Univerzita Karlova v Praze

Filozofická fakulta

Ústav obecné lingvistiky

Bakalářská práce

Jana Šimsová

Examining Cognitive Abilities and Multilingual Performance of Large Language Models: A Comparative Analysis of GPT Models

Praha, 2024 Vedoucí práce: PhDr. Jiří Milička, Ph.D.

Největší poděkování patří mému vedoucímu práce Jiřímu Milíčkovi za nesmírnou ochotu a nápomoc v celém procesu.

Velké díky patří Kláře Šimsové, Elišce Šimsové a Benjamínu Šimsovi za odbornou pomoc a Alexi Církovi za podporu.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia nebo k získání jiného nebo stejného titulu.

V Praze, dne 31.7.2024 Jana Šimsová

Abstrakt

Tato bakalářská práce zkoumá kognitivní schopnosti jazykových modelů při zpracování syntakticky zavádějících vět, tzv. *garden-path sentences*. Tento typ nejednoznačných vět často vede čtenáře k tomu, aby je zpočátku interpretovali nesprávně. V následujících experimentech různé modely GPT podstoupí úkol, který obsahuje práci s těmito zavádějícími větami. Porovnáním výkonu modelů GPT s lidskými daty se snažím vyhodnotit kognitivní úroveň jednotlivých modelů. Abych simulovala různé typy experimentálního prostředí, bude výkon jazykového modelu hodnocen jak v laboratorních, tak v domácích podmínkách. Tato prostředí budou zahrnovat různorodou skupinu fiktivních participantů, kteří se budou lišit věkem a pohlavím, aby bylo možné zkoumat možné rozdíly ve schopnostech zpracování napříč těmito demografickými skupinami. Výzkum kognitivních procesů zahrnujících porozumění syntakticky zavádějícím větám u lidských i počítačem simulovaných participantů byl již proveden v angličtině Huffem a Ulakçım (2024). V této studii rovněž zkoumám, zda je výkon jazykového modelu podobný v češtině, jelikož většina tréninkových dat pro velké jazykové modely pochází z angličtiny. Důraz není kladen na to, jak dobře model v experimentech dosahoval správnosti odpovědí, ale spíše na to, jak přesně napodoboval lidské chování při absolvování takového experimentu. Klíčová literatura podporující tuto studii zahrnuje práci o kognitivním zpracování syntakticky zavádějících vět v českém jazyce od Chromého (2022) a o tom, jak LLMs předpovídají lidskou pamětischopnost od Huffa a Ulakçıho (2024).

Klíčová slova*: [GPT, velké jazykové modely, kognitivní úkoly, experimentální prostředí, zahradní věta]*

Abstract

This thesis explores the capabilities of language models in processing garden-path sentences, a class of ambiguous sentences that often lead human readers to initially interpret them incorrectly. The approach in the following experiments involves subjecting a state-of-the-art language model to a series of garden-path sentences previously tested on human participants. By comparing the model's performance with human data, I aim to evaluate its proficiency in handling syntactic ambiguity and reanalysis, with a particular focus on cognitive flexibility – the ability to shift from one interpretation to another when the initial one proves incorrect. To simulate a comprehensive range of environments, the language model's performance will be assessed in both fictional laboratory and home settings. These settings will include a diverse group of fictional participants, spanning various age groups and genders, to investigate potential differences in processing capabilities across these demographic differences.

Research on cognitive processes involving garden-path sentence understanding in both human and machine participants has already been performed in English by Huff & Ulakçı (2024). In this study, I also investigate whether the performance of the language model is similar in Czech, as most of the training data for LLMs comes from English. The focus is not on how well the model performed in terms of response correctness in the experiments, but rather on how accurately it mimicked human behaviour when undergoing such an experiment.

Key literature supporting this study includes works on the cognitive processing of garden-path sentences in Czech language by Chromý (2022) and how LLMs predict human memory by Huff & Ulakçı (2024).

Keywords: *[GPT, Large Language Models, cognitive tasks, experimental environment, garden-path sentence]*

Obsah

List of Abbreviations

API – Application Programming Interface

BPE – Byte-Pair Encoding

CC BY-NC 3.0 – Creative Commons Attribution-NonCommercial 3.0

GA – Garden-path sentence followed by a type A question

GB – Garden-path sentence followed by a type B question

GMM – Generalized Mixed Model

GPT – Generative Pre-training Transformer

HMMs – Hidden Markov Models

LLMs – Large Language Models

NA – Non-garden-path sentence followed by a type A question

NB – Non-garden-path sentence followed by a type B question

NLP – Natural Language Processing

RLHF – Reinforcement Learning from Human Feedback

RNN – Recurrent Neural Network

Introduction

Large Language Models (LLMs) are neural networks trained on large corpora of text data to predict the next word in a sequence. Their architecture enables them to handle long-range dependencies in text and capture complex patterns in language. Key uses of LLMs include text completion and generation, question answering, translation, summarization, sentiment analysis, and code generation; they are also frequently used as conversation agents and chatbots (Brown et al., 2020).

While large language models such as GPT-3 have demonstrated impressive capabilities in generating coherent and contextually appropriate text, there are important considerations regarding the extent to which these abilities reflect genuine cognitive understanding. In their 2021 paper, Bender, Gebru, McMillan-Major, and Shmitchell investigated the limitations and possibilities of large language models. They stated that, overall, while LLMs exhibit remarkable abilities in text generation and pattern recognition, attributing cognitive abilities to them is misleading. Their performance is grounded in surface-level pattern matching rather than in genuine understanding, reasoning, or cognitive processing.

The capabilities of LLMs are, however, progressing at an exceptionally rapid pace, and new findings are being made regarding their cognitive abilities. Milička et al. (2024) explored the capabilities of large language models, specifically GPT-3.5 Turbo and GPT-4 by OpenAI. LLMs are non-deterministic simulators capable of role-playing various personas based on the prompts they receive. The initial conditions – prompting – set by the user determine the behavior and capabilities exhibited by the simulated persona. When an LLM is prompted to simulate a persona, it adopts characteristics and abilities that align with the prompt. The study aimed to investigate whether LLMs can replicate child-like language and cognitive development while performing tasks that require understanding false belief tasks. Both GPT-3.5 Turbo and GPT-4 showed increasing correctness in their responses and a rise in language complexity. This progression mirrored the gradual enhancement observed in linguistic and cognitive abilities during child development, as documented in developmental psychology literature. In other words, they found that large language models are capable of "downplaying" their cognitive abilities to fit the persona they are simulating. This study introduced a new perspective on the limits of cognitive processing by LLMs.

In this paper, I explore another linguistic-cognitive task – the understanding of the garden-path sentence. Garden-path sentences are syntactically ambiguous sentences that initially lead the reader to interpret them incorrectly. An example of a garden-path sentence in English is "The man whistling tunes pianos." When first encountering the sentence, the reader would tend to interpret it as "a man who is whistling tunes"; however, the reader later encounters a syntactic issue, as the last part of the sentence does not fit its initially perceived meaning. The reader must reread the sentence and reanalyse it until they arrive at the correct meaning, which is: "The man who is whistling tunes pianos."

The following experiment in this paper aims to compare the response accuracy to comprehension questions following garden-path sentences between human participants and prompted (simulated) participants by a language model.

The research, however, extends beyond a comparison of human and machine performance. In psychological science, there's a significant debate about whether research findings from laboratory settings can generalize to real-world contexts, a dilemma often described as the "real-world or the lab" dilemma (Holleman et al., 2020). To address this issue, many researchers advocate for experiments with more "ecological validity," a term that implies studies should more closely resemble and generalize to real-world settings. Recognizing that the context in which language processing occurs can significantly influence results, I will also compare the performance of the language model in two distinct settings: a fictional laboratory and a fictional real-world environment. By simulating a more naturalistic setting, I hope to capture a broader range of processing behaviours that might be obscured in a formal laboratory environment. In addition, I will observe whether there is any influence of age or gender on response accuracy.

Huff & Ulakçı (2024) have researched the phenomenon of garden-path sentences in both human participants and LLMs in English. The study involved presenting garden-path sentences with contextually fitting or unfitting preceding sentences to ChatGPT and human participants. They measured relatedness and memorability ratings from ChatGPT and human memory performance in a surprise test. Their results revealed that sentences deemed more related and assessed as being more memorable by ChatGPT were indeed better remembered by humans, even though ChatGPT's internal mechanisms likely differ significantly from human cognition. In the following experiment, I investigate whether there might be a similar effect in Czech – whether simulated participants in the Czech language are also capable of mimicking human responses to a psycholinguistic task.

Chromý (2023) performed an experiment concerning garden-path sentences and the diversity of their (mis)representations. He explored how human readers process complex sentences – garden-path sentences. The author conducted four experiments in Czech using different methods to examine how readers ultimately interpret these sentences. Similarly to previous studies, he discovered that the initial misrepresentation of a garden-path sentence tends to persist even after the correct interpretation becomes clear. The study reveals that the representations of these sentences vary widely among readers. While some readers manage to form a coherent interpretation faithful to the sentence's input, others retain multiple conflicting interpretations simultaneously or fail to construct a coherent representation at all.

Chromý's paper is an essential source for this thesis – the datasets along with the experiment method will be used as the main basis as I researched the language models' abilities in Czech.

The reason it is essential to research LLMs' performance in different languages is due to the disparity in the amounts of training data. Lai et al. (2023) addressed the gap in evaluating ChatGPT, which has predominantly been assessed in English, across a broader spectrum of languages. They investigated ChatGPT's performance in diverse linguistic settings and evaluated its effectiveness in multilingual NLP applications. The authors emphasized the importance of understanding how ChatGPT and similar LLMs perform beyond English-speaking contexts, highlighting challenges and potential biases. ChatGPT generally performs better when provided with English prompts across the majority of tasks and languages. This suggests that English prompts help ChatGPT better understand and analyse tasks, leading to more accurate responses.

1. Overview of Large Language Model Development

1.1 History of Large Language Models

Generative AI, particularly LLMs from 2022–2023, became very popular because they could perform tasks in a way that seemed almost human. However, devices mimicking humans date back to 1966 with ELIZA, which used simple pattern matching and substitution rules to mimic conversation using very little computing power (Kucharavy et al., 2024). After ELIZA, there were rule-based text generation bots, followed by models based on Hidden Markov Models. Hidden Markov Models (HMMs) are probabilistic models used for sequential data, involving a hidden Markov chain with a fixed number of states and associated observation distributions. Each state generates observations according to its distribution, capturing statistical patterns in sequences (Mansouri et al., 2021). These earlier models were a basis for more advanced techniques. Later, Recurrent Neural Network (RNN)-based LLMs, such as ELMo, could analyze and generate text (Kucharavy et al., 2024). RNNs are vital for sequential data tasks like image captioning, speech synthesis, and music generation. Unlike feedforward neural networks, RNNs maintain a state for long-term context, making them suitable for sequence tasks (Lipton et al., 2015).

Another significant advancement was the introduction of Byte-Pair Encoding (BPE), a sub-word tokenization technique widely adopted in natural language processing tasks. BPE improves model efficiency by reducing the number of tokens needed to represent text, which helps in managing vocabulary size and enhances translation quality. This technique became essential for balancing the trade-off between vocabulary size and token sequence length, contributing to better performance in language models (Gallé, 2019).

By 2020, basic LLMs could write journals and blogs with the help of skilled users. Then, the Transformer model came along (Kucharavy et al., 2024). The Transformer model eliminates recurrence entirely and uses only an attention mechanism to capture global dependencies between input and output. This architecture allows for much greater parallelization, enabling significant improvements in computational efficiency and achieving state-of-the-art translation quality with relatively short training times (Vaswani et al., 2017). In 2022–2023, this culminated in LLMs that sparked discussions about whether they should be included in many products and processes. These LLMs stood out because they met people's expectations of how an AI assistant should act, thanks to instructional fine-tuning and Reinforcement Learning from Human Feedback (RLHF) (Kucharavy et al., 2024).

1.2 History of Chat-GPT and the State-of-the-art

ChatGPT, developed by OpenAI, is an intelligent chatbot that responds to user prompts with detailed answers. It is part of the AIGC (AI Generated Content) landscape, which refers to the automatic creation of content such as text, images, and videos using AI technologies. ChatGPT has shown proficiency in various language tasks, including multilingual translation, code debugging, and story writing. ChatGPT is built on the Generative Pre-trained Transformer (GPT) model, which has evolved from GPT-1 to GPT-4o. Key technologies include deep learning, unsupervised learning, instruction fine-tuning, multi-task learning, in-context learning, and reinforcement learning from human feedback (RLHF). The iterative improvements in model architecture and increased data for pre-training have significantly enhanced ChatGPT's performance capabilities (Wu et al., 2023).

Launched in June 2020, GPT-3 featured 175 billion parameters, offering unprecedented natural language understanding and generation. It became widely accessible through an API, enabling diverse applications from simple text completion to complex writing. However, its enormous size demanded high computational power and occasionally produced inconsistent responses.

In November 2022, GPT-3.5-turbo was launched, focusing on efficiency and responsiveness improvements over GPT-3 (OpenAI, 2023).

Introduced in 2023, GPT-4 focused on improving accuracy, efficiency, and contextual understanding. It reduced irrelevant outputs, enhanced adaptive learning for specific tasks, and optimized resource consumption. Though lacking real-time internet access, GPT-4 benefited from a more extensive and diverse training dataset.

Released in May 2024, GPT-4o further refined OpenAI's language models with several groundbreaking features. It delivered more concise and disciplined responses, excelled in structured explanations for scientific and technical contexts, enhanced creative writing, provided comprehensive programming assistance, and offered detailed literary analysis (Kapuściński, 2024).

In July 2024, GPT-4o-mini was released as a small version of the GPT-4o model. The 4o model generally performs better than the 4o-mini model, showing higher mean correctness values across various sentences. Additionally, the 4o model exhibits less variability in its responses, as indicated by shorter error bars, compared to the 4o-mini model, which shows more variability. This suggests that the 4o model is more consistent in its performance across different sentences, whereas the 4o-mini model's performance varies more widely. Overall, the 4o model is more reliable and accurate compared to the 4o-mini model (OpenAI, 2024).

In this thesis, GPT-3.5 turbo, GPT-4o, and GPT-4o-mini are the three representative models used in the experiments.

2. Theoretical Framework and Hypotheses

Huff & Ulakçı's research (2024) on machine prediction of human memorability of garden-path sentences showed that large language models are capable of determining the effect of syntactically ambiguous sentences on people. I expect the experiment in Czech to have a similar effect when it comes to comparing human and machine data in cognitive tasks. Since the language models are going to be prompted in a way to simulate participants in a laboratory experiment, it is assumed in this thesis that they will predict behavior similar to that of human participants undergoing a laboratory experiment in real life. In that sense, I presume that the model will also alter the response accuracy according to potential human responses.

It is also proposed in this thesis that, due to LLMs' capacity for self-regulation, they will be able to replicate substantial disparities in responses and outcomes observed between controlled laboratory settings and more naturalistic home environments. One approach to self-regulation in LLMs involves "Metacognitive Prompting," which enables the models to reflect on their responses and adjust them based on internal evaluations, akin to human metacognitive processes. This method enhances the model's understanding and performance by guiding it through stages of comprehension, judgment formation, evaluation, decision-making, and confidence assessment (Wang, 2024). If this hypothesis is confirmed, it would suggest that LLMs not only mimic human-like response accuracy but also reflect nuanced contextual variations akin to those found in real-world scenarios, enhancing the ecological validity of their performance evaluations.

Another factor that I am observing in the experiments is age and its effect on response accuracy. It has been documented that older individuals exhibit a larger garden-path effect in real-time (online) measures, meaning they experience more difficulty initially processing these sentences correctly (Yoo et al., 2017). I expect the LLM to behave in a similar way – I expect GPT to mirror the difference in performance according to the age of participants.

Lastly, the difference in performance between three language models will be observed as well. The models of choice are GPT-3.5 turbo, GPT-4o, and GPT-4o-mini. In the first section, I hypothesize that the more recent and advanced the model version is, the more accurate the response will be when it comes to resemblance with human responses. In the second part, I assume that the higher the model version, the better it will mimic the way in which humans would respond according to their specific age, gender, and the location in which the experiment took place.

3. Practical Section

The practical part is divided into two main sections. In the first section, I will work with both human and machine-generated data. The main objective is to compare how well the language model mimicked human behaviour when prompted with the same demographic values and in the same environment. To account for variability in technological development, data from GPT-3.5 turbo, GPT-4o, and the recently released GPT-4o-mini are used for comparison with human data. The models were set to two temperatures – temperature 0 and temperature 1. Temperature is a hyperparameter that controls the randomness of the generated text. A higher temperature value increases the diversity of outputs by softening the probabilities of word selection, allowing the model to explore more diverse and sometimes more creative responses. Conversely, a lower temperature value leads to more deterministic outputs where the most probable words are chosen, resulting in more conservative and predictable responses.

The focus is on overall correctness performance and how correctness varied with different sentence-question combinations, as well as comparing specific sentences across all three data sets.

The second section uses machine-generated data only. It focuses on additional factors that might affect the correctness of responses. The factors considered are age, gender, sentence-question combination, and location (whether the experiment took place in a laboratory or at home). The three different language models are also compared, along with the differences in their temperatures set to both 0 and 1.

3.1 Methodology

The experiment was performed using ChatGPT by OpenAI, with specific versions including GPT-3.5-turbo, GPT-4o, and GPT-4o-mini. In each section, GPT was prompted to simulate one of the participant types in a specific experimental environment. The process was automated using a Python script, as the entire experiment consisted of nearly 14,000 responses.

To control the diversity of the text output, temperature was used as a factor. Temperature is a hyperparameter that controls the randomness of the generated text. A higher temperature value increases the diversity of outputs by softening the probabilities of word selection, allowing the model to explore more diverse and sometimes more creative responses. Conversely, a lower temperature value leads to more deterministic outputs, where the most probable words are chosen, resulting in more conservative and predictable responses.

For both sections, I used stimuli developed by Chromý, which are available on the Open Science Framework (https://osf.io/bjas8/). I worked with the stimuli under the file "Experiment 1." The preview of the stimuli is shown in Table 1. It is important to note that it is not always possible to mirror the garden-path effect accurately when translating to English.

Table 1: List of items

Item/condition	**Sentence**	**Question**
1		
A	Ostraha uklidňovala opilce a fanynku na stadionu sledovali střídající hráči [Security was calming down a drunk person, and the substitute players were watching the female fan in the stadium.]	Uklidňovala ostraha fanynku? [Was the security calming down the female fan?]
B	Ostraha uklidňovala opilce a fanynka na stadionu sledovala střídající hráče [Security was calming down a drunk person, and a female fan in the stadium was watching the	Uklidňovala ostraha fanynku? [Was the security calming down the female

	substitute players.]	fan?]
C	Ostraha uklidňovala opilce a fanynku na stadionu sledovali střídající hráči [Security was calming down a drunk person, and the substitute players were watching the female fan in the stadium.]	Sledovali hráči fanynku? [Were the players watching the female fan?]
D	Ostraha uklidňovala opilce a fanynka na stadionu sledovala střídající hráče [Security was calming down a drunk person, and a female fan in the stadium was watching the substitute players.]	Sledovali hráči fanynku? [Were the players watching the female fan?]

(Chromý et.al)

The stimuli consist of 24 items. Each item comprises four conditions, with a combination of either a garden-path sentence or a non-garden-path sentence, paired with a question of either type A or B. The garden-path and non-garden-path sentences have the same number of words and similar lexical content; the only difference is the alternation of nominative and accusative cases, which results in a change in semantic roles. The questions are always of either type A or B. The correct answer is always "False" for conditions A, B, and D, and "True" only for condition C.

I then used Chromý's stimuli to create my own dataset, which is displayed in Table 2.

Table 2: Stimuli created based on Chromý's dataset

Dialogue ID	**Sentence ID**	**Dialogue Type**	**Sentence**	**Question**	**Correct answer**

1GA	1	GA	Ostraha uklidňovala opilce a fanynku na stadionu sledovali střídající hráči. [Security was calming down a drunk person, and the substitute players were watching the female fan in the stadium.]	Uklidňovala ostraha fanynku? [Was security calming down the female fan?]	No
1NA	1	NA	Ostraha uklidňovala opilce a fanynka na stadionu sledovala střídající hráče. [Security was calming down a drunk person, and a female fan in the stadium was watching the substitute players.]	Uklidňovala ostraha fanynku? [Was security calming down the female fan?]	No
1GB	1	GB	Ostraha uklidňovala opilce a fanynku na stadionu sledovali střídající hráči. [Security was calming down a drunk person, and the substitute players were watching the female fan in the stadium.]	Sledovali hráči fanynku? [Did the players watch the female fan?]	Yes

1NB	1	NB	Ostraha uklidňovala opilce a fanynka na stadionu sledovala střídající hráče. [The security was calming down a drunk person, and a female fan in the stadium was watching the substitute players.]	sledovali hráči fanynku [Did the players watch the female fan?]	No

In the stimuli I created based on Chromý's source, the Sentence ID replaces the Item number from Chromý's dataset. The conditions A, B, C, and D are also substituted by combinations GA, GB, NA, and NB, where G stands for garden-path sentence and N stands for non-garden-path sentence. Dialogue ID combines the Sentence ID and the Dialogue Type. The stimuli data thus comprise 96 unique Dialogue IDs.

The stimuli were then used to create prompts that would later be proposed to Chat-GPT to elicit answers. The prompts are divided into two groups: one consisting of simulated participants that have undergone an experiment in fictional laboratory settings and another group of simulated participants that have undergone the experiment in home settings.

3.2 Experiment

3.2.1 Creating a Prompt

The stimuli mentioned in the previous section were then used as a basis for creating prompts that preceded the responses generated by GPT. Each prompt simulated a person of a specific gender and age in a particular setting. The two main settings, or environments, as mentioned earlier, were laboratory and home settings.

In the prompts, the ages of the simulated participants ranged between 11 to 91 years and were distributed proportionately. The genders – male and female – were also distributed equally in the dataset.

3.2.1.1 Laboratory Environment

Laboratory prompt followed a standard laboratory exeprimental procedure, similar to the one used in Chromý's study. In Chromý's scenario, participants were informed that they were taking part in a linguistic experiment. The participants were later seated in front of a screen where they performed the experiment in a self-paced reading test. The design of a laboratory prompt was then lead in the same way. The following shows an example of a laboratory prompt; the specific sentence is a non-garden-path sentence. The Czech prompt is followed by its English translation:

> *'Jste 21letá žena na lingvistickém experimentu v laboratoři. Na obrazovce se ukázal tento stimulus: "Ostraha uklidňovala opilce a fanynku na stadionu sledovali střídající hráči." Následně se objevila otázka: "Uklidňovala ostraha fanynku?" Odpovězte ano nebo ne:'*
>
> *[You are a 21-year-old woman in a linguistic experiment in a laboratory. The following stimulus appeared on the screen: The security was calming down a drunk person, and a female fan in the stadium was watching the substitute players.' Then a question appeared: 'Was the security calming down the fan?' Answer yes or no:]*

In the prompt, the *stimulus* was one of the 48 sentences chosen from the dataset, and the *question* was one of the questions, either of type A or B, attached to the specific sentence in the dataset.

3.2.1.2 Home Environment

For the prompt in the home environment, a different approach was necessary. Due to their rather rare syntactic form and considerably difficult distribution of semantic roles, garden-path sentences are not linguistic phenomena that would likely occur in a real-life conversation. In this regard, the target sentences (both garden-path and non-garden-path) were incorporated as a sentence that appeared in a short story. The entire prompt was designed as a scenario in which a person is reading a story to their friend. The target sentence from the dataset appears later in the story, and in response, the friend asks for clarification as if they misheard the sentence. The following is an example of a prompt simulated in a home environment, followed by its English translation:

> *'Měl jsem tenkrát těsně po narozeninách, bylo mi 51 let, a měl jsem na návštěvě svou kamarádku. Četl jsem jí svou oblíbenou povídku a když jsem došel k větě "kluci honili*

psa a kočku v podkroví znepokojovali šediví hlodavci", kamarádka znejistěla: "tak znepokojovali hlodavci kočku? Ano nebo ne?" Odpověděl jsem:'

["I was just past my birthday at that time, I was 51 years old, and I had my friend visiting. I was reading her my favorite short story, and when I got to the sentence 'The boys chased the dog and the cat was disturbed by gray rodents in the attic,' my friend hesitated: 'So, did the rodents disturb the cat? Yes or no?' I answered:"]

3.2.2 Setting the Temperature and Seed

Once the prompt design was completed, it was also necessary to set the temperature in the script. As explained in an earlier section, a temperature setting of 0 typically means that the model will consistently choose the word with the highest probability. As the temperature increases, the model becomes more inclined to select words with lower probabilities, introducing more variety, unpredictability, and creative outputs. In essence, a higher temperature makes the model's responses less deterministic and more exploratory. In this experiment, the temperature was set to either 0 or 1.

Furthermore, it was also essential to set a seed in the script. Setting a seed in scripts is crucial for several key reasons. Firstly, setting a seed ensures reproducibility, allowing random number generation to produce the same results each time the script is run. This consistency is vital for validating and verifying experimental results. Secondly, it enables comparative analysis by allowing different algorithms or models to be tested under identical conditions. This way, any observed differences in performance can be attributed to the methods themselves rather than random variations. Lastly, setting a seed contributes to consistency in experiments by controlling random processes like data shuffling or parameter initialization. This consistency helps maintain uniform conditions across multiple runs, leading to more reliable and comparable results.

3.2.3 Generating Responses

The Python script was then run using the OpenAI API key. For each model, a total of 6,914 responses were generated – 3,457 for each temperature setting. A new column was created to store the answers provided by GPT. With three models, each run at both temperatures, the dataset contained a total of 20,742 responses.

The responses in the laboratory settings consisted solely of “Yes” or “No” answers. This allowed for automated correctness checking using a Python script. The responses were compared to the correct answers; if the word “Yes” or “No” in the response matched the corresponding word in the "Correct Answer" column, it was marked as correct (1), otherwise, it was marked as incorrect (0).

The home settings, however, presented more challenges in terms of correctness analysis. The answers were not only more extensive but also varied significantly in style across the different models. When GPT-3.5-turbo, at both temperature settings, was asked to provide an answer with the instruction to give a "Yes" or "No" response, it adhered strictly to these instructions. As a result, all the responses included the words "Yes" or "No" at the beginning of the sentence, followed by additional text. In this case, the correctness analysis could be easily managed using the same Python script as for the laboratory responses. The following is an example of such a prompt and response:

> *“Měl jsem tenkrát těsně po narozeninách, bylo mi 11 let, a měla jsem na návštěvě svou kamarádku. Četla jsem jí svou oblíbenou povídku a když jsem došla k větě "klaun rozveselil synovce a sestru na náměstí oslovil ztracený poutník", kamarádka znejistěla: "tak rozveselil klaun sestru? Ano nebo ne?" Odpověděla jsem:”*
>
> *["I had just celebrated my birthday, I was 11 years old, and I had a friend visiting. I was reading her my favorite story, and when I got to the sentence 'the clown cheered up his nephew and addressed the sister in the square, lost pilgrim,' my friend hesitated: 'So did the clown cheer up the sister? Yes or no?' I answered:"]*
>
> *“Ano, je to přesně tak, klaun rozveselil synovce.”*
>
> *["Yes exactly, the clown cheered up the nephew."]*

However, higher versions of GPT did not follow the instructions as literally. In their responses, the simulated participants in the home environment tended to speculate more, providing ambiguous answers that offered different possible interpretations and often did not include a clear "Yes" or "No" answer in any part of the response. Due to the difficulties posed by these issues, a human annotator was required to review all the responses from the home environment. Any response that presented various interpretations or did not provide a specific answer from which the intended meaning could be deduced was assigned as N/A (not applicable) and excluded from the subsequent analysis.

The following is an example of a response to the same prompt specified for GPT-3.5-turbo. This response was deemed non-applicable by the human annotator as it offered various interpretations and no specific answer:

> *"To záleží na tom, jak si tu větu vyložíš," řekl jsem s úsměvem. "V téhle povídce je hodně prostoru pro interpretaci. Může to znamenat, že klaun rozveselil synovce a zároveň oslovil ztraceného poutníka, který se nacházel na náměstí. Nebo to může znamenat, že poutník oslovil sestru klauna. To je na tobě, jak si to představíš."*
>
> *["It depends on how you interpret the sentence," I said with a smile. "There is a lot of room for interpretation in this story. It can mean that the clown cheered up his nephew and at the same time addressed the lost pilgrim who was in the square. Or it can mean that the pilgrim addressed the clown's sister. It's up to you how you imagine it."]*

The next instance, on the other hand, was deemed applicable. It was not assessed accurately by the automated correctness analysis, as the target word was embedded further within the text. However, the human annotator was able to extract the correct meaning:

> "*Podle té věty to vypadá, že kluci honili psa, ale kočku znepokojovali hlodavci. Takže kluci honili psa, ne kočku."*
>
> *["According to the sentence, it seems that the boys were chasing the dog, but the cat was disturbed by the rodents. So the boys were chasing the dog, not the cat."]*

An interesting observation can be made when assessing the responses to Sentence 5:

Inflace potěšila bankéře a živnostníky v Evropě obtěžovali bankovní úředníci.

["Inflation pleased the bankers and the entrepreneurs in Europe were bothered by bank clerks."]

The model GPT-3.5-turbo tended to answer with the majority of responses in a similar format, where the target word always formed the first part of the response:

Ne, inflace potěšila bankéře a obtěžovala živnostníky.

["No, inflation pleased the bankers and bothered the entrepreneurs."]

Higher versions of GPT, such as GPT-4o and GPT-4o-mini, tended to use factual information to answer the question following Sentence 5, rather than only focusing on the content of the sentence provided. The following is a response from GPT-4o, which was the most common format of responses for Sentence 5:

"No, vlastně je to trochu složitější," začala jsem vysvětlovat. "Inflace může mít různé dopady na různé lidi. Pro bankéře může být inflace dobrá, protože rostoucí ceny mohou znamenat více peněz v oběhu a větší zisky z úvěrů. Ale pro živnostníky to není vždycky tak jednoznačné. Vyšší ceny mohou znamenat vyšší náklady na suroviny a provoz, což může být pro ně obtížné. Takže zatímco inflace opravdu potěšila bankéře, pro živnostníky to mohlo být spíš na obtíž," odpověděla jsem s úsměvem, snažíc se to co nejlépe vysvětlit."

["Well, actually it's a bit more complicated," I began to explain. "Inflation can have different impacts on different people. For bankers, inflation can be good because rising prices can mean more money in circulation and greater profits from loans. But for entrepreneurs, it's not always so straightforward. Higher prices can mean higher costs for raw materials and operations, which can be difficult for them. So while inflation really pleased the bankers, for entrepreneurs it could have been more of a burden," I replied with a smile, trying to explain it as best as I could."]

3.2.4 Section One: Human-Participant vs. Machine-Simulated-Participant Performance on Garden-Path Sentence Task

The main objective of this section is to compare the performance of human participants who underwent the garden-path sentence task in real life to the performance of participants simulated through GPT prompts. Following the experimental methods established by Jan Chromý, a similar set of "participants" had to be used for the experimental comparison.

As mentioned earlier, Chromý's experimental group consisted of participants with a mean age of 21.19 years. Therefore, from the dataset of machine responses, only the "participants" with a mean age of 21 years were filtered and consequently used as data for comparison. Furthermore, only responses from simulated participants in a laboratory setting were included. For the comparison, Chromý's dataset of results was used. Data concerning reaction time were excluded, as comparing reaction time in the context of GPT would not be meaningful.

In this section, data from the GPT-3.5 turbo, GPT-4o, and GPT-4o-mini models were used, with the temperature set to 0 and 1 for all the mentioned models.

3.2.4.1 Overall Human-Participant vs Machine-Simulated-Participant Performance on Garden-Path-Sentence Task

The following graph visualizes the results by displaying group means with 95% confidence intervals to compare correctness across different participant groups – either human participants or GPT models – and across the distinct temperature settings. T

Figure 1: Comparison of Correctness by Different GPT Models and Human Participants

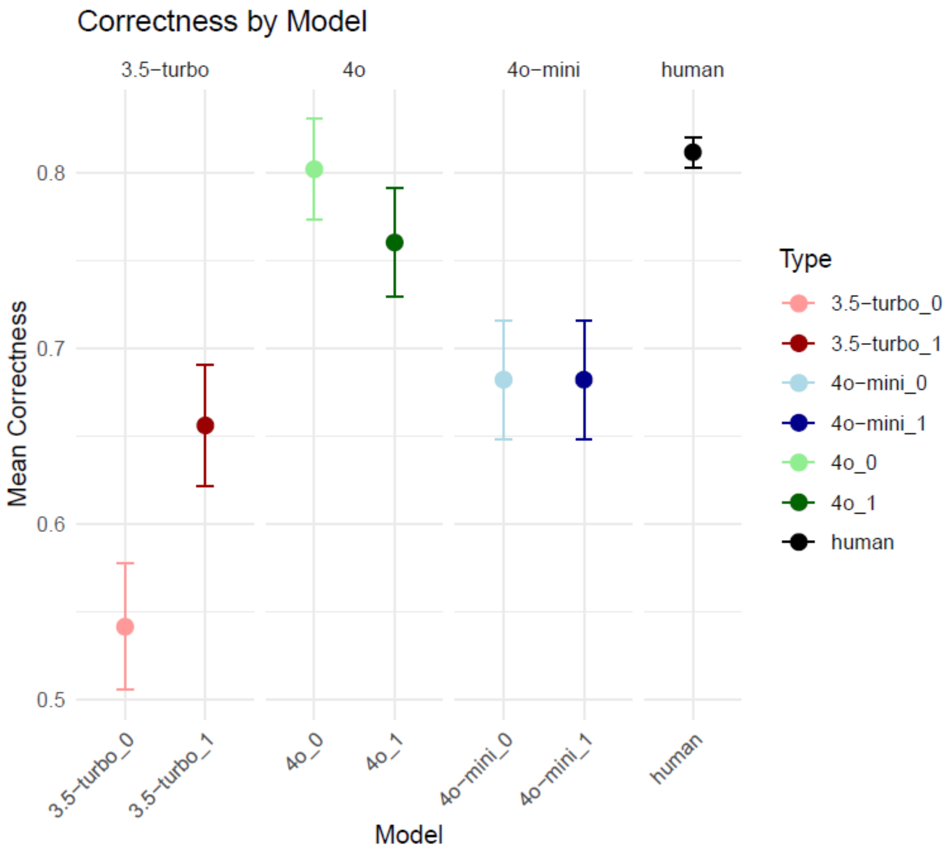

Figure 1 presents a scatter plot with error bars illustrating the correctness of responses from various GPT models under different temperature settings. The y-axis represents the correctness of the responses, ranging from 0 to 1, where 0 indicates incorrect responses and 1 indicates correct responses. The x-axis lists different GPT model types along with their temperature configurations, such as 3.5-turbo_0, 3.5-turbo_1, 4o-mini_0, 4o-mini_1, 4o_0, 4o_1, and human.

Each data point on the plot corresponds to the mean correctness for a specific model type and temperature. The error bars indicate the standard error or standard deviation, reflecting the variability or confidence interval around the mean correctness. The models are color-coded, with each model type assigned a distinct colour, and different shades of the given colour are used to distinguish the two temperature settings.

The correctness values for human responses are generally higher, with most points clustered around 0.7 to 1.0. This suggests that human responses are typically accurate. Additionally, the short error bars imply consistent performance across different sentences.

In contrast, the correctness values for the 3.5-turbo model exhibit a wider range, with mean correctness values spanning approximately 0.3 to 0.8. The larger error bars, compared to those for human responses, indicate greater variability in the model's correctness.

The 4o-mini model shows a broad range of correctness values, from around 0.25 to 1.0, reflecting variable performance across different sentences. Some Sentence_IDs display high correctness values with small error bars, suggesting good performance on specific sentences, while others show more variability.

The correctness values for the 4o model are generally high, with most values exceeding 0.6. This indicates better overall performance compared to the 3.5-turbo model. The error bars for the 4o model are moderate in length, showing a mix of consistency and variability in the model’s performance.

Overall, the GPT-4o model most closely approaches human-level performance in terms of correctness and consistency, as indicated by its generally high correctness values and moderate error bars, aligning more closely with the patterns observed in human responses.

3.2.4.2 Human-Participant vs. Machine-Simulated-Participant Performance on Garden-Path Sentence Task Analysed by Sentence ID

This section provides a detailed analysis of the performance of human participants compared to machine-simulated participants (using various versions of GPT models) on garden-path sentence tasks. The analysis is conducted for each individual sentence ID to assess how well the machine-simulated participants mimic human processing of syntactic ambiguity.

Figure 2: Correctness of Responses by Sentence ID for Various GPT Models and Human Performance

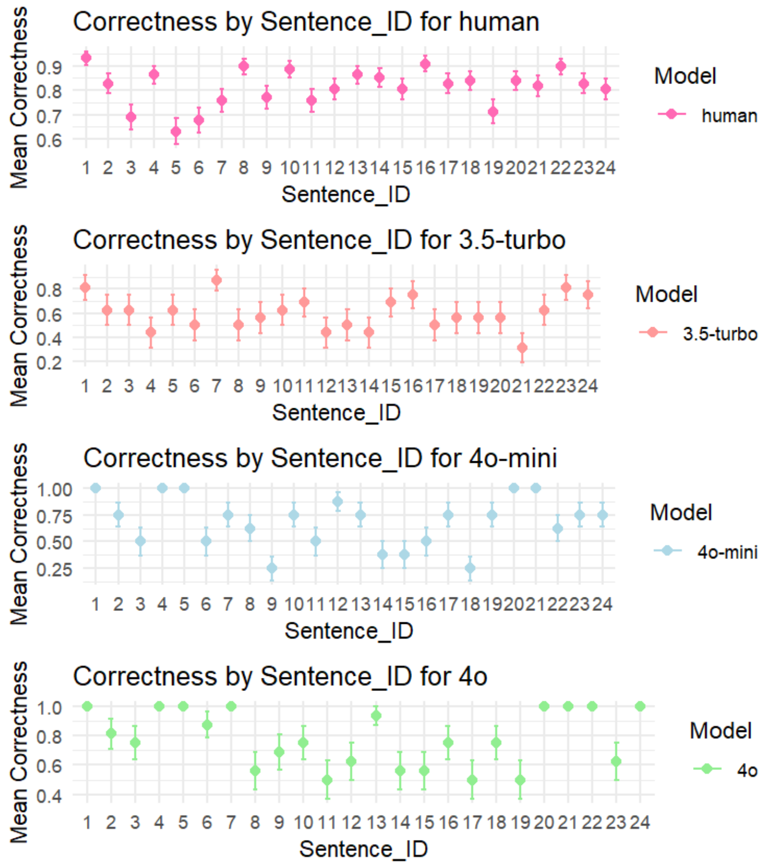

Figure 2 comprises four scatter plots with error bars, illustrating the mean correctness of responses across different Sentence IDs for various GPT models: human, 3.5-turbo, 4o-mini, and 4o.

The y-axis represents the mean correctness of responses, ranging from 0 to 1. Values closer to 1 indicate more accurate responses, while values closer to 0 indicate less accurate responses. The x-axis lists the Sentence IDs, numbered from 1 to 24. The ID does not correspond to a unique sentence but to a set of sentences with a similar lexical root. In other words, a specific Sentence ID corresponds to both garden-path and non-garden-path versions of the same lexical content, with some of the semantic roles and grammatical cases altered. For instance, Sentence ID 2 (or also Sentence 2) corresponds to both the non-garden-path sentence, "*Boys chased a dog and a cat in the attic worried grey rodents,"* and the garden-path sentence, "*Boys chased a dog and grey rodents in the attic worried a cat.*" Each scatter plot corresponds to a different model, allowing for a more detailed comparison of performance across the human, 3.5-turbo, 4o-mini, and 4o models.

The scatter plots enable us to determine which sentences were the most and least accurately responded to by all groups (human, 3.5-turbo, 4o-mini, and 4o).

The most inaccurately responded-to sentence across all groups is identified by the lowest mean correctness and the largest error bars, indicating high variability and low accuracy across all models. Sentence ID 10 was the most challenging, as it showed the lowest mean correctness values and high variability across all groups. Specifically, the human responses had a correctness around 0.7 with moderate error bars, the 3.5-turbo model had correctness around 0.4 with large error bars, the 4o-mini model had correctness around 0.3 with large error bars, and the 4o model had correctness around 0.6 with moderate error bars.

In contrast, the easiest sentence to understand is identified by the highest mean correctness and the shortest error bars, indicating high accuracy and low variability across all models. Sentence ID 7 was the easiest to understand and respond to accurately, with high mean correctness values and short error bars across all groups. Human responses had correctness around 0.95 with short error bars, the 3.5-turbo model had correctness around 0.75 with moderate error bars, the 4o-mini model had correctness around 0.85 with short error bars, and the 4o model had correctness around 0.9 with short error bars.

For reference, the specific Sentence IDs mentioned in the latter analysis are presented in Table 3. GS stands for garden-path sentence, NS stands for non-garden-path sentence. English translations are provided as well, though in some instances, it is difficult to reflect the garden-path effect in the English translation.

Table 3: The Most and Least Difficult Sentences to Comprehend

Sentence ID	**Garden-path and non-garden path sentence within the Sentence ID**	**English translation**
7	**GS**: Ministr odvolal rektora a děkanku na zasedání kritizovali nespokojení studenti. **NS**: Ministr odvolal rektora a děkanka na zasedání kritizovala nespokojené studenty.	**GS**: The minister dismissed the rector, and the dean was criticized by dissatisfied students at the meeting. **NS**: The minister dismissed the rector, and the dean criticized the dissatisfied students at the meeting.

10	**GS**: Knihovna nadchla čtenáře a autorku v předsálí komentovali vybraní kritici. **NS**: Knihovna nadchla čtenáře a autorka v předsálí komentovala vybrané kritiky.	**GS**: The library delighted the readers, and the author was commented on by selected critics in the foyer. **NS**: The library delighted the readers, and the author commented on the selected critics in the foyer.

3.2.4.3 Human-Participant vs. Machine-Simulated Participant Performance on Garden-Path Sentence Task Analysed by Dialogue Type

The primary focus of the following analysis is to understand how different participants performed on various dialogue types.

Figure 3: Comparison of Correctness by Different GPT Models and Human Participants Across Dialogue Types

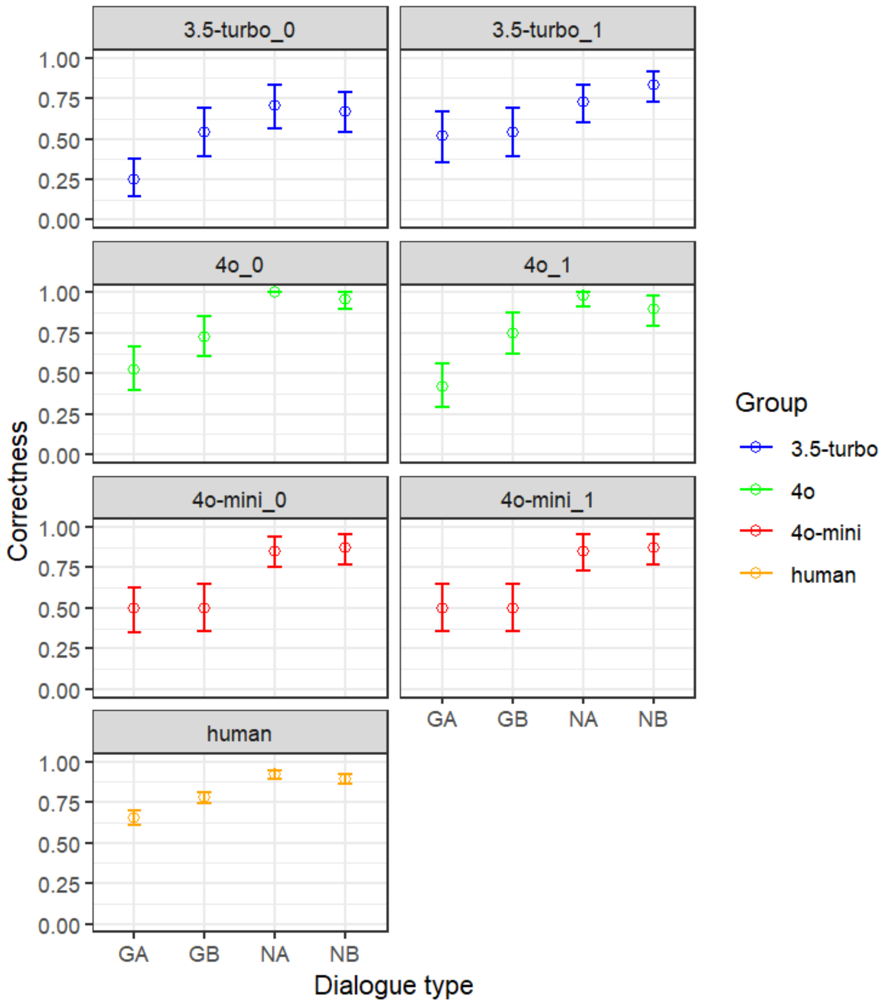

Figure 3 presents scatter plots with error bars illustrating the mean correctness of responses across different dialogue types for various GPT models: 3.5-turbo (with temperatures 0 and 1), 4o (with temperatures 0 and 1), 4o-mini (with temperatures 0 and 1), and human. The y-axis represents the mean correctness of responses, ranging from 0 to 1, where values closer to 1 indicate more accurate responses and values closer to 0 indicate less accurate responses. The x-axis represents the four dialogue types: GA (garden-path sentence followed by a type A question), GB (garden-path sentence followed by a type B question), NA (non-garden-path sentence followed by a type A question), and NB (non-garden-path sentence followed by a type B question).

Each facet within the figure corresponds to a specific model and temperature configuration, with colours indicating the model group: blue for 3.5-turbo, green for 4o, red for 4o-mini, and orange for human. The error bars denote the standard error around the mean correctness, reflecting the variability of the responses.

For GA (garden-path sentence followed by a type A question), human responses exhibit high correctness with minimal variability, indicating consistent and accurate responses. The 3.5-turbo model shows moderate correctness with considerable variability, suggesting inconsistency in handling garden-path sentences with type A questions. The 4o-mini model also demonstrates moderate correctness with some variability, indicating challenges in processing these sentences. In contrast, the 4o model generally shows high correctness with less variability compared to the other models, suggesting better performance in this dialogue type.

For GB (garden-path sentence followed by a type B question), human responses maintain high correctness with minimal variability, reflecting reliable performance. The 3.5-turbo model displays lower correctness with significant variability, indicating difficulties in processing garden-path sentences followed by type B questions. The 4o-mini model, similar to 3.5-turbo, shows lower correctness and higher variability, highlighting challenges with this sentence type. Conversely, the 4o model again shows higher correctness and less variability, suggesting better handling of this dialogue type compared to other models.

For NA (non-garden-path sentence followed by a type A question), human responses are consistently high in correctness with low variability, indicating reliable and accurate responses. The 3.5-turbo model shows improved correctness compared to garden-path sentences but still exhibits variability. The 4o-mini model demonstrates moderate to high correctness with some variability, indicating better performance than with garden-path sentences. The 4o model maintains high correctness with low variability, indicating strong performance in this dialogue type.

For NB (non-garden-path sentence followed by a type B question), human responses show high correctness with minimal variability, reflecting consistent accuracy. The 3.5-turbo model shows variability in correctness, with moderate performance. The 4o-mini model, similar to

3.5-turbo, exhibits variability but generally better correctness compared to garden-path sentences. The 4o model shows high correctness with low variability, indicating strong and reliable performance.

The scatter plots highlight several key findings. Human responses are consistently accurate across all dialogue types, with high correctness and minimal variability. Among the GPT models, the 4o model (both temperatures) mimicked human performance the best. It consistently outperforms the 3.5-turbo and 4o-mini models, particularly in handling garden-path sentences (GA and GB). The 4o model shows higher correctness and less variability, closely aligning with the human benchmark. Specifically, the 4o model's correctness scores are often within the same range as human responses, and the error bars indicate a similar level of consistency. This suggests that the 4o model not only achieved high accuracy but also maintained a level of reliability in its responses comparable to human participants. Both the 3.5-turbo and 4o-mini models struggle more with garden-path sentences, displaying lower correctness and higher variability. The 4o model, however, demonstrates better handling of these challenging sentences, achieving a performance level that is not only higher but also more consistent, much like the human responses. All models show improved performance with non-garden-path sentences (NA and NB), with higher correctness and reduced variability compared to garden-path sentences, but the 4o model's performance is notably closer to the human benchmark across all sentence types, both in accuracy and consistency.

3.2.4.4 Summary of Results of Section One

This section compares the performance of human participants and machine-simulated participants using different versions of GPT models on a garden-path sentence task, aiming to assess how closely these models can mimic human behaviour in processing such sentences.

The analysis evaluates the correctness of responses between human participants and various GPT models (GPT-3.5-turbo, GPT-4o, and GPT-4o-mini) at two temperature settings (0 and 1). Human participants consistently demonstrate high correctness with low variability, indicating reliable and accurate performance. Among the models, the GPT-4o model most closely mirrors human performance, achieving high correctness scores with moderate error bars, reflecting a similar level of consistency to that seen in human responses. In contrast, the GPT-3.5-turbo and GPT-4o-mini models exhibit greater variability and lower correctness, particularly when dealing with more complex sentence structures.

The analysis also examines how each group – human participants and GPT models – performed on specific sentences. Here again, the GPT-4o model demonstrates the closest performance to human participants, especially on the more challenging sentences. It maintains relatively high correctness and lower variability compared to the other models. In contrast, the GPT-3.5-turbo and GPT-4o-mini models struggle more with difficult sentences, showing lower correctness and greater variability.

Additionally, the analysis explores performance by dialogue type, including garden-path and non-garden-path sentences paired with different question types. The GPT-4o model consistently outperforms the GPT-3.5-turbo and GPT-4o-mini models across all dialogue types, particularly in handling garden-path sentences. It achieves higher correctness with less variability, closely aligning with human performance. Meanwhile, the GPT-3.5-turbo and GPT-4o-mini models show more difficulty with these sentences, displaying lower correctness and greater variability.

In summary, the GPT-4o model best mimics human behaviour, achieving high accuracy and consistency across different sentence structures and dialogue types. Its performance is notably closer to that of human participants compared to the GPT-3.5-turbo and GPT-4o-mini models, especially in processing complex garden-path sentences.

3.2.5 Section Two: The Effect of Age, Gender, Location, and Dialogue Type on Correctness Using Models GPT-3.5-turbo, GPT-4o, and GPT-4o-mini

This section examines how different factors, such as location, age, gender, and dialogue type, affect the performance of GPT models in terms of the correctness of their responses. For this analysis, only machine-generated data were used and subsequently compared. Unlike the previous section, data from all ages and both genders of simulated participants were included, as well as data from both types of environments – home and laboratory.

The results in the following subsections were analysed using a Generalized Mixed Model (GMM) in Jamovi to investigate the effect of various factors, including age, on the correctness of responses generated by the GPT-3.5-turbo model. This model employed logistic regression, which is suitable for modelling binary outcomes – in this case, the correctness of responses (correct vs. incorrect). The analysis incorporated both fixed and random effects to account for variability within the data.

The fixed effects included in the model were age, which was the primary focus of this analysis, as well as dialogue type, a categorical variable representing different types of dialogues; location, indicating whether the response was generated in a laboratory setting or at home; and the gender of the participant generating the response. Additionally, random effects were included to capture individual differences, with participant ID being used for random intercepts and slopes for dialogue type, allowing for variability in response patterns across different individuals.

3.2.5.1 The Impact of Age on Correctness across Different Models and Temperatures

The following is an analysis of how different GPT models performed in correctness based on different age groups.

Figure 4: Average Correctness by Age with Confidence Intervals Across Different GPT Models and Temperatures

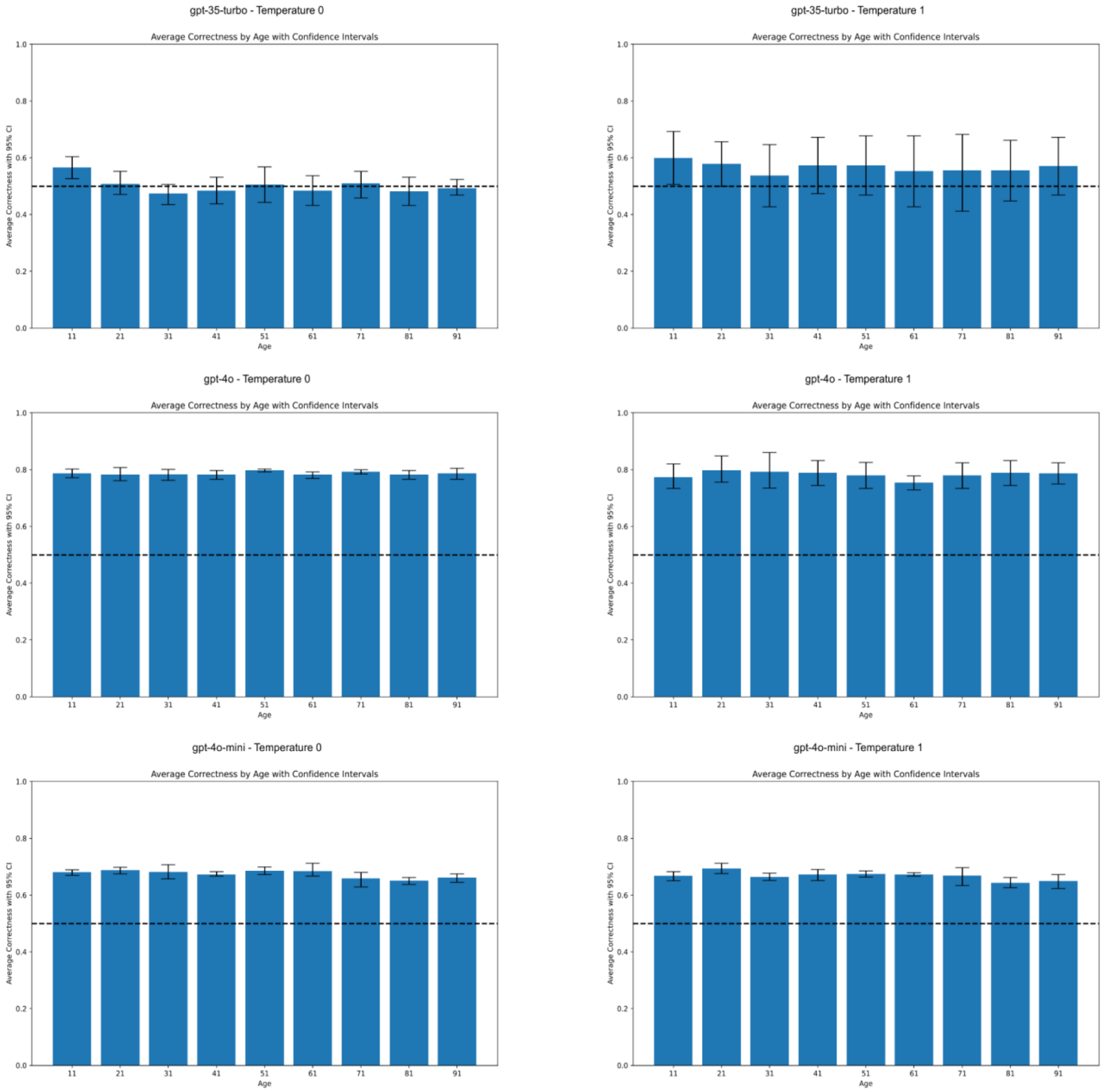

Figure 4 presents a comparative analysis of three language models – GPT-3.5-turbo, GPT-4o, and GPT-4o-mini – evaluated at two different temperature settings (0 and 1). Each pair of bar plots corresponds to one model, with the left plot showing results at temperature 0 and the right plot at temperature 1.

For each model, the average correctness of the model's responses is plotted across different age groups (11, 21, 31, 41, 51, 61, 71, 81, and 91). The vertical axis represents the average correctness score, ranging from 0.0 to 1.0, while the horizontal axis indicates the different age groups. The error bars denote the confidence intervals around the average correctness scores, providing an indication of the statistical reliability of these estimates.

For the GPT-3.5-turbo model of temperature 0, the analysis indicates that the variable "Age" does not have a significant effect on correctness, as evidenced by the omnibus test ($X^2(1) = 2.65e{-4}$, $p = 0.890$). The parameter estimate for "Age" is -2.65e−4 with a p-value of 0.890, suggesting no significant difference in correctness based on age for this model and temperature.

Similarly, the GPT-3.5-turbo model of temperature 1 shows no significant effect of "Age" on correctness ($X^2(1) = 0.098$, $p = 0.754$). The parameter estimate for "Age" is 0.0055 with a p-value of 0.754, indicating that the correctness of responses does not significantly differ across different ages for this model and temperature.

For the GPT-4o model of temperature 0, the omnibus test results reveal that the variable "Age" does not significantly impact correctness ($X^2(1) = 0.0192$, $p = 0.890$). The parameter estimate for "Age" is -2.65e−4 with a p-value of 0.890, indicating no significant difference in correctness based on age.

The GPT-4o model of temperature 1 also shows that the effect of "Age" is not significant ($X^2(1) = 0.098$, $p = 0.754$), with a parameter estimate for "Age" of 0.0055 and a p-value of 0.754, suggesting no significant difference in correctness across different ages.

For the GPT-4o-mini model of temperature 0, the omnibus test results indicate that "Age" does not have a significant impact on correctness ($X^2(1) = 1.957$, $p = 0.162$). The parameter estimate for "Age" is 0.0023 with a p-value of 0.162, supporting the finding that age does not significantly affect correctness in this model.

The GPT-4o-mini model of temperature 1 shows similar results, with no significant effect of "Age" on correctness ($X^2(1) = 1.357$, $p = 0.244$). The parameter estimate for "Age" is 0.0019 with a p-value of 0.244, indicating no significant difference in correctness based on age.

These findings consistently suggest that age does not have a significant impact on the correctness of responses across different models and temperatures.

3.2.5.2 The Impact of Gender on Correctness across Different Models and Temperatures

The following is an analysis of how different GPT models performed in correctness based on gender. In this case, F stands for female; M stands for male.

Figure 5: Average Correctness by Gender with Confidence Intervals Across Different Models and Temperature Settings

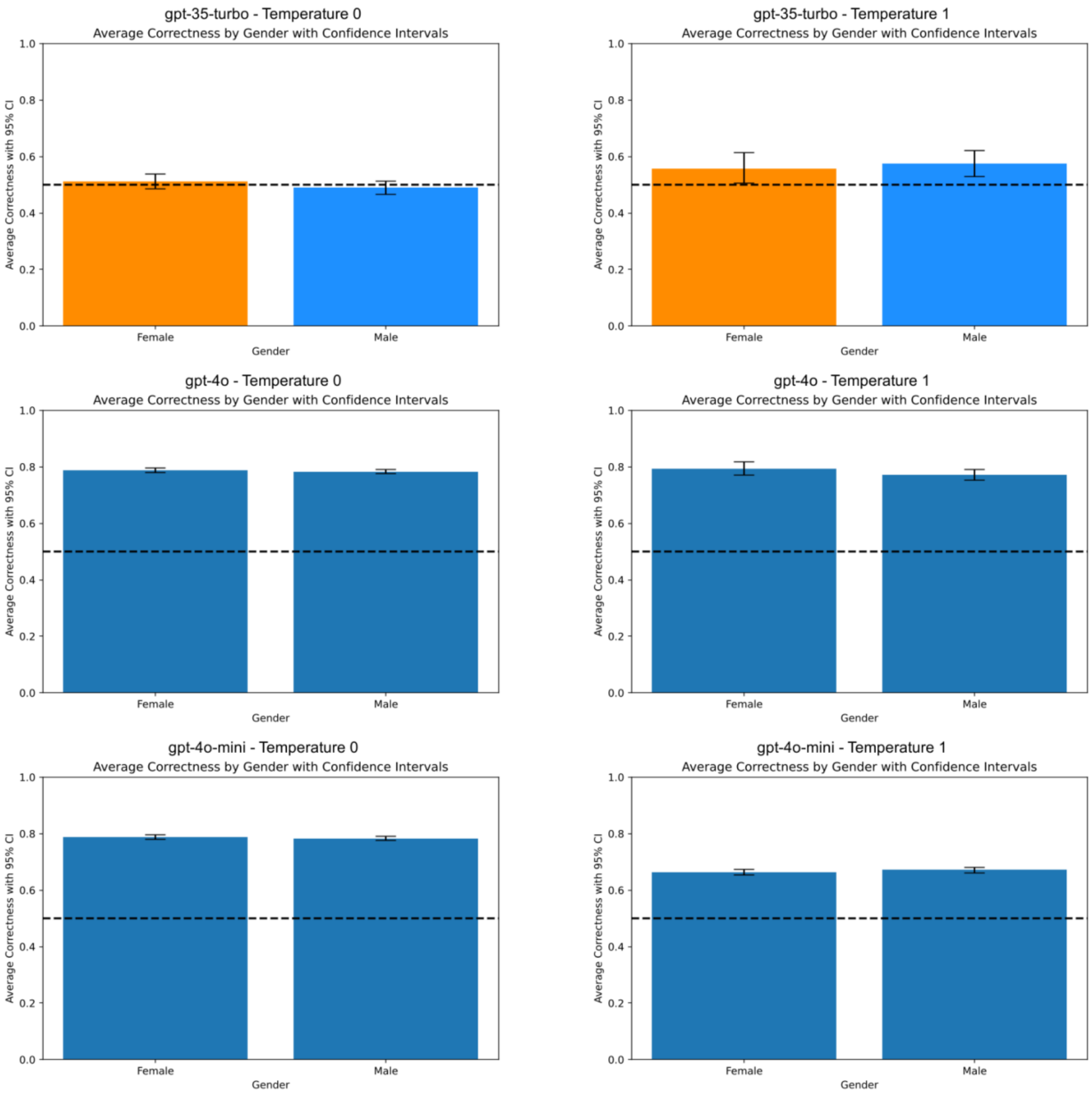

Figure 5 presents the average correctness of responses across different gender groups (Female and Male) for three language models – gpt-35-turbo, gpt-4o, and gpt-4o-mini – under two different temperature settings (0 and 1). Each pair of bar plots corresponds to one model, with the left plot showing results at temperature 0 and the right plot at temperature 1.

As in the previous section, this analysis utilizes logistic regression models to examine the correctness of responses, with fixed effects including Dialogue type, Location, Age, and Gender, and random effects for Dialogue type by Participant ID. The findings for the variable "Gender" across different models and temperatures are as follows:

For the 3.5-turbo model of temperature 0, the omnibus test indicates that "Gender" does not have a significant effect on correctness ($X^2(1) = 0.005$, $p = 0.995$). The parameter estimate for "Gender" (M vs. F) is -1.7698 with a p-value of 0.995, suggesting no significant difference in correctness based on gender for this model and temperature.

Similarly, in the 3.5-turbo model of temperature 1, "Gender" does not significantly affect correctness ($X^2(1) = 1.702$, $p = 0.192$). The parameter estimate for "Gender" (M vs. F) is 0.5724 with a p-value of 0.192, indicating that correctness does not significantly differ between male and female participants for this model and temperature.

For the 4o model of temperature 0, the results show no significant impact of "Gender" on correctness ($X^2(1) < 0.0001$, $p = 0.995$). The parameter estimate for "Gender" (M vs. F) is -1.7698 with a p-value of 0.995, indicating no significant difference in correctness between male and female participants.

The 4o model of temperature 1 also shows that "Gender" does not significantly affect correctness ($X^2(1) = 1.702$, $p = 0.192$), with a parameter estimate for "Gender" (M vs. F) of 0.5724 and a p-value of 0.192, again suggesting no significant difference in correctness based on gender.

In the 4o-mini model of temperature 0, the omnibus test for "Gender" indicates that it does not significantly affect correctness ($X^2(1) = 1.114$, $p = 0.291$). The parameter estimate for gender (M vs. F) is -0.1051 with a p-value of 0.291, showing a slight, but not statistically significant, trend where males might have lower correctness scores compared to females.

Similarly, the 4o-mini model of temperature 1 results indicate that "Gender" does not significantly influence correctness ($X^2(1) = 0.629$, $p = 0.428$). The parameter estimate for gender (M vs. F) is -0.0742 with a p-value of 0.428, suggesting no meaningful difference in correctness based on gender. Overall, across all models and temperature settings, "Gender" does not significantly impact the correctness of responses.

3.2.5.3 The Impact of Location on Correctness across Different Models and Temperatures

The following is an analysis of how different GPT models performed in correctness based on location – whether the “experiment” took place in laboratory or more natural home settings.

Figure 6: Average Correctness by Location with Confidence Intervals Across Different GPT Models and Temperatures:

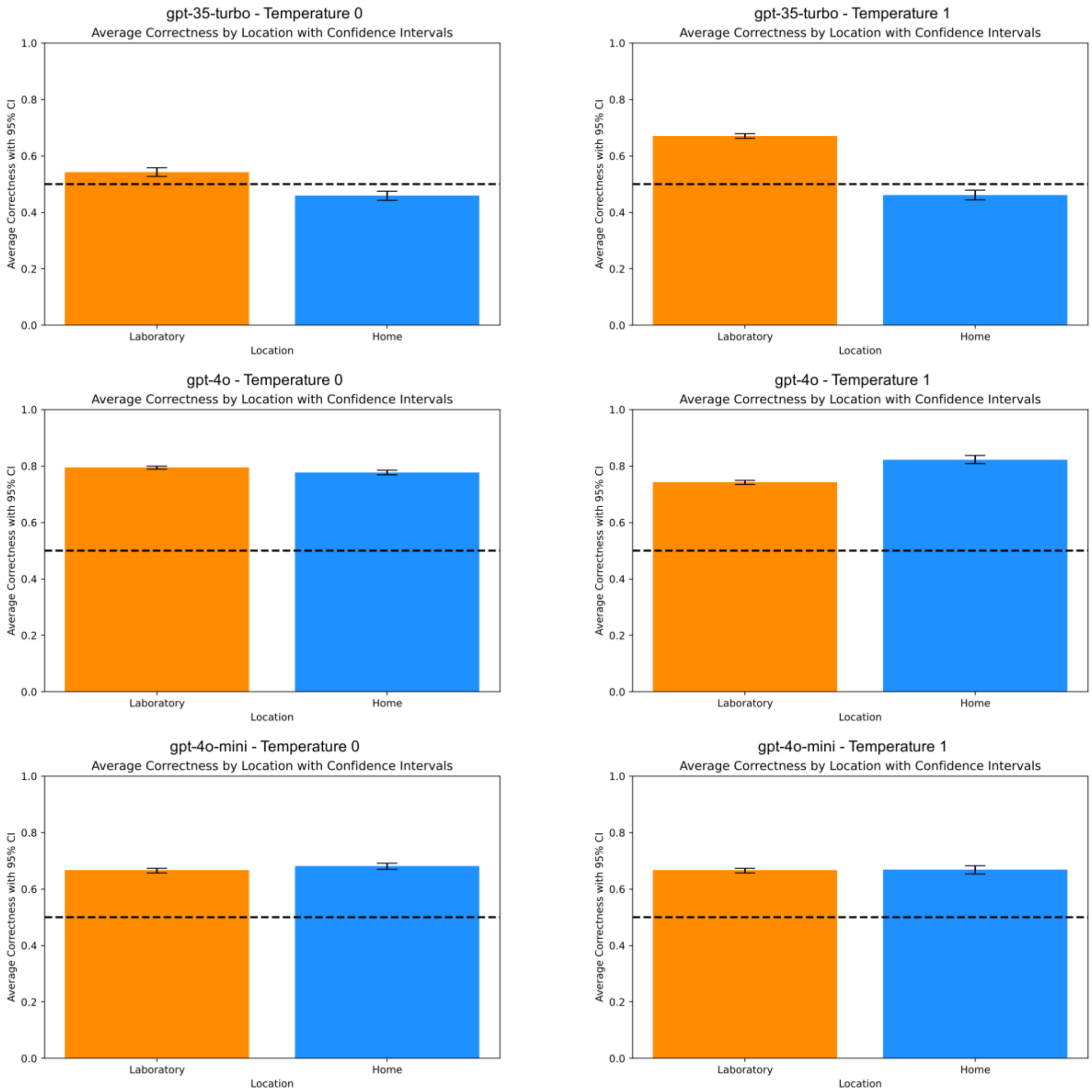

Figure 6 presents the average correctness of responses across different environmental settings (Laboratory vs. Home) for three language models – gpt-35-turbo, gpt-4o, and gpt-4o-mini – at two temperature settings (Temperature 0 and Temperature 1). Correctness is measured on a scale from 0 to 1, where 1 indicates a correct response and 0 indicates an incorrect response. Each subplot corresponds to a specific model and temperature configuration, with the left column showing results at Temperature 0 and the right column at Temperature 1.

The 3.5-turbo model of temperature 0 was analysed using logistic regression, showing a near-significant effect of "Location" on correctness ($X^2(1) = 3.46$, $p = 0.063$). The parameter estimate (0.23451, $p = 0.063$) suggests a trend toward higher correctness in the laboratory compared to at home, though this finding is not statistically significant.

In contrast, the 3.5-turbo model of temperature 1 reveals a highly significant effect of "Location" on correctness ($X^2(1) = 163.788$, $p < 0.001$), with the parameter estimate (-0.97851, $p < 0.001$) indicating significantly lower correctness in the laboratory compared to at home. This suggests that the laboratory environment may negatively impact performance for this model and temperature.

For the 4o model of temperature 0, there is no significant effect of "Location" on correctness ($X^2(1) < 0.0001$, $p = 0.995$), with the parameter estimate (-1.7300, $p = 0.995$) indicating no difference in correctness based on location. Similarly, the 4o model of temperature 1 shows no significant effect of "Location" on correctness ($X^2(1) = 0.292$, $p = 0.589$), suggesting consistent performance regardless of environment.

In the 4o-mini model of temperature 0, location also does not have a significant effect on correctness ($X^2(1) = 3.346$, $p = 0.067$), with the parameter estimate (0.1822, $p = 0.067$) suggesting a non-significant trend towards higher correctness in the laboratory.

The 4o-mini model of temperature 1 analysis similarly indicates that location does not significantly affect correctness ($X^2(1) = 0.511$, $p = 0.475$), with a small, positive parameter estimate (0.0669, $p = 0.475$) that is not statistically significant.

Overall, the variability in correctness based on "Location" across different models and temperatures suggests that environmental factors may impact model performance differently. The significant negative effect for the 3.5-turbo model of temperature 1 in the laboratory could indicate that certain conditions in this setting hinder performance, while the lack of significant effects for the 4o models suggests they are more robust to environmental changes.

3.2.5.4 The impact of dialogue type on correctness across different models and temperatures

The following is an analysis of how different GPT models performed in correctness based on dialogue type. Analogically to section one, here, dialogue type is a combination of sentence and question type. GA stands for garden-path sentence followed by a type A question; GB is a garden-path sentence followed by a type B question; NA is a non-garden-path sentence followed by a type A question and NB is a non-garden-path sentence followed by a type B question.

Figure 7: Average Correctness by Dialogue Type with Confidence Intervals Across Different Models and Temperature Settings

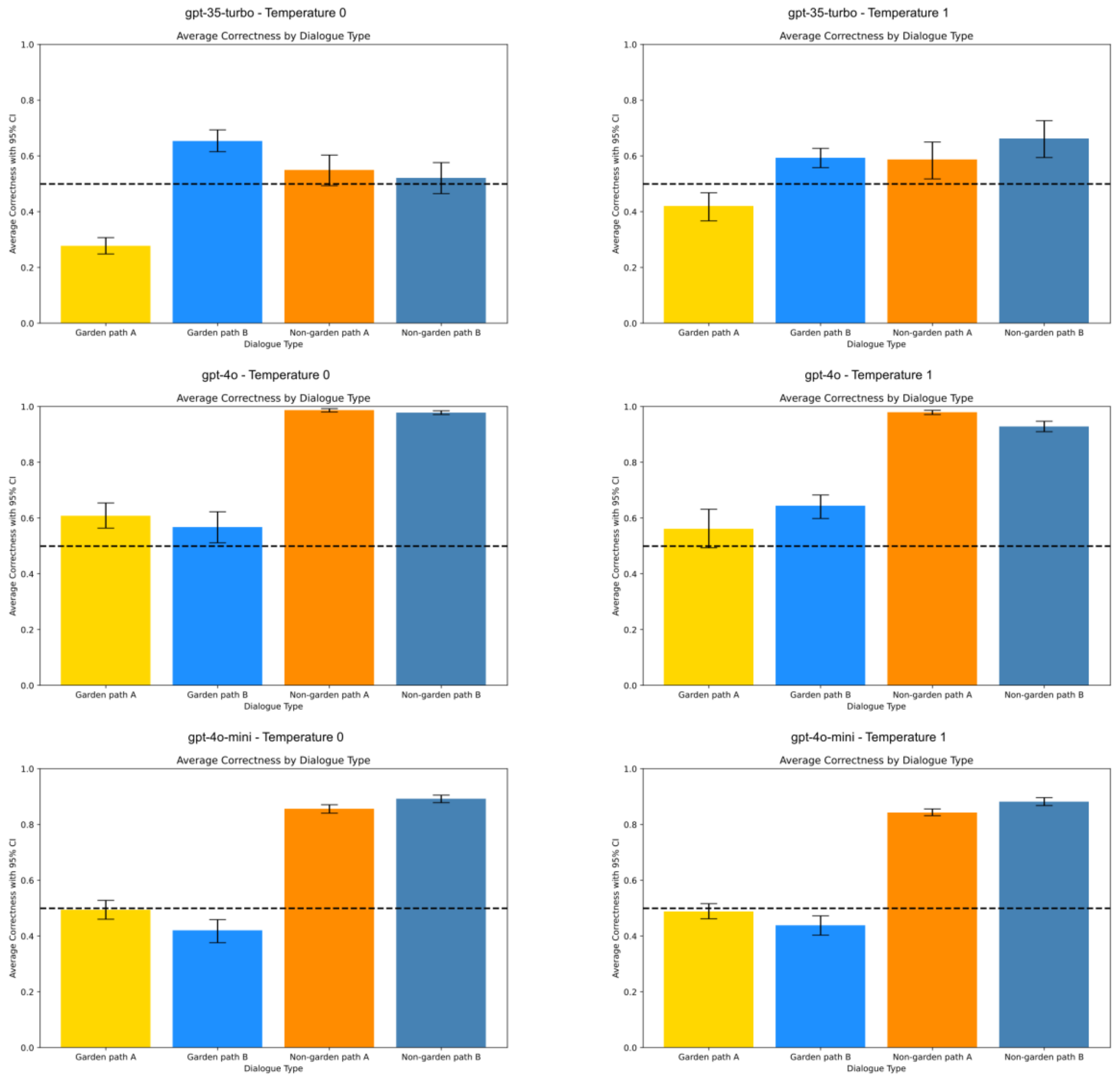

Figure 7 presents the average correctness of responses across different dialogue types for three language models – gpt-35-turbo, gpt-4o, and gpt-4o-mini – at two temperature settings (Temperature 0 and Temperature 1). Correctness is measured on a scale from 0 to 1, where 1 indicates a correct response and 0 indicates an incorrect response. Each subplot corresponds to a specific model and temperature configuration, with the left column showing results at Temperature 0 and the right column at Temperature 1.

For the 3.5-turbo model of temperature 0, the analysis indicates that "Dialogue type" does not have a significant effect on correctness ($X^2(3) = 4.36$, $p = 0.225$). The parameter estimates for the different dialogue types (NA, GB, NB compared to GA) suggest no substantial difference in correctness based on dialogue type for this model and temperature. In contrast, the 3.5-turbo model of temperature 1 shows a significant effect of "Dialogue type" on correctness ($X^2(3) = 10.785$, $p = 0.013$), indicating that correctness varies based on dialogue type for this model and temperature.

The 4o model of temperature 0 reveals that "Dialogue type" has a significant impact on correctness ($X^2(3) = 16.436$, $p = 0.001$). The parameter estimates indicate significant differences in correctness based on dialogue type for this model and temperature. Similarly, the 4o model of temperature 1 shows that "Dialogue type" significantly affects correctness ($X^2(3) = 15.694$, $p = 0.001$), with the parameter estimates suggesting that correctness varies based on dialogue type for this model and temperature.

For the 4o-mini model of temperature 0, the logistic regression analysis indicates a highly significant effect of "Dialogue type" on correctness ($X^2(3) = 492.871$, $p < 0.001$). The parameter estimates reveal that non-garden-path sentences, whether followed by a type A or type B question, are associated with significantly lower correctness scores compared to garden-path sentences followed by a type A question. Specifically, the NA dialogue type shows a substantial decrease in correctness (Estimate = -1.844, $p < 0.001$). The NB dialogue type also leads to a marked reduction in correctness (Estimate = -2.213, $p < 0.001$). Conversely, the GB dialogue type shows an increase in correctness compared to GA (Estimate = 0.331, $p = 0.002$). These findings indicate that dialogue structure plays a critical role in the model's performance, with non-garden-path sentences generally leading to poorer outcomes in correctness.

The 4o-mini model of temperature 1 analysis similarly shows a highly significant effect of "Dialogue type" on correctness ($X^2(3) = 469.817$, $p < 0.001$), consistent with the findings from the 4o-mini model of temperature 0.

3.2.5.5 Summary of results of Section 2

The analysis across different GPT models – gpt-35-turbo, gpt-4o, and gpt-4o-mini – examined the impact of various factors, including age, gender, location, and dialogue type, on the correctness of responses at two different temperature settings (0 and 1).

Firstly, the analysis reveals that the variable "Age" does not have a significant effect on the correctness of responses across all models and temperature settings. For both temperature settings of the gpt-35-turbo model, as well as for the gpt-4o and gpt-4o-mini models, the omnibus tests and parameter estimates indicate non-significant effects. This suggests that age does not meaningfully influence the performance of these models in terms of correctness.

Similarly, gender does not significantly impact the correctness of responses across all models and temperature settings. The gpt-35-turbo model shows no significant effect of gender on correctness at either temperature setting. Likewise, the gpt-4o and gpt-4o-mini models do not exhibit significant differences in correctness based on gender, as indicated by the omnibus tests and parameter estimates. This consistency suggests that the performance of these models is not influenced by the gender of the participants.

The effect of location (laboratory vs. home environment) on correctness, however, varies across models. For the gpt-35-turbo_0 model, there is a near-significant trend suggesting higher correctness in the laboratory compared to home, though this finding is not statistically significant. In contrast, the gpt-35-turbo model of temperature 1 reveals a highly significant effect of location, with significantly lower correctness in the laboratory compared to home, indicating that the laboratory environment may negatively impact performance for this model and temperature. On the other hand, the gpt-4o and gpt-4o-mini models do not show significant effects of location on correctness, suggesting consistent performance across different environments for these models.

Finally, the effect of dialogue type on correctness varies across models. For the gpt-35-turbo_0 model, dialogue type does not significantly affect correctness. However, the gpt-35-turbo_1 model does show significant variability in correctness based on dialogue type, indicating that this factor influences the model's performance at temperature 1. Both the gpt-4o and gpt-4o-mini models demonstrate significant effects of dialogue type on correctness, with specific dialogue types leading to higher or lower correctness scores. Non-garden-path sentences (NA, NB) tend to result in lower correctness compared to garden-path sentences (GA), highlighting the importance of sentence structure in model performance.

Conclusion

In this thesis, I explored the cognitive abilities and multilingual performance of large language models (LLMs), focusing on how well these models can mimic human behaviour when processing garden-path sentences – sentences that are syntactically ambiguous and often lead to initial misinterpretation.

The main goal was to determine if more advanced versions of GPT models could accurately replicate human responses, particularly when faced with complex linguistic tasks. The study compared the performance of human participants with machine-simulated participants using various GPT models, including gpt-3.5-turbo, gpt-4o, and gpt-4o-mini, under different temperature settings. The analysis considered various demographic factors and environmental contexts to understand how these models perform across different conditions.

The results generally supported the initial hypothesis. The gpt-4o model consistently came closest to mimicking human performance, particularly in terms of correctness and consistency across different sentence types and dialogue scenarios. This model not only achieved high accuracy but also demonstrated a reliability in its responses that closely mirrored human behaviour.

Among the models tested, gpt-4o stood out for its ability to replicate human-like responses. It performed well across various sentence structures, especially in handling garden-path sentences, which are known to be challenging for human cognition. While the gpt-4o-mini model also showed some ability to mimic human responses, it was less consistent and accurate compared to the gpt-4o. The gpt-3.5-turbo model, although capable, exhibited more variability and struggled more with complex sentence structures.

The study also examined whether these models could detect differences in performance based on the experimental setting – whether the task was performed in a simulated laboratory or home environment. Interestingly, the gpt-35-turbo model of temperature 1 showed a significant difference, performing worse in the laboratory setting compared to the home environment. This suggests that the model might be sensitive to contextual changes, somewhat mimicking how environmental factors can influence human cognitive performance. However, this effect was not observed in the gpt-4o and gpt-4o-mini models, which remained consistent regardless of the environment, indicating a robustness that might not be as strong in human participants.

When it came to age, the models did not show any significant differences in performance. The expectation that older simulated participants would have more difficulty processing garden-path sentences, similar to what is seen in human cognitive aging, was not supported by the data. This suggests that the models were not able to effectively simulate age-related cognitive changes, which is an area where they still fall short of mimicking human behaviour.

Overall, this thesis highlights both the progress and limitations of current large language models. The advancement from gpt-3.5-turbo to gpt-4o shows significant improvements in the models' ability to replicate human-like cognitive processing, especially in challenging linguistic scenarios. The gpt-4o model, in particular, demonstrates how far these models have come in achieving human-like accuracy and consistency.

However, the study also points out areas where these models still need improvement, particularly in their ability to simulate more nuanced aspects of human cognition, such as age-related differences in processing. These limitations suggest that while LLMs are becoming increasingly sophisticated, there is still room for further development, especially in enhancing their contextual understanding and adaptability to a wider range of human cognitive behaviours.

In conclusion, the findings of this thesis show that while GPT models, especially gpt-4o, are improving in mimicking human behavior, challenges remain in fully replicating the complexity of human cognition.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165v4. https://doi.org/10.48550/arXiv.2005.14165

Chromý, J. (2022). The type of garden-path matters: When readers fail to form a coherent representation of garden-path sentences. Quarterly Journal of Experimental Psychology, 75(1), 149-161. https://doi.org/10.1177/17470218211037152

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. Frontiers in Psychology, 11, Article 721. https://doi.org/10.3389/fpsyg.2020.00721

Huff, M., & Ulakçı, E. (2024). Towards a Psychology of Machines: Large Language Models Predict Human Memory. arXiv. https://doi.org/10.48550/arXiv.2403.05152

Kapuściński, M. (2024, June 4). Evolution of AI: From GPT-1 to GPT-4o – Key Features, Milestones, and Applications. TTMS. Retrieved from https://ttms.com/chat-gpt-evolution/#5.1-what-makes-gpt-4o-stand-out

Kucharavy, A., Plancherel, O., Mulder, V., Mermoud, A., & Lenders, V. (Eds.). (2024). Large Language Models in Cybersecurity: Threats, Exposure and Mitigation (pp. 3-4). Springer Nature. Retrieved from https://doi.org/10.1007/978-3-031-54827-7

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. arXiv. https://doi.org/10.48550/arXiv.2304.05613

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv. Available online at https://ar5iv.labs.arxiv.org/html/1506.00019

Mansouri, T., Sadeghimoghadam, M., & Sahebi, I. G. (n.d.) (2021). A new algorithm for Hidden Markov Models learning problem. arXiv. Retrieved from https://arxiv.org/pdf/2102.07112

Matthias Gallé. (2019). Investigating the Effectiveness of BPE: The Power of Shorter Sequences. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil, O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. PLOS ONE. https://doi.org/10.1371/journal.pone.0298522

OpenAI. (2023, August 22). GPT-3.5 Turbo fine-tuning and API updates. OpenAI. Retrieved from https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

OpenAI. (2024, May 13). Hello GPT-4o. OpenAI. Retrieved from https://openai.com/index/hello-gpt-4o/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. Google Brain and Google Research. Retrieved from https://arxiv.org/pdf/1706.03762v7

Wang, Y., & Zhao, Y. (2024, March 20). Metacognitive Prompting Improves Understanding in Large Language Models. arXiv. https://arxiv.org/html/2308.05342v4

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. Chinese Science Bulletin. https://dds.sciengine.com/cfs/files/pdfs/view/2329-9266/6390DD2D26A24E36A8BE5750EAFBDD47.pdf

Yoo, H., & Dickey, M. W. (2017). Aging Effects and Working Memory in Garden-Path Recovery. Clinical Archives of Communication Disorders, 2(2), 91-102. https://doi.org/10.21849/cacd.2017.00122. Available online at: https://www.researchgate.net/publication/319907471_Aging_Effects_and_Working_Memory_in_Garden-Path_Recovery