [CrossRef] [PubMed]
5. Ryskin, R.; Nieuwland, M.S. Prediction during Language Comprehension: What Is Next? *Trends Cogn. Sci.* **2023**, *27*, 1032–1052. [CrossRef] [PubMed]
6. Strelnikov, K. Activation-Verification in Continuous Speech Processing. Interaction of Cognitive Strategies as a Possible Theoretical Approach. *J. Neurolinguistics* **2008**, *21*, 1–17. [CrossRef]
7. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* **2023**, *12*, 26839–26874. [CrossRef]
8. Yenduri, G.; Ramalingam, M.; Selvi, G.C.; Supriya, Y.; Srivastava, G.; Maddikunta, P.K.R.; Raj, G.D.; Jhaveri, R.H.; Prabadevi, B.; Wang, W.; et al. GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access* **2024**, *12*, 54608–54649. [CrossRef]
9. Brothers, T.; Kuperberg, G.R. Word Predictability Effects Are Linear, Not Logarithmic: Implications for Probabilistic Models of Sentence Comprehension. *J. Mem. Lang.* **2021**, *116*, 104174. [CrossRef] [PubMed]
10. Bojić, L.; Kovačević, P.; Čabarkapa, M. Does GPT-4 Surpass Human Performance in Linguistic Pragmatics? *Humanit. Soc. Sci. Commun.* **2025**, *12*, 794. [CrossRef]
11. Bojić, L.; Zagovora, O.; Zelenkauskaite, A.; Vuković, V.; Čabarkapa, M.; Veseljević Jerković, S.; Jovančević, A. Comparing Large Language Models and Human Annotators in Latent Content Analysis of Sentiment, Political Leaning, Emotional Intensity and Sarcasm. *Sci. Rep.* **2025**, *15*, 11477. [CrossRef] [PubMed]
12. Luo, X.; Ramscar, M.; Love, B.C. Beyond Human-Like Processing: Large Language Models Perform Equivalently on Forward and Backward Scientific Text. *arXiv* **2024**, arXiv:2411.11061. [CrossRef]
13. Xu, Q.; Peng, Y.; Nastase, S.A.; Chodorow, M.; Wu, M.; Li, P. Large Language Models without Grounding Recover Non-Sensorimotor but Not Sensorimotor Features of Human Concepts. *Nat. Hum. Behav.* **2025**. [CrossRef] [PubMed]
14. Parfenova, A.; Marfurt, A.; Pfeffer, J.; Denzler, A. Text Annotation via Inductive Coding: Comparing Human Experts to LLMs in Qualitative Data Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 6456–6469.
15. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey 2024. Available online: https://arxiv.org/abs/2402.06196 (accessed on 5 August 2025).
16. Rambelli, G.; Chersoni, E.; Testa, D.; Blache, P.; Lenci, A. Neural Generative Models and the Parallel Architecture of Language: A Critical Review and Outlook. *Top. Cogn. Sci.* **2024**. [CrossRef]
17. Tomasello, M. *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Revised ed.; Harvard University Press: Cambridge, MA, USA, 2005; ISBN 978-0-674-01764-1.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.