SAARLAND UNIVERSITY

Faculty of Humanities
Department of Language Science and Technology
Master Thesis

Large Language Models and their failure to draw Atypicality Inferences

Charlotte Kurch
Saarbrücken
August 2024

Supervisor:
Prof. Dr. Vera Demberg

Advisor:
Margarita Ryzhova

Declarations

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged. Partial results of this thesis are also to be published as: Large Language models fail to derive atypicality inferences in a human-like manner. Charlotte Kurch, Margarita Ryzhova, and Vera Demberg. To appear in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

The tool ChatGPT by OpenAI was employed only for the generation of LaTeX code, specifically the generation of tables. The code for the violin plots that can be found in this work was provided by Margarita Ryzhova.

I assure that the electronic version is identical in content to the printed version of the Master's thesis.

Charlotte Kurch

Saarbrücken, 06.08.2024

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Vera Demberg, for providing me with the opportunity to undertake this research and for guiding me through the process. Your suggestions and feedback have been invaluable in shaping this thesis. The chance to collaborate on turning this work into a paper has provided me with insights that have significantly informed and improved my research.

My sincerest thanks go to my advisor, Margarita Ryzhova, who offered guidance and advice on everything from high-level ideas to the small details of data annotations and statistical analysis. Knowing that I could come to you with any question, no matter how small, and receiving your assistance whenever I struggled with R or LaTeX, was truly invaluable.

I would also like to thank Mayank Jobanputra for his help with the Llama model. Additionally, I am beyond thankful to Jonas Elflein, not only for being a fresh pair of eyes on my code whenever I got stuck, but also for constantly encouraging me.

Finally, I would like to thank all my friends and family for their support and encouragement, and especially my fellow students Tahiira Zimmerman and Lynn Zhou – we've got this!

Thank you all for your support and contributions to this thesis.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG), Funder Id: http://dx.doi.org/10.13039/501100001659, Project-ID 232722074 – SFB1102: Information Density and Linguistic Encoding.

Abstract

Recent studies have claimed that large language models (LLMs) are capable of drawing pragmatic inferences (Qiu et al., 2023, Hu et al., 2022, Barattieri di San Pietro et al., 2023) and that reasoning abilities are emerging in LLMS (Huang et al., 2022a, Han et al., 2024). This master thesis sets out to test LLMs' abilities on atypicality inferences, a type of pragmatic inference that is triggered through informational redundancy that is identified based on script knowledge. Several state-of-the-art LLMs ares tested in a Zero-Shot setting and found to fail systematically at deriving atypicality inferences. A robustness analysis indicates that when inferences are seemingly derived in a Few-Shot setting, these results can be attributed to shallow pattern matching and not pragmatic inferencing. An analysis of the performance of the LLMs at the different derivation steps required for drawing atypicality inferences indicates that the failure seems to stem from not reacting to the subtle maxim of quantity violations introduced by the informationally redundant utterances. Consequently, this thesis then tests two additional prompting methods (adaptations of Generated Knowledge Prompting (Liu et al., 2022) and Three-hop Reasoning (Fei et al., 2023) that, respectively, aim at making the redundancy more explicit, and guiding the models through the necessary reasoning / derivation process. Even with these additional prompting methods, none of the LLMs derive atypicality inferences consistently, showing that this derivation is currently outside the scope of their pragmatic and reasoning abilities.

Contents

Chapter 1
Introduction

Large Language Models (LLMs) have shown impressive performance improvements in recent years, not only excelling on a variety of natural language understanding and processing tasks, but also garnering public attention and fascination. Especially 'chat-bot' AI's, i.e. generative LLMs such as OpenAI's ChatGPT, have firmly and quickly taken a place in society, proving themselves useful assistants to all kinds of people with all kinds of tasks and problems. Notably, these LLMs can produce language (but also to some extent multi-modal) output, that appears to some degree indistinguishable from human language production. While some argue that these models are not more than advanced stochastic parrots (c.f. Bender et al. (2021)) that reproduce only what they have previously seen – which in itself has brought up questions regarding copyright issues [1] – the models continue to impress, and more intelligent or human-like behavior such as reasoning appears to emerge.

Consequently, there have been investigations into how much actual language understanding these models are capable of, how much of the linguistic information innate to humans is learned by them (Malfa et al., 2023, Tenney et al., 2019), and to what extent their language understanding parallels human understanding on a pragmatic level (Hu et al., 2022, Pandia et al., 2021).

This master thesis follows this line of research by investigating the abilities of four LLMs (GPT-3.5-turbo, GPT-4, Llama 3 and Mixtral) with regard to drawing atypicality inferences, a pragmatic phenomenon that has been observed and studied in humans (Kravtchenko and Demberg, 2022, Ryzhova et al., 2023). As the research has shown, there is strong evidence that people notice informational redundancies that are rooted in script knowledge, and are able to accommodate these redundancies through changing their belief about the common ground or typical behavior. This drawing of an atypicality inference is not only reflected in typicality ratings of a habitual activity, but also by the explanations people offer for their typicality rating. However, not everybody draws such inferences and even people who draw them might reject them. Interestingly, the drawing and accepting of atypicality inferences seems to correlated with reasoning abilities,

[1] As of December 2023, the New York Times has sued OpenAI for copyright infringement, showing a version of ChatGPT-having reproduced their articles word for word https://apnews.com/article/openai-new-york-times-chatgpt-lawsuit-grisham-nyt-69f78c404ace42c0070fdfb9dd4caeb7, accessed: 14.01.2024

implying that the (implicit) steps taken to actually accommodate the redundancy in such a way are akin to other thought or reasoning processes (Ryzhova et al., 2023).

This thesis presents the results of six different experiments, that are laid out in more detail below in the overview over the chapters (see Section 1.1). These experiments not only showed that the tested LLMs fail to derive atypicality inferences when tested analogously to humans, but also when different prompting methods are employed to facilitate the necessary reasoning. Further, this thesis provides evidence that the models' failure can be related to their inability to identify informational redundancies as conversationally infelicitous, and the tendency to accommodate redundancy without considering atypicality. Additionally, some insights into the robustness of the LLMs behavior and the faithfulness of generated output are provided. Ultimately, this master thesis provides solid evidence for the inability to derive atypically inferences exhibited by the tested models and valuable insights into the LLMs' behaviors that potentially cause this[2].

1.1 Overview over the chapters

Chapter 2 introduces the relevant literature in regard to atypicality inferences and how they relate to script knowledge and reasoning. Further, research into investigating the abilities of LLMs, specifically with regard to human-like behavior such as reasoning, is discussed. Chapter 3 then provides more detailed insights into the previous studies on atypicality inferences in humans, the employed experimental materials, and how these experiments are adapted for LLMs in this thesis. Additionally, the LLMs that are tested are introduced, and an overview over the application of prompts and the process of prompt engineering is provided.

In Chapter 4 the first experiment (Exp. 1) is presented, where the models are prompted in a Zero-Shot manner, which can be considered most closely analogous to human experiments. After the LLMs failure to derive atypicality inferences in this setting, Chapter 5 presents a Few-Shot prompting approach for the same task (Exp. 2). For this prompting method, the models are presented with exemplars that model the expected behavior, in hopes that the models cannot only emulate it, but also understand and apply the necessary reasoning process. This chapter also includes a perturbation analysis that aimed to test the robustness of the observed model behavior. With the Few-Shot experiment also confirming that the models do not derive atypicality inferences, Chapter 6 then aims to analyze the model behavior and the potential points of failure more closely, by breaking the derivation of atypicality inferences into four distinct steps, and testing the models' performance on each step in isolation (Exp. 3).

While taking the previous findings into consideration, Chapter 7 then tests an additional prompting method called Generated Knowledge Prompting (Liu et al., 2022) with the goal of eliciting atypicality inferences from the models (Exp. 4). This is followed up with Chapter 8, where a prompting method called Three-hop Reasoning (Fei et al., 2023) is adapted and applied with the same goal (Exp. 5). Chapter 9 then takes a step back from actually aiming to elicit atypicality inferences in LLMs, and instead considers the robustness and faithfulness of the exhibited model behavior by applying a different scale for assessing the derivation of atypicality inferences (Exp. 6.1), and a self-calibration method (Tian et al. (2023), Exp. 6.2), before finally presenting a probe analysis into the

[2]All code for data generation, collection, processing and statistical analysis, as well as all data files can be found at: `https://github.com/Lotta-K/Thesis_Atypicality`

different frequency expressions the models employ and how those are tied to different ratings (Exp. 6.3). Chapter 10 then presents the general conclusions that can be drawn from this work, and Chapter 11 details the limitations of this thesis and proposes future avenues of research for extending this work.

Chapter 2
Related Work

In this section, I firstly define atypicality inferences and illustrate their relation to script knowledge and reasoning. Then I focus on LLMs and the concept of natural language prompting, before offering a brief discussion on LLMs' abilities in regard to script knowledge, pragmatics and reasoning. Afterwards I introduce prompting methods designed specifically for reasoning tasks, and further outline some recent work that relied on natural language prompting for investigating model behavior analogous to human behavior. Finally, the practice of having LLMs work with concrete numbers as well as having the models externalize their reasoning is discussed, to reveal some potential problems this might cause.

2.1 Atypicality Inferences, Script Knowledge and Reasoning

Atypicality inferences, as identified and investigated by Kravtchenko and Demberg (2022), are a type of inference that people can draw to accommodate informational redundancies at the background world knowledge level. While redundancies in language production are generally viewed as infelicitous (Grice, 1975), comprehenders are frequently able to accept them by either tolerating (Davies and Katsos, 2010) or accommodating them. Specifically, the redundancies Kravtchenko and Demberg studied occur in utterances that are seen as over-informative and consequently redundant because of related script knowledge.

Script knowledge (Schank and Abelson, 1975) refers to the shared knowledge of typical event sequences for common everyday activities such as *going to a restaurant* or *going grocery shopping*. For example, the event sequence for going to a restaurant might include: asking for a table – looking at the menu – ordering food and drinks – eating and drinking – paying the bill. In fact, the knowledge about these common activities is so prevalent to people that they can easily predict the next event in the sequence (Zwaan et al., 1995), and tend to recall the events as part of stories even if they were not specifically mentioned (Bower et al., 1979). With such a strong association between the stereotypical activity,

and the individual events in that activity's event sequence, it seems that utterances such as:

(1) *Mary went to a restaurant.* ***She ate there!***

are in fact providing redundant information to the comprehender, as the matter of eating is implied through the script of going to a restaurant. Kravtchenko and Demberg (2022) investigated the behavior of humans encountering such redundancies by having subjects provide typicality ratings for habitual activities taken from the script. When the provided context followed the script-schema, subjects assigned high typicality ratings, meaning that subjects believed that Mary usually eats in restaurants. However, when the utterance in (1) was present in the story, the subjects' ratings about Mary typically eating when going to a restaurant were significantly lower (no utterance: 85.79 vs. utterance: 72.37; $p < .001$) – see also Figure Figure 2.1[1].

This shows that such overt mention of an event triggers an atypicality inference, where the comprehender accommodates the redundancy by adjusting their previous belief about the typicality of the mentioned activity. As stated above, producing a redundant utterance is infelicitous according to Grice (1975), as it violates the conversational maxim of quantity by being over-informative. Violating a conversational norm can lead the comprehender to reason above the literal content of the utterance. In this case, the comprehender attempts to 'repair' the redundancy by inferring that the behavior is actually atypical in the presented context, which make the utterance no longer over-informative.

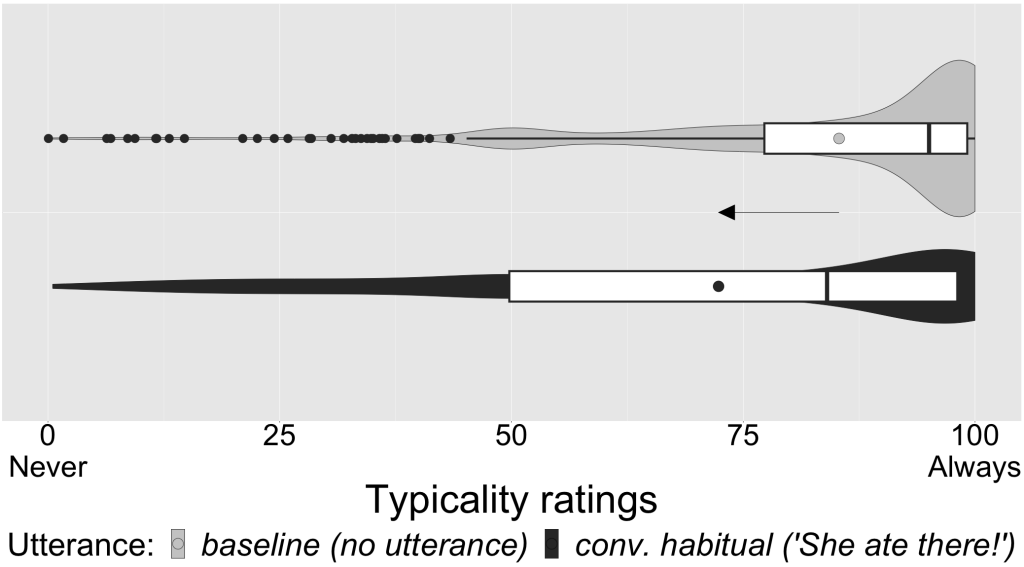

Figure 2.1: Human ratings of event typicality (e.g., eating when going to a restaurant) taken from Kravtchenko and Demberg (2022). Violin plots, overlaid with box plots, show the distribution of ratings. Circles represent mean values. The arrow indicates a statistically significant difference in ratings between conditions.

This effect crucially depends on informational redundancy – it disappeared when the context is manipulated to make the utterance no longer informationally redundant, i.e. because the context states that Mary doesn't like to eat out, and is also not present when

[1]For detailed information on the experiment setup and the experimental materials, please see Section 3.1

the context includes an utterance that is not referring to a predictable script event such as: "She got to see their kitchen!"

While Kravtchenko and Demberg measured atypicality inferences only through these typicality ratings, recent work by Ryzhova et al. (2023) builds on this study by additionally asking the participants to explain their given rating. The provided explanations offer insights into the concrete inferences the participants drew, and how they potentially accommodated the atypicality. Furthermore, they administered individual differences test for cognitive factors to all participants. The participants explanations were then classified according to 5 different annotation labels. In addition to of course assigning *atypicality* when the subjects derive atypicality inferences, the label *normal* was introduced for when the explanation showed no atypicality inference, i.e. the event of eating is considered to be the 'normal' thing to do at a restaurant, Further, *not_sure* was assigned when participant expressed being unsure. Explanations were classified as *notice_reject* when they indicated that participants drew the inference, i.e. considered atypicality, but then rejected it in favor of 'normal' behavior. Lastly, responses were classified as *other* when they could fall into multiple categories or none of them.

In most cases, subjects derived atypicality inference (*atypicality*). As expected, Ryzhova et al. found that responses that indicated the drawing of an atypicality inference corresponded with low typicality ratings. Interestingly, subjects oftentimes effectively augmented the common ground to make the informationally redundant utterance informative with respect to the context. In doing so, they provided justification of **why** Mary does not usually eat ("...because she interviews people there"). When subjects did not derive atypicality inferences, their explanations included various formulations of stating that the activity is typical human behavior (*normal*). Such answers were associated with high typicality ratings, and comprised a second-biggest annotation category. Additionally, *notice_reject* corresponded with similar ratings as the *normal* responses, hence showing that they indeed rejected the inference, and that having rejected the inference has similar implications as not drawing an inference at all. When looking at the individual participants, they also found that most were consistent in their behavior, with the majority of their responses either being labeled *atypicality*, or *normal*/*notice_reject*. This indicates that individual participants exhibited tendencies for dealing with over-informativeness, i.e. they either tended to be literal comprehenders (no inferences beyond the literal meaning) or pragmatic comprehenders (derivation of pragmatic inference, i.e. the atypicality inference). Finally, Ryzhova et al. showed that the cognitive properties identified through the individual differences test battery exhibited an effect on the type of responses, participants gave, i.e. influenced whether they comprehended the over-informativeness in a literal or pragmatic way. Specifically, they identified a significant effect of reasoning ability as measured by this test, implying that participants with higher reasoning abilities draw more atypicality inferences, i.e. comprehend the over-informativeness in a pragmatic way.

Based on their findings, Ryzhova et al. (2023) proposed that the process of deriving atypicality inferences can be broken down into four distinct steps:

1) identify the redundancy based on script knowledge;

2) realize that redundancy is infelicitous, as it violates conversational norms;

3) infer activity atypicality;

4) explicitly accommodate atypicality in situational context

Ultimately, results of Ryzhova et al. (2023) confirm that informationally redundant utterances lead subjects to infer atypical behavior, and that they go through an accommodation process: in order to obtain a consistent picture, they come up with a circumstance leading to the activity being worth mentioning (e.g., usually ordering only drinks makes eating noteworthy). In combination with the observed effect of individual reasoning ability as identified by their test battery, these observations of human behavior provide a basis for comparison with the behavior and reasoning processes of LLMs.

2.2 Large Language Models

In recent years, and especially after Vaswani et al. (2017) introduced the transformer architecture, new language models with impressive performance improvements continue to be developed. Many of them build directly on transformers, such as Devlin et al. (2019)'s BERT (Bidirectional Encoder Representations from Transformer), which spawned a whole family of related language models (RoBERTa (Liu et al., 2019), sentence-BERT (Reimers and Gurevych, 2019), prompt-BERT (Jiang et al., 2022) *inter alia*) that each offer an adjusted architecture or fine-tuning mechanisms and overall performance gains on specific tasks. Similarly based on transformers, decoder-based autoregressive generative models such as GPT-3 (Brown et al., 2020) or T5 (Raffel et al., 2020) were conceived, that treat every task as a generative problem and conceptually only rely on pre-training, potentially eliminating the need for cost- and labor-intensive fine-tuning. In addition to encoder- or decoder-based models, there is also BART (Lewis et al., 2020) which offered an attempt to unify decoder- and encoder-based approaches.

Focusing on recent decoder-based generative models, they have at least at surface level impressed by their seeming ability to comprehend natural language and generate an appropriate natural language output. Furthermore, the jumps in output quality have been considerable with each new model. As mentioned previously, this has led to the question of how much actual language understanding these models are capable of, and how their linguistic and pragmatic competencies compare to humans. Beyond linguistic competencies, further emergent cognitive abilities such as reasoning are garnering more and more interest (Huang et al., 2022a).

As the importance of script knowledge for drawing atypicality inferences and the role of reasoning has been detailed above, I will now focus on previous research on those factors in LLMs. Specifically, we will first take a look at natural language prompting, before moving onto script knowledge, and pragmatic and reasoning abilities in LLMs. Afterwards, prompting methods designed for reasoning tasks will be introduced. Finally, an overview over recent work investigating model behavior analogous to human behavior will be given, and problems with dealing with concrete numbers and externalized model reasoning will be discussed.

2.2.1 Natural Language Prompting for LLMs

With the introduction of new LLMs, the idea of prompting models with natural language input gained prominence. While initial prompting methods leaned heavily into the next-word prediction abilities by following a fill-in-the-blank style (Petroni et al., 2019), recently the focus has been on end-to-end prompting with unrestrained natural language input as popularized by Brown et al. (2020), and made widely publicly known and

available through Chat-GPT[2]. The most recent generation of LLMs (GPT-3 (Brown et al., 2020), Claude (Bai et al., 2022), Gemini (Anil et al., 2023) *inter alia*) are interactive, offering near-human like generated output when prompted with just about any conceivable natural language (or even multi-modal) input. While the input does not need to be restrained, the process of prompting has also given rise to the field of prompt engineering, i.e. the study of carefully optimizing wording and structure of input in order to optimize the output. In a way, developing the most effective prompt method for a specific task has taken the place of fine-tuning.

As with any newly emergent concept, there are a multitude of different approaches and supposed best practices, as well as ongoing research into the working of natural language prompts. It is, however, agreed upon, that how you say things matters, i.e. the words chosen and the order in which information is presented in a prompt affect the performance (Reynolds and Mcdonell, 2021, Zhao et al., 2021). Hence, most recent works detail their prompt development process to some extent, or include robustness testing with alternative prompts (cf. Wei et al., 2022, Han et al., 2024, Sancheti and Rudinger, 2022), and a catalog of different prompting approaches and best practices for optimizing performance has been proposed (White et al., 2023). Generally, natural language prompting can be separated into Zero-, One-, and Few-Shot prompting methods (Brown et al., 2020). With Zero-Shot prompting the model is presented with just instructions and a task, and with One- and Few-Shot prompting the model is presented with instructions with one example and a task, and instructions with k examples and a task respectively. Upon introducing GPT-3, Brown et al. note that the model performs best in the latter setting (though the other two settings are also considered promising).

2.2.2 LLMs and script knowledge

To investigate whether LLMs learn script knowledge and have an internal representation of it that they can access, research focuses on probing tasks that have the model externalize script knowledge. Sancheti and Rudinger (2022) tasked GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020) and BART((Lewis et al., 2020), with generating Event Sequence Descriptions, i.e. scripts. Of the three models, only GPT-2 was able to produce usable event sequence descriptions for tasks such as baking a cake. However, these descriptions would often skip over necessary steps, or order steps incorrectly, showing a lack of temporal understanding of the event sequence, and overall only very surface level script knowledge.

GPT-3 on the other hand, already showed a more promising understanding of script knowledge, when Huang et al. (2022b) investigated its ability to generate a set of actionable steps for performing a high-level task. Here, given a fixed task description of common tasks such as making breakfast, the model generated a plausible plan, hence demonstrating script knowledge. Similarly Yuan et al. (2023) found GPT-3 capable of generating scripts for abstract goals, such as baking a cake, but unable to produce adequate scripts when a constraint is imposed, e.g. when the goal is more concrete such as baking a cake for a diabetic.

While the extent to which newer models appear to have script knowledge remains to be studied, there appears to be a trend of improvement, with newer models exhibiting more script knowledge – at least when specifically probed for it. However, it is not only relevant if the script knowledge is learned by the model, but also if the model is able to access it and integrate it into the task solving process. In that line of thought, Hong et al.

[2]https://openai.com/

(2023)'s work on causal relations between two script-related events finds that models, unlike humans, are unable to infer or predict a cause/event from script knowledge, if it is omitted in the text. This does imply that the models they tested (including variants of GPT-3.5 and Vicuna) either exhibit evidence of an insufficient representation of script knowledge or an inability to integrate it.

2.2.3 Pragmatic and reasoning abilities, and how to assess them in LLMs

Recent studies have shown that LLMs can oftentimes provide responses that are consistent with pragmatic interpretations, e.g., Qiu et al. (2023). LLMs also have exhibited the ability to infer discourse connectors based on pragmatic cues (Pandia et al., 2021), but without substantial human-like temporal preferences, and only in naturally occurring data (as opposed to contexts that isolate high-level pragmatic cues). Lin et al. (2024) found that LLMs encode rich lexical-semantic information about scalar adjectives, which does, however, not necessarily transfer into a good performance on pragmatic reasoning tasks about their scalar diversity.

An analysis of seven different pragmatic phenomena (including humor, coherence and irony) by Hu et al. (2022) found LLMs to exhibit similar accuracy and error patterns as humans. Further, Barattieri di San Pietro et al. (2023) reported LLMs performing well on tests developed for assessing the pragmatic ability of humans, and Strachan et al. (2024)'s battery of tests aimed at theory of mind abilities and mentalistic inferences found GPT-4 models to perform on par with or better than humans.

Similarly, LLMs have started to exhibit abilities in regard to reasoning tasks. While overarching emergent reasoning abilities in language models and the nuances of what actually constitutes reasoning are out of the scope of this work (for a detailed discussion into emergent reasoning in large language models and the definitions of reasoning see Huang et al., 2022a), LLMs' performance on reasoning tasks such as logical or common sense reasoning improved drastically through introducing specialized prompt methods. These prompt methods often build on the idea that reasoning is a step-by-step process, with the correct the answer manifesting at the end of a chain of thoughts.

With LLMs impressive performance on a variety of (reasoning) tasks, the question of how similar their cognitive abilities are to human ones, or as in how far these models offer a good representation of general intelligence, is being posed. This question has been addressed by performing experiments with LLMs analogous to human experiments, i.e. by treating the models like participants and having them perform the same task.

Han et al. (2024) took the reasoning task of property induction and tested the performance of GPT-3.5 with the Completion API and GPT-4 with the Chat API. The models were tested analogous to human participants, and their judgements were compared against those of the participants. For obtaining judgements from GPT-4 they set the temperature to t=0 and only obtained one response per problem, as they did not see a large enough variance in the final ratings to warrant obtaining a distribution. Overall, they found that GPT-4 performance was remarkably similar to human performance in all tested phenomenons besides non-monotonicity, while GPT-3.5-turbo struggled overall.

Webb et al. (2023) also tested GPT-3 analogous to human participants by giving it tasks for reasoning by analogy. They find that the models' ability matches or outperforms that of humans on several tasks. It does however, unsurprisingly, fall short in any tasks that require a good physical understanding of the world (i.e. transfer problems for simple

tools), and evaluating causal relation analogies. Overall, they did find the model's ability to generalize from relational patterns to be surprisingly human-like.

Dasgupta et al. (2022) compare the performance of human and LLMs for abstract reasoning, arguing that the imperfect reasoning found in models actually mirrors the imperfections seen in human reasoning. They focused specifically on presence of content effects, where familiarity, believability and grounded-ness of the situation influences humans reasoning ability, i.e. arguments with believable conclusions are more likely to be judged as valid, than ones with unbelievable conclusions; even if both arguments are logically equivalent. When comparing humans and state-of-the-art LLMs' performances across three tasks, they found very similar content effects, with the more complex tasks showing some deviations in model error patterns and some LLMs outperforming humans.

The following section Subsection 2.2.4 will provide an in depth discussion of different prompt methods for reasoning tasks.

2.2.4 Prompting LLMs for reasoning tasks

There are various prompt methods aimed specifically at various reasoning tasks, many of which aim to emulate or evoke human-like reasoning chains or thought processes. The most basic application of the idea that complex problems are usually solved through a chain of thoughts or thinking step by step, is a Zero-Shot method introduced by Kojima et al. (2022). They propose simply appending "Let's think step by step" at the end of a prompt, which quite successfully causes the model to externalize (i.e. output) reasoning steps and improves accuracy, for example on arithmetic reasoning tasks (Chen et al., 2023).

This concept was extended with Chain-of-Thought prompting introduced by Wei et al. (2022), which is a Few-Shot prompting method, where the model is provided with exemplars that consist of a task/questions, a reasoning chain, and the correct answer. This was successfully applied to arithmetic, symbolic and common sense reasoning, but can be labor-intensive, as it relies on using quality exemplars with reasoning chains. Notably though, Wei et al. found Chain-of-Thought prompting to be robust, with different exemplars crafted by different people all improving significantly on the baseline, despite the previously discussed sensitivity of LLMs to specific wording and information ordering. Additionally, efforts to automatize the reasoning chain crafting process have been introduced by Zhang et al. (2022). Wang et al. (2022) introduced another variant of this method using self-consistency as a decoding strategy. With self-consistency, the prediction/answer is chosen by determining which answer is supported by multiple different reasoning chains. This leverages the idea that there is more than one way to reach a conclusion, and that a conclusion that was reached in multiple independent ways is more likely to be correct. Their method led to improved performance on arithmetic and common sense reasoning tasks.

Another prompting method introduced by Fei et al. (2023) also follows the step-by-step process, but does not provide the model with exemplars to emulate. For the task of identifying implicit sentiment, they propose a Three-hop Reasoning strategy, where the task is solved in three steps, i.e. hops, with each hop being increasingly difficult. In each hop, one step of the reasoning process is elicited from the model by asking a specifically crafted question. The model's response is then appended to the prompt before the next step, allowing the model to consider this knowledge when then answering the next, more difficult question. For the task of implicit sentiment classification, the first hop

seeks to identify the fine-grained aspect of the input text, the second hop then infers the opinion on that fine-grained aspect, so that the third and final hop can correctly identify the polarity of the input text.

Obviously, not all prompt methods aimed at reasoning tasks rely so heavily on emulating human-like thought processes or decomposing a problem into steps. Liu et al. (2022) introduced Generated Knowledge Prompting, which focuses on the knowledge that is necessary specifically for common sense reasoning tasks. To ensure that a model accesses the relevant knowledge for arriving at the correct answer, the information is made explicit by first using a Few-Shot approach to having the model generate knowledge that is relevant to the task. This generated knowledge is then appended before the actual question, allowing the model to correctly predict the answers for common sense reasoning tasks.

2.2.5 Uncertainty in LLMs: dealing with concrete numbers and unfaithful externalized reasoning

While LLMs have been exhibiting impressive behavior across a variety of tasks, it has been previously noted that they tend to struggle with task that involve concrete numbers, such as e.g. tasks involving arithmetic (Schwartz et al., 2024). Consequently, research has also focused on improving handling of numbers in LLMs and specifically improving performance on tasks that rely on numbers (Schwartz et al., 2024, Hong et al., 2024). However, the work of Tian et al. (2023) actually saw no difference between expressing probability in natural language or numbers. Ultimately LLMs' abilities to handle numbers, and to what tasks the potential problems extend does remain under investigation, and should consequently be kept in mind when dealing with numbers as is the case with obtaining typicality ratings on a 0 to 100 scale.

With successful performance on reasoning tasks and ascribing reasoning abilities to LLMs, there comes of course an interest in how far the models are actually reasoning. It should be noted that Wei et al. (2022) themselves point out that their aforementioned Chain-of-Thought prompting improves model performance by having the model emulate human thought processes, which in itself just proves the model's ability to do such an emulation, and doesn't actually prove that the model is reasoning.

Arguably, the impression that LLMs are reasoning is not only created by final performance on reasoning tasks, but specifically by the models' abilities to externalize a reasoning that to the human eye appears like a valid equivalent of a thought process. When Turpin et al. (2023) analyzed in how far the externalized reasoning the model gave when using Chain-of-Thought prompting actually corresponded with the model prediction, they found the reasons for the prediction to be misrepresented. While the reasoning appeared plausible, it was systemically unfaithful, and the predictions were influenced not only by biased input (i.es the correct answer is always 'a', or an answer is suggested in the context), but also implicit social stereotypes – neither of which were mentioned as contributing to the prediction.

Similarly, Wang et al. (2023) and Schaeffer et al. (2023) found that the exemplars given to a model in Chain-of-Thought prompting do not actually need to contain correct or valid reasoning, to achieve performances almost on par with valid reasoning. Instead, Wang et al. found the importance to lie in the exemplars displaying the correct ordering of the steps, and being relevant to the actual task, i.e. using numbers for arithmetic reasoning. Madaan and Yazdanbakhsh (2022), however, found successful arithmetic reasoning was possible even when replacing numbers in the exemplars with symbols, showing

that the components and interaction contributing to the success of Chain-of-Thought prompting are even more complex. Finally, Lanham et al. (2023) offers a nuanced task-focused perspective, finding that the faithfulness and relevance of the Chain-of-Thought prompting to the actual performance varies across different tasks. They also found the faithfulness to decrease with an increase of model size.

Chapter 3
Experiment Setup

In this chapter, the methodology of the experiments by Kravtchenko and Demberg (2022) and Ryzhova et al. (2023) introduced in Section 2.1 are explained in more detail, and the adaptation of their methods for the work with LLMs is described. Furthermore, the LLMs that were tested for their ability to derive atypicality inferences are introduced, and some information about their functionality and differences is provided. Lastly, this chapter describes the prompting process and the way various prompt methods were applied in this work, and the process of engineering the actual prompts is elaborated.

3.1 Methods

Kravtchenko and Demberg (2022) performed their analysis using 24 stories describing common everyday scenarios, such as *going to a restaurant* or *going shopping*. In each of these scenarios, the script knowledge consists of specific sequences of events, i.e. for a story using the scenario of going to a restaurant this sequence would include reaching the restaurant, getting a table, looking at the menu, ordering food, eating, paying, and leaving the place (Bower et al., 1979, Wanzare et al., 2016).

For each event sequence, they selected one activity from the script schema as the conventionally (conv.) habitual activity; in the above example, that activity was **eating**. Additionally, they chose a non-habitual activity not taken from this event sequence; for the above scenario, that was **seeing the kitchen**. For each activity, an utterance was devised ("She ate there!", "She got to see the kitchen!").

Initially, the 24 stories are set in an ordinary common ground context, where the script knowledge event sequence is considered the most likely event sequence. Additionally, a wonky common ground context with the aim of overwriting this script knowledge was devised, i.e. the context stating that a person does not like eating out, hence making it unlikely that they would.

It thereby follows that each story underwent a 2 (ordinary vs. wonky common ground context) x 2 (conventionally habitual vs. non-habitual utterance) manipulation (see an example of an item in all conditions in Table Table 3.1).

Table 3.1: An example of a "restaurant" story by context (ordinary vs. wonky) and utterance condition (conv. habitual vs. non-habitual activity is mentioned in the utterance). A baseline for both context conditions does not include an utterance block.

Context	**ordinary**	**wonky**
	Mary is a journalist **who often goes to restaurants after her interviews**.	Mary is a journalist **who often interviews restaurant waiters, but doesn't like eating out**.
Yesterday, she went to a popular Chinese place. As she was leaving, she ran into her friend David, and they started talking about the restaurant. After they parted, David continued on his way when he suddenly ran into Sally, a mutual friend of him and Mary.		
Utterance	**conventionally habitual activity**	**non-habitual activity**
	David said to Sally: "I ran into Mary leaving that Chinese place. **She ate there!**"	David said to Sally: "I ran into Mary leaving that Chinese place. **She got to see their kitchen!**"
Q habitual	How often do you think Mary usually eats, when going to a restaurant?	
Q non-habit	How often do you think Mary usually gets to see the kitchen, when going to a restaurant?	

3.1.1 Typicality Ratings

In Kravtchenko and Demberg (2022)'s experiments, the subjects were presented a version of the story as a stimulus. After reading that story, subjects were asked to express their beliefs about the typicality of the conv. habitual activity, on a scale ranging from 0 to 100: *How often do you think Mary usually eats, when going to a restaurant?* (Never-Sometimes-Always). Additionally, they were asked to express their beliefs about the non-habitual activity as a control activity on the same scale: *How often do you think Mary usually gets to see the kitchen, when going to a restaurant?*

For each experiment described in this work, the models are given these same stimuli and questions in either the full set or a subset of conditions described above (see Table 3.2). The models are instructed with a prompt to generate their answers in a certain output format, giving a typicality rating on a scale from 0% to 100% of the time (for details on this prompt see Section 3.3, or the prompts section of each respective chapter). The ratings are extracted and compared across the different conditions using a paired t-test.

3.1.2 Explanations/Reasoning

To investigate how subjects accommodate atypicality inferences in the situational context of a story and to better understand the underlying derivation processes of atypicality inferences, Ryzhova et al. (2023) conducted a follow-up study, in which they asked participants to explain a given rating. The explanations were tagged according to whether they provided evidence for an atypicality inference having been drawn (see Table 3.3 for the complete annotation scheme)[1].

[1]Ryzhova et al. (2023) report a substantial inter-annotator agreement (Cohen's κ = 0.74 (p<.0001), 95% CI (0.7, 0.77)).

Table 3.2: Overview of the experiment materials used for each experiment

Experiment	Context	Utterance	Question	Explanation
Exp. 1 (Chapter 4)	ordinary, wonky	no utterance, conv. habitual utterance, non-habitual utterance	habitual, non-habitual	full set of conditions, analogous to humans
Exp. 2 (Chapter 5)	ordinary	no-utterance, conv. habitual utterance, non-habitual utterance	habitual, non-habitual	reduced set without the wonky context
Exp. 3 (Chapter 6)	ordinary	conv. habitual utterance	–	does not collect atypicality ratings; uses specifically designed and targeted question probes
Exp. 4 (Chapter 7)	ordinary	no-utterance, conv. habitual utterance, non-habitual utterance	habitual	reduced set without the wonky context
Exp. 5 (Chapter 8)	ordinary	conv. habitual utterance	habitual	reduced set without the wonky context, without collecting data for the baseline and non-habitual utterance; the prompting process is designed specifically for the presences of a redundancy
Exp. 6.1 (Section 9.2)	ordinary	no-utterance, conv. habitual utterance, non-habitual utterance	habitual	reduced set without the wonky context
Exp. 6.2 (Section 9.3)	ordinary	no-utterance, conv. habitual utterance, non-habitual utterance	habitual	reduced set without the wonky context
Exp. 6.3 (Section 9.4)	–	–	–	does not apply; analysis of previously collected data

They considered a subset of conditions from Kravtchenko and Demberg (2022) - namely the ordinary context's baseline (where no utterance was included in the story) and the ordinary context with the utterance about a conv. habitual activity ("She ate there!"). In addition to asking a question about the habitual activity (How often do you think Mary usually eats, when going to a restaurant?) they asked subjects to explain a given rating.

To include this type of investigation into the experiments presented in this thesis, the

models were also instructed to provide a reasoning or explanation for the rating they provided. The explanations provided in the conv. habitual utterance condition were annotated using an extended version of the annotation scheme Table 3.3[2]. The added annotation labels served to cover types of answers that were typical in LLMs, but had not been observed in humans. This includes the label *reinforced_utterance* as a subtype of *no_atypicality*, for explanations where the redundant utterance was considered a reinforcement of the typicality, and the labels *hallucination* and *incorrect_reasoning* to capture erroneous and nonsensical model generated explanations. Furthermore, the *no_atypicality* subtype *non-coop* for explanations that attribute the atypicality to the non-cooperativeness of the speaker was omitted, as it did not occur in the model explanations.

Table 3.3: Annotation scheme for explanations, adapted from Ryzhova (2024); categories that were added specifically for LLMs or adjusted significantly from the original are marked in *bold*; any two subcategories under the same annotation category can co-occur which is marked by a '+', e.g. *reinforced_utt +sk*

Annotation		**Explanation**	**Example**
atypicality	*noutt_concise*	Atypicality of the target event X is not related to the utterance. No explanation of atypical behavior is provided.	She doesn't usually eat in restaurants
	noutt_elab	Atypicality of the target event X is not related to the utterance. Explanation of atypical behavior is provided.	Maybe she also sometimes interviews people in restaurants
	utt_concise	Atypicality of the target event X is related to the utterance. No explanation of atypical behavior is provided.	Since David mentioned it, it sounds like she doesn't always eat at restaurants
	utt_elab	Atypicality of the target event X is related to the utterance. Explanation of atypical behavior is provided.	Since David mentioned it, it sounds like she doesn't always eat at restaurants. Maybe she also sometimes interviews people in restaurants

Continued on next page

[2]A subset of answers (GPT-4, Few-Shot) was annotated with two annotators with a substantial inter-annotator agreement (Cohen's κ = 0.73 (p<.0001), 95% CI (0.52, 0.93)).

Table 3.3 – continued from previous page

Annotation		Explanation	Example
no-atyp	*normal*	The explanation considers the behavior normal for that specific character due to non-script knowledge reasons and/or confirms the normalcy by pointing out that a deviation from the norm has not been stated.	"...as Susan didn't mention any exceptions or different payment methods"; "she rinses the dishes every time because she is considerate"
	reinforced-utt	The explanation states that the activity is always done because it was done and/or mentioned specifically that it was done this time.	"She often goes to restaurants to eat after her interviews. This is evident by her friend David's comment to Sally: 'I ran into Mary leaving that Chinese place. She ate there!'"
	sk	The explanation explicitly or implicitly states that the target activity is a normal or common occurrence, necessary step, or usually done.	"This is because the usual purpose of going to a restaurant is to eat"; "it can be assumed that she eats every time she goes to a restaurant"
	incorrect-reasoning	The answer does not follow an internal logic or provides no logical relation to the question.	"Based on the context, it is stated that Jim threw the can away after feeding the dog."
	not-sure	The answer expresses uncertainty with no additional explanation provided.	"You can't tell from the passage how often Mary eats in restaurants."
	atyp-reject[3]	The answer states/implies that the main character does not usually do the target activity X, BUT this statement/implication is further rejected.	"It is strange that David mentioned this, so maybe she doesn't always eat? But after interviews, Mary will be tired – she cannot just go to a restaurant for a drink after a long day. So she should eat."
	other	Otherwise.	
error		The explanation does not relate to the question at all.	

The annotations of the obtained explanations are compared against the distribution of actual ratings, as well as the ratings each category was associated with, to provide further insights into the model performance on the task of drawing atypicality inferences.

[3]This category is equivalent to the category *notice-reject* that was introduced in Section 2.1

3.2 The Models

Below, the models that were used for the experiments in this work are introduced. All of them are transformer based, decoder-only models. While there were considerations to test further open-source LLMs with differing architectures, it was ultimately decided to focus on the models that had to be considered the most promising models due to their reported performance on other tasks.

3.2.1 GPT-Models

At the time of this work, the OpenAI GPT-Models were not only the most well known LLMs through the introduction of ChatGPT, but they were also considered state-of-the-art. GPT stands for Generative Pre-trained Transformer, and uses only the transformer decoder blocks. Since the original GPT model (Radford et al., 2018), OpenAI has scaled up their model and made some modifications, i.e. in regard to layer normalization and tokenization. The amount of training data was also increased. GPT-3.5-turbo is a version of OpenAI's 175 billion parameter model GPT-3 (Brown et al., 2020) that was fine-tuned using supervised learning and optimized for chat using Reinforcement Learning from Human Feedback. At the time of this work, GPT-4 was the most recent and best-performing OpenAI model. Compared to GPT-3.5-turbo, it leverages even more training data and is larger in size (though the exact number of parameters is not officially known), and it is also optimized for chat tasks. OpenAI advertises the model for it multi-modal and multilingual abilities, and promises improved reasoning abilities compared to earlier models.

In this work, I present the performance of GPT-3.5-turbo and GPT-4. More specifically, this refers to the version *gpt-3.5-turbo-0125* and *gpt-4-0613*. The performance of GPT-4-turbo, which was released as this research was ongoing, was also evaluated. While it showed the same improvements over GPT-3.5-turbo as GPT-4 in terms of fluency and world knowledge, it did not show the improved reasoning abilities that were observed in GPT-4 and did not perform better on the task at hand. Consequently, I did not pursue the full range of experiments with GPT-4-turbo. The collected results are, however, provided in Appendix B

For both GPT-3.5-turbo and GPT-4 (as well as the experiments with GPT-4-turbo) the default temperature t=1, the default presence_penalty = 0, and the default top_p = 1 were used. I performed some experiments with adjusted temperature. A temperature increase to t >1 did lead to notably less coherent results and even fully nonsensical generations. Additionally, setting t=0 does not lead to deterministic output for the GPT-Models, hence the 'recommended' default temperature of t=1 was chosen going forward.

3.2.2 Open Source Models

In addition to the commercial OpenAI models, I also tested two open source models: Llama 3 and Mixtral – more specifically *llama3:instruct (8B)* and *mixtral:8x7b*.

Llama 3 is another transformer-based, decoder-only model, courtesy of Meta. It improves on previous Llama models with an optimized 128k tokenizer, and grouped query attention (a combination of transformer multi-head- and multi-query- attention), and is considered one of the strongest open-source LLMs. The model is also trained on comparatively large amounts of data, especially considering the relative small size of

only 8 billion parameters.

Mixtral is a sparse mixture of experts model by Mistral AI. It is another open-source decoder-only transformer and has 45 billion parameters. The "experts" in this model are 8 distinct groups of parameters, and the feed-forward block picks two "expert"-groups to process every token at every layer.

Similar to the GPT-Models, I decided to keep the default, or at least most commonly employed/recommended temperature of t=0.6, as well as the repeat_penalty = 1.2 and top_p = 0.9 for both Llama 3 and Mixtral.

3.3 Prompting and Prompt Engineering

As laid out in Subsection 2.2.3, there is precedent for prompting LLMs and performing experiments such as this analogous to human experiments. For this, the models are provided with varying instructions, and potentially exemplars, before being presented with the same material and task as a human participant. Instead of filling out a survey, as a human participant would, the model then generates a text output that optimally answer the question/solves the task.

For this, a fresh instance of the model is called up for every single request. The models are not provided with any context, or any data beyond the stimulus and questions, and the information required by the chosen prompt method. Consequently, all data collection is independent, and the model is never aware of previous answers it has given. This work uses Zero-Shot prompting for Exp. 1 (Chapter 4), and a derivation thereof for Exp. 3 (Chapter 6). Few-Shot Prompting is used for Exp. 2 (Chapter 5)[4]. Exp. 4 (Chapter 7) makes use of Generated Knowledge Prompting and Exp. 5 (Chapter 8) uses an adapted version of Three-hop Reasoning. For each of the prompting methods, a detailed description of the final prompt(s) used and all additional info provided to the model will be given in the respective chapter. Here, I will lay out the general process for engineering and choosing the prompt(s) that was applied for all prompts and prompting methods.

The specific prompt formulations used in this thesis underwent varying degrees of iterative prompt engineering. For this, the effects of different formulation were tested using GPT-3.5-turbo, and consequent adjustments were made. Once a prompt was optimized to the best ability, it was tested on the other models to confirm an adequate performance.

Creating the prompts came with three distinct challenges:

1. Consistently eliciting an actual response
2. Returning definite and committed answers in comparable values
3. Answers adhering to a consistent output format

To achieve the first point, the behavioral instructions for the model had to be tweaked. The models needed to be explicitly told to speculate and make assumptions, as they would else refuse a response on the grounds of a lack of necessary context or information.

[4]The Few-Shot Exemplars used for Exp. 2 already bared a high similarity to the type of exemplars used in Chain-of-Thought prompting. Given the results of Exp. 2, Chain-of-Thought prompting was not pursued further in this work (see Section 5.4)

Telling the models that a definitive answer was required further facilitated their ability to commit to a response, though occasionally a definitive answer was still not given. Initially, frequent problems were encountered with the model refusing to answer because it was "just a language model", which led to additionally request a second response, where the model pretends to be a human who knows the characters in the stimulus. Ultimately, the other tweaks to the prompt improved this behavior to the point where the model also consistently provided answers as "itself". Since a paired t-test revealed no significant difference between the two types of responses (i.e. responses as the model and responses pretending to be a human), the distinction between those data points was not upheld in the further analysis.[5]

For the scale, the experiments by Kravtchenko and Demberg (2022) used a continuous sliding scale from Never to Always, mapping to values of 0 and 100 respectively. Attempts at similar scale failed to elicit consistent response categories, and ultimately the model was instructed to output its ratings directly in values. Hence, a scale of 0% o 100% of the time was established, which closely corresponds to the initial scale, but appeared more consistent and accessible to the model[6].

Furthermore, the model output needed to be constrained to a format from which the ratings and reasoning could easily be retrieved. I experimented with different instructions as well as template designs (i.e. different placeholders, separators etc.) and found the simple and concise variant presented throughout this work (see e.g. Section 4.1 or Section 5.1) to be most consistently adhered to.

It should be noted that especially Mixtral did struggle to adhere to the desired output format in a Zero-Shot setting (as would Llama 3 and GPT-3.5-turbo on occasions). Different template styles did not change this issue, and for the most part the generations were parse-able, meaning the amount of ratings that needed to be extracted manually was kept low. Furthermore, all models except GPT-4 struggled with responding to two questions in the same conversational pass, and the ability to adhere to the template improved when there was only one question.

For the probing experiment presented in chapter Chapter 6 the system prompt was changed to one that was adapted from Han et al. (2024)[7]. This adjusted system prompt was also tested on GPT-3.5-turbo for Zero-Shot prompting during the iterative prompt engineering phase, where it did not perform better or more consistently than the presented prompt. The formulation used for the probing questions employed in this experiment were also varied and only well performing ones were pursued. The same is true for the prompt design of the Multi-hop Reasoning (Chapter Chapter 8) where a broad variety of reasoning hops was tested on GPT-3.5-turbo and only the three best performing were pursued for all models.

[5]The data for the recently released Llama 3 and Mixtral was collected after this distinction had been deemed unnecessary, hence the relevant sentence was removed from the prompt when prompting Llama 3 and Mixtral, and only one data point was collected for each stimulus.

[6]Chapter Chapter 9 also presents experiments using a different scale as some research has indicated that LLMS tend to struggle with task involving numbers (Schwartz et al., 2024, Hong et al., 2024), see Subsection 2.2.5

[7]The exact prompt formulations can be found in the respective chapters.

Chapter 4
Exp. 1: Zero-Shot Prompting

Zero-Shot prompting is the most straightforward adaptation of the task performed by human participants for LLMs. The models are not provided with any additional information or guidance beyond the instructions and the task. Consequently, the model was provided a system prompt detailing instructions for the task, and an output template to adhere to. Then the model was given one stimulus at a time and two questions (Q1, Q2), asking to assess the frequency of the conv. -habitual and non-habitual activity. The model was presented the stimulus in each of the modifications described above. Following the methodology detailed in Section 3.1, the ratings provided by the models across all condition are collected and compared using paired t-tests. The reasoning provided for each rating is annotated according to the scheme, and the distribution of the different labels is compared against the actual ratings.

4.1 Prompt

The following system prompt was given to the model each time prior to showing a stimulus and task[1]:

(4.1) *You will receive a context (C) and two questions (Q1, Q2). Answer the questions by rating the frequency on a scale from 0% of the time to 100% of the time. Explain your answer in no more than two sentences. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior.*

(Additionally, tell me how a person that knows the people mentioned in the context would answer the below questions, using the same scale and explaining their answer in no more than two sentences.)

Use the following template for your output, where '<>' is a placeholder for content:

[1] As detailed in Section 3.3, the sentence in parentheses was dropped from the prompt eventually as it no longer served a distinct purpose, and not only did not yield significantly different results, but also at times hindered the models' ability to stick to the output template.

X: <Responder: AI or Human>

Q: <Question>

A: <Answer>

R: <Reasoning>

4.2 Results

As humans draw atypicality inferences, there's a significant decrease in typicality rating. This experiment, however, showed no significant typicality rating changes between the baseline and the conv. habitual utterance condition across the models (see Figure 4.1). There was a non-significant change observable for GPT-3.5-turbo (mean: 94.40→97.04; t(23) = -1.94, $p > .05$) and Llama 3 (mean: 87.8→94.5; t(23) = -0.965, $p > .05$) in the opposite direction, i.e. activities are judged to be more frequent, when the utterance is seen.

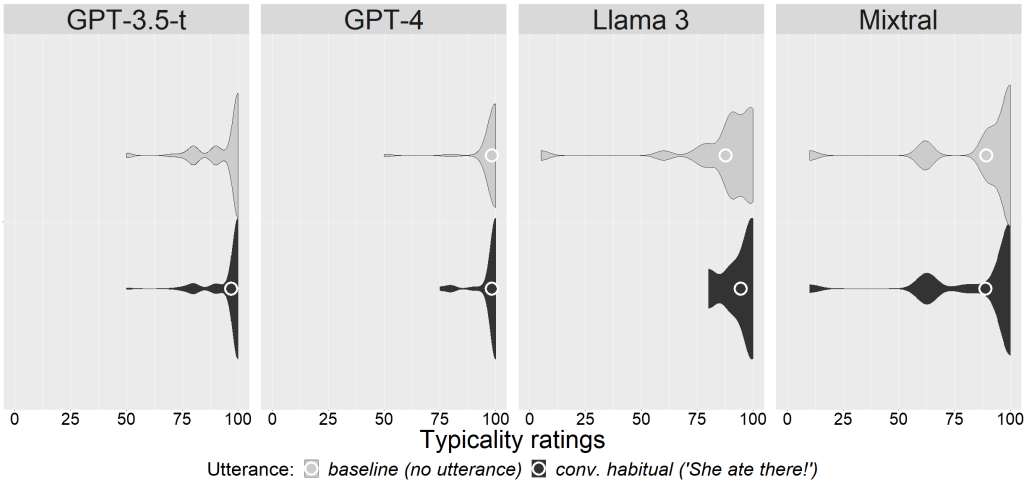

Figure 4.1: Zero-Shot, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

The typicality ratings of the con. habitual activity when the non-habitual utterance was present in the story were also checked as a sanity check. Similarly to human results, the ratings in this condition were relatively high and not significantly different from the baseline. Mixtral stands out here as the rating is decreased, though also not significantly (89.2→78.6; t(23) = 1.85, $p > .05$)[2]. The other three models exhibit the expected behavior, where the presence of the non-habitual utterance does not affect their interpretation of the habitual event typicality. In other words, the fact that Mary got to see the kitchen does not influence the typicality of her eating in the restaurant.

The manipulation of the context to state atypical behavior (wonky context) reduced the baseline typicality ratings in all models. The rating change after encountering redundancies was minimal for GPT-3.5-turbo and Mixtral, only somewhat higher for GPT-4 and almost double for Llama 3. (see Table Table 4.1) Encountering only a minor

[2]A brief manual inspection of these results showed that the drop in the mean rating was caused by a few outliers, and overall the model did still consider the typicality similarly high as previously observed.

	Wonky baseline		Wonky habitual utterance	
model	**mean**	**sd**	**mean**	**sd**
GPT-3.5-turbo	52.71	37.34	50.20	41.84
GPT-4	35.89	39.19	41.46	40.29
Llama 3	16.2	24.8	31.9	38.9
Mixtral	39.2	41.8	41.9	39.4

Table 4.1: Typicality ratings for the habitual activity in the wonky context conditions

	Baseline		Non-conventional utterance	
model	**mean**	**sd**	**mean**	**sd**
GPT-3.5-turbo	21.19	24.27	27.25	31.54
GPT-4	33.96	25.19	44.39	25.15
Llama 3	36.2	36.2	23.3	23.8
Mixtral	34.5	28.9	39.6	28.9

Table 4.2: Typicality ratings for the non-habitual activity in the normal context condition

rating change is in line with the human results obtained by Kravtchenko and Demberg (2022). As indicated by a very high standard deviation, the effect of overwriting script knowledge did vary greatly across stimuli, i.e. not all activities were equally strongly influenced by the manipulated background.

Next, the typicality ratings of the non-habitual activity were analyzed. At baseline the activity was indeed rated to be very atypical, with a high standard deviation, again showing differences across the stimuli. If compared against the typicality ratings in the non-habitual utterance condition, there was a relatively high rating change (see Table Table 4.2). While the low baseline rating is in line with the observations in Kravtchenko and Demberg (2022), the effect size is larger in the models than in humans.

The annotations of the models' explanations were in accord with the high ratings – see Table Table 4.3. The majority of responses were classified as *no_atypicality*, and largely either referred to script knowledge (*sk*; the action is typically done), a reinforcement effect of the present utterance (*reinforced_utterance*, the action is always done because they said it was done this time), or a combination of the two. Only a very small number of responses were classified as *atypicality*, and they were mostly associated with high ratings. Finally, some responses also contained hallucinated facts or incorrect or confused reasoning, or were simply erroneous generations.

4.3 Discussion

With Zero-Shot prompting, which placed the models in the setting that most closely resembled the experiment conditions for human participants, no atypicality inferences were observed. While the provided explanations occasional pointed towards a derivation, these were not associated with lower ratings. Similarly, occasional lower ratings did not come with explanations that indicated atypicality.

Given this apparent failure of the various models to draw atypicality inferences, the question is where this discrepancy between humans and models stems from. As detailed previously, Kravtchenko and Demberg (2022) and Ryzhova et al. (2023) have proposed that for the deriving of this inferences four distinct steps are necessary:

Table 4.3: Zero-Shot Reasoning Distribution (in %)

Annotation		GPT-3.5-turbo	GPT-4	Llama 3	Mixtral
atypicality	*noutt_concise*	0.92	2.08	0.00	0.00
	noutt_elaborative	3.21	2.08	4.17	0.00
	utt_elaborative	0.00	0.00	4.17	0.00
	utt_concise	0.00	0.00	0.00	4.17
no-atypicality	*normal*	5.00	10.33	16.67	0.00
	reinforced-utt +sk	5.50	8.33	8.33	4.17
	reinforced-utt	43.12	33.33	37.50	50.00
	sk	32.57	35.42	25.00	33.33
unclear	*hallucinated-facts*	5.96	6.25	0.00	0.00
	incorrect-reasoning	0.92	0.00	0.00	0.00
	not-sure	1.83	0.00	0.00	0.00
	atyp-reject	0.00	2.08	0.00	0.00
	other	0.00	0.00	4.17	4.17
error		0.00	0.00	0.00	4.17

1) identify the redundancy based on script knowledge;

2) realize that redundancy is infelicitous, as it violates conversational norms;

3) infer activity atypicality;

4) explicitly accommodate atypicality in situational context.

Provided with these steps, a likely first point of failure would be if the models lacked the relevant script knowledge, and hence could not apply it to identify the redundancy. However, in the baseline condition (no activity mentioned) typicality ratings of the conv. habitual activity are high, and the models' explanations frequently refer to script knowledge. The subsequent assessment that the models have script knowledge is further supported by the reduced typicality ratings that were obtained in a wonky context's baseline (see Table 4.1). With the wonky context, the script knowledge is overwritten by stating atypical behavior and therefore introducing it into the conversational context. This change was captured and applied by all models as they lowered their beliefs accordingly. Finally, the typicality ratings the models provided for the non-habitual activity (see Table 4.2) further solidify the models' understanding of script knowledge, and the understanding of what activities that are not part of a script and hence atypical. While the provided explanations for non-habitual activity were not annotated, the observation that the typicality is higher in the utterance condition appears to be in line with the reinforced utterance reasoning that was observed for habitual activity, i.e. something being rated as typical because it was mentioned.

While this shows that the models have the relevant script knowledge, and are able to access it to some degree, it is still possible that they fail to recognize that the observed utterance is informationally redundant, i.e. despite the script knowledge being accessible for the models, they do not actively consider it when observing the utterance and evaluating the typicality. Furthermore, it is possible that the models do recognize the redundancy, but fail at the second step of understanding that the conversational norm of quantity is violated by this over-informativeness – either because they do not know this norm, or because they don't access it while performing the task. Either of these explanations would be consistent with the fact that model justifications for high typicality ratings referred to the mentioning of the activity confirming the typicality (*reinforced_utterance*),

a type of reasoning that was typically not found in human justifications.

The following Chapter 5 presents Exp. 2 where the models are given the same task but are prompted in a Few-Shot manner, i.e. provided examples of the desired output prior to performing the task themselves, in hopes of then emulating the modeled behavior. As this has previously improved model performances on NLP tasks, this is a logical next step in trying to elicit atypicality inferences from the models. Since Chapter 5 will show that this does not lead to the derivation of atypicality inferences, Exp. 3 (Chapter 6) then analyses the potential points of failure in detail by assessing model performance on each of the four steps detailed above.

Chapter 5
Exp. 2: Few-Shot Prompting

Few-Shot prompting (Brown et al., 2020) is a popular technique in which the prompt is enriched with a small number of examples that demonstrate how to do the target task correctly. This has often been found to improve model performance on other NLP tasks (Schick and Schütze, 2020, Zhao et al., 2021).

The hypothesis was that seeing examples of atypicality inferences being drawn would facilitate the models' abilities to draw such inferences themselves. Furthermore, it was considered a possibility that the model performance would improve in ratings and reasoning distributions, but that this improvement would not account for the ability to draw the inferences, and instead relies on emulating exactly what the model sees.

To create the exemplars, a total of 4 of the stimuli were selected: specifically, the stimuli with conv. activities that were, respectively, rated most and least habitual by the human participants. For each stimulus, responses that follow the output template while mimicking human behavior in the conv. habitual utterance condition were crafted, i.e., the responses showing a lower rating and providing a justification that alluded to an atypicality inference being drawn. The models were prompted twice with two exemplars each (paired according to their ratings), responses were only collected for the conv. habitual activity (Q habitual in Table Table 3.1) in the ordinary condition. This analysis presents the combined results collapsing across exemplars[1], using the same analysis as in Exp. 1.

In addition to the Few-Shot experiment above, it was aimed to test the robustness of the inferencing ability in the Few-Shot setting in order to determine whether the model shallowly copies over and adapts the provided exemplars, or whether it uses the exemplars to pick up on the task more deeply. Results are again presented collapsing across exemplars.

[1]It was theorized that using these two different sets of stimuli as exemplars would influence the model performance, i.e. that using exemplars that are rated more typical by humans would have a stronger impact / would be more likely to cause the LLMs to adapt the modeled reasoning. While some differences were observed, e.g. the degree of significance varying, the main tendencies and findings are conserved when collapsing across exemplars. The analysis differentiating between the exemplar sets can be reproduced using the code provided at: `https://github.com/Lotta-K/Thesis_Atypicality`

5.1 Prompt

The following system prompt was given to the model each time prior to showing the two exemplars, a stimulus and task[2]. It is the same system prompt as seen in Chapter 4, amended with the information that prior to the stimulus, the model will be provided with two examples, and that there will only be one question (Q1):

(5.1) *You will receive a context (C) and one question (Q1).*

Answer the questions by rating the frequency on a scale from 0% of the time to 100% of the time. Explain your answer in no more than two sentences. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior.

(Additionally, tell me how a person that knows the people mentioned in the context would answer the below questions, using the same scale and explaining their answer in no more than two sentences.)

You will be provided with 2 examples (Ex1, Ex2).

Use the following template for your output, where '<>' is a placeholder for content:

X: < Responder: AI or Human >

Q: <Q1>

A: <Answer>

R: <Reasoning>

5.2 Results

In the Few-Shot setting, a significant difference in typicality ratings between the baseline and habitual utterance conditions was observed for GPT-4 (mean: 96.2 →84.1; t(23) = 5.82, $p < .0001$) and GPT-3.5-turbo (mean: 96.5 →89.4; t(23) = 2.98, $p < .01$). For Llama 3 there is no change (mean: 85.0 →81.2; t(23) = 0.192, $p > .05$), and the same is true for Mixtral (mean: 89.4 →90; t(23) = -0.283, $p > .05$).

The ratings being on average lower for GPT-4 and GPT-3.5-turbo when the habitual utterance was present is in line with the derivation of an atypicality inference – see Figure Figure 5.1.

In contrast to Exp. 1, the presence of the non-habitual utterance ("She got to see their kitchen!") did not have an effect on the ratings only for GPT-4 (mean: 96.2 →95.0; t(23) = 1.50, $p > .05$). For GPT-3.5-turbo, however, there was a significant change (mean 96.5 →84.0; t(23) = 3.50, $p < .01$), meaning that the ratings were on average lower in the presence of any utterance (even the one not related to the activity mentioned in the question), indicating that the model does not actually derive atypicality inferences. Interestingly, we also see a significant rating change for Llama 3 (mean 85.0 →73.1; t(23) = 2.61, $p < .05$) and Mixtral (89.4 →78.9; t(23) = 2.81, $p < .01$) further solidifying the models' failure at deriving atypicality inferences.

Next, the number of explanations in favor of atypicality inferences increased strongly in GPT-4, where *atypicality* is the most frequent annotation tag (there's a small increase

[2]As detailed in Section 3.3, the sentence in parentheses was dropped from the prompt eventually as it no longer served a distinct purpose, and not only did not yield significantly different results, but also at times hindered the models' ability to stick to the output template.

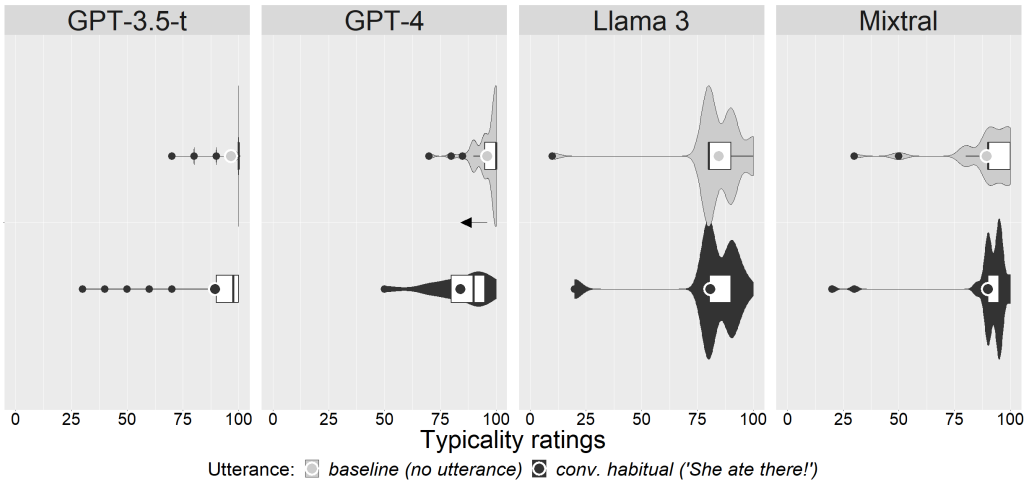

Figure 5.1: Few-Shot, habitual activity analysis

for GPT-3.5-turbo and Mixtral, and no for Llama 3, Table 5.1). All explanations pointing towards atypicality did reference the redundant utterance. It should be noted though that the atypicality justifications were sometimes inconsistent with the numerical ratings given by the model: a very cautious explanation stating a slightly decreased typicality would co-occur with a large decrease in the typicality rating. For GPT-3.5-turbo, Llama 3 and Mixtral the majority of responses are again classified for exhibiting *no_atypicality*. Within this category, there are fewer explanations classified as exhibiting script knowledge than in the Zero-Shot experiment (see Section 4.2), and more ratings that are annotated as belonging to two categories.

Table 5.1: Few-Shot Reasoning distribution (in %)

	Annotation	GPT-3.5-turbo	GPT-4	Llama 3	Mixtral
atyp	*utt_concise*	13.64	40.91	4.55	2.27
	utt_elab	2.27	11.36	2.27	11.36
no_atyp	*normal*	2.27	0	2.27	2.27
	reinforced-utt	18.18	2.27	27.27	15.91
	reinforced-utt+normal	0	0	9.09	2.27
	reinforced-utt+sk	36.36	2.27	40.91	18.18
	sk	18.18	11.36	4.55	13.64
	sk+normal	4.55	0	2.27	2.27
unclear	*atyp-reject*	4.55	27.27	2.27	27.27
	incorrect-reasoning	0	0	0	2.27
	other	0	0	4.55	2.27

Overall, the models now also show more responses that were classified as *atypicality_reject*, where the atypicality is brought up but dismissed in the justification. Notable, both Mixtral and GPT-4 do this at about the same rate and more frequently than the other two. For GPT-4 reinforced utterance has almost disappeared as a reasoning, but for the other three models it remains a frequent category, frequently in combination with script knowledge.

5.3 Perturbation Analysis

In addition to the Few-Shot experiment above, I aimed to test the robustness of the inferencing ability in the Few-Shot setting in order to determine whether the model shallowly copies over and adapts the provided exemplars, or whether it uses the exemplars to pick up on the task more deeply.

5.3.1 Perturbation 1

Firstly, the models were prompted using the same items as exemplars, but this time, only one exemplar modeled the conv. habitual utterance condition, while the second one modeled the non-habitual utterance condition. This aimed at the models' ability to differentiate between the utterances and apply only the relevant exemplar to the problem it was presented with.

This manipulation meant that the results for GPT-4 became less clear: ratings in the conv. habitual utterance condition still vary significantly from the baseline (mean: 96.2 $\rightarrow$91.9, t(23) = 2.47, $p < .05$), but the non-habitual utterance condition also varies significantly from the baseline (mean: 94.8, t(23) = 2.49, $p < .05$), and the two utterance conditions no longer vary significantly from each other. This decline in atypicality inferences is supported by the explanations, where we see *no-atypicality* for most stimuli (atypicality is only classified 8 times) and three instances of atypicality_reject. The main category of annotations is script knowledge, and the number of explanations relating to reinforced utterance remains very low.

GPT-3.5-turbo, Llama 3 and Mixtral are showing less evidence of drawing atypicality inferences in both in ratings and reasoning, with both Llama 3 (mean: 84.5 $\rightarrow$94.2; t(23)= -4.84, $p < .0001$) and Mixtral (mean: 90.1 $\rightarrow$91.2; t(23)= -0.467, $p > .05$) reverting back to assigning very high ratings in the conv. utterance condition. GPT-3.5-turbo did keep with the previous trend of lower ratings that were not statistically significant (mean: 96.8 $\rightarrow$89.7; t(23) = 1.87, $p > .05$).

In the non-habitual utterance condition, the ratings did not increase or decrease for Llama 3 (mean: 84.5 $\rightarrow$84.2; t(23) = 0.231, $p > .05$). For Mixtral there are significantly lower ratings in this condition (mean: 90.1 $\rightarrow$82.4; t(23) = 2.46, $p < .05$), and for GPT-3.5-turbo the ratings are similar to the conv. habitual utterance condition and actually significantly lower than the baseline (mean: 96.8 $\rightarrow$89.1; t(23) = 2.25, $p < .05$)

In line with the ratings, the annotations of the explanations provided by these three models also show that fewer atypicality inferences are being drawn. Interestingly, the rate of *atypicality_reject* remained very high for Mixtral. Script knowledge accounts for most annotations, with a combination of script knowledge and reinforced utterance as a close second.

5.3.2 Perturbation 2

Next, intentionally misleading and incongruent exemplars were crafted where 100% ratings paired with reasoning expressing atypicality. Two variations of that reasoning were tried: (A) expresses atypicality due to the utterance implying a change from habitual behavior, (B) simply states atypicality without any reference to habitual behavior. Notably, GPT-4 matches the exemplars the majority of the time in setting B, where we do not introduce the concept of habituality due to script knowledge. In setting A, however,

Table 5.2: Few-Shot Perturbation Reasoning distribution (in %)

	Annotation	GPT-3.5-turbo	GPT-4	Llama 3	Mixtral
atyp	*noutt_concise*	2.27	4.55	0	0
	utt_concise	4.55	9.09	0	6.82
	utt_elab	0	4.55	2.27	2.27
no_atyp	*normal*	4.55	2.27	4.55	6.82
	reinforced-utt	9.09	0	4.55	9.09
	reinforced-utt+normal	0	4.55	0	2.27
	reinforced-utt+sk	18.18	2.27	36.36	15.91
	sk	59.09	61.36	43.18	34.09
	sk+normal	0	0	9.09	0
unclear	*atyp-reject*	2.27	6.82	0	20.45
	incorrect-reasoning	0	0	0	2.27

it replicates the exemplar less than half the time, and the remaining times rejects the atypicality or assigns no atypicality. For the latter, it will frequently assign a different purpose to the utterance, explicitly stating that it does not imply atypicality.

For GPT-3.5-turbo the difference between A and B is not as pronounced, and *atypicality_reject* is observed much less frequently than in GPT-4. Both the modeled rating with the modeled atypicality reasoning (i.e. copying down), and the modeled rating with an adjusted matching *no_atypicality* reasoning appear approximately equally frequently. For B there is notably an increased number of the modeled atypicality reasoning being paired with an adjusted rating. Llama 3, on the other hand, rarely emulated both the modeled reasoning and rating. For B it matches the rating of 100% with an appropriate reasoning more than half the time, and for A, this happens the majority of the time. When the model does replicate the modeled atypicality reasoning, it is most of the time paired with a typicality rating of 0%. Notably, Llama 3 does actually generate a few nonsensical or incongruent responses in this setting.

Finally, Mixtral only "copies down" the modeled atypicality reasoning with a high rating once in setting B, and never in setting A. In both settings it can however be observed that the model provides internally inconsistent reasoning, i.e. starts out emulating the atypicality reasoning but then adds a no-atypicality reasoning. Most frequently, the model emulates the high rating and adds an appropriate no-atypicality reasoning based on script knowledge. Once again, the model also shows a high rate of explicitly rejecting the atypicality and consequently providing a high rating (in both settings). A few instances of emulating the atypicality reasoning and providing a matching low rating can also be observed; additionally, it has to be noted that inexplicably the model failed to generate responses several times, i.e. returned an empty template.

Additionally, I also crafted a second set of intentionally misleading and incongruent exemplars, modeling the reverse behavior, i.e. pairing a low rating that implies atypicality with a reasoning that models no atypicality. Again, two variants were tried; firstly providing a low rating (<40%) and then a rating of 0%. In these experiments, all four models very consistently reproduce the reasoning but ignore the modeled rating and assign a high matching rating instead.

5.4 Discussion

First of all, this Few-Shot experiment had no positive effect on the performance of Llama 3 and Mixtral, i.e. these models do not draw atypicality inferences, even when they are modeled for them. The performance of GPT-3.5-turbo seemed promising at first, but the sanity check revealed that the model cannot differentiate between the conv. habitual utterance and the non-habitual utterance, strongly indicating that the seemingly successful performance can be attributed to shallowly adapting the examples, e.g., copying down a low rating and adapting the explanation to the new target.

While the initial results for GPT-4 seemed very promising as they also held up in the sanity check, the question remained whether these responses are given for the "right reasons" (i.e., whether the examples provided in the prompt clarified the task to the model) or whether the model is adapting aspects of the answers given in the prompt in a shallow way.

The first perturbation analysis showed that GPT-4 cannot consistently differentiate between redundant and non-redundant utterances, or apply the conversational norm leading to atypicality. Unsurprisingly, this also holds true for GPT-3.5-turbo, as it had been the case during regular Few-Shot prompting. Llama 3 and Mixtral continue not to derive atypicality inferences, with Mixtral continuing the previous behavior of lowering the ratings in the non-habitual utterance condition.

With the second perturbation analysis, two different behaviors can be observed for GPT-4: (1) matching both reasoning and rating to the exemplar even if they are incongruent, and (2) copying of the rating and adjusting the reasoning. While (1) mostly implies some degree of blind copying, the occurrence of (2) shows the model applying some level of reasoning or knowledge. Interestingly, this behavior is prevalent when the exemplars provide the script knowledge and resulting habituality, and how it is overwritten by the utterance, leading the model to explicitly disagree with this modeled reasoning. This leads to the hypothesis that model does not see a problem with redundancies and hence does not apply the conversational norm that leads to the derivation of atypicality inferences, even to the point of rejecting it.

The results of Mixtral also align with this analysis, as it frequently rejects atypicality. The fact that Mixtral barely emulates the reasoning further makes the case that the model is not prone to shallow adaption; it's continued failure to derive atypicality inferences in the regular Few-Shot experiment can be taken as further evidence of non-adaptation, e.g. even seeing non-misleading exemplars modeling atypicality was not enough for the model to adapt is responses.

For GPT-3.5-turbo this analysis further confirms a tendency for shallow adaptation, while Llama 3 shows more evidence of not actually adapting the exemplars at all.

Finally, the second half of this perturbation analysis shows all models emulating the no-atypicality reasoning that aligns with their non-manipulated behavior in the Zero-Shot experiment, and adapting the rating accordingly. This confirms that none of the models emulate blindly, as they do adjust the ratings accordingly. Further, all models can emulate and adapt an exemplar to a degree; it does appear as though the emulation is facilitated when the exemplar models behavior that is similar to the behavior of the models when performing the task without exemplars, i.e. similar to the output the models produced in the Zero-Shot setting in Chapter 4.

In order to obtain better insights on the performance of all models on the reasoning steps that were previously hypothesized to be part of human reasoning for this task, the

performance of all models on the component steps of atypicality reasoning is tested in Exp. 3.

While it was initially considered to also perform an experiment using Chain-of-Thought prompting (Wei et al., 2022), this idea was disregarded based on the results of this Few-Shot experiment. Notably, the explanations provided to the models in the exemplars already closely resemble a reasoning chain for deriving atypicality inferences, and the only way to build on this would have been to craft very schematic reasoning chains, that would also not resemble any reasoning a human would potentially externalize. With the two GPT-models exhibiting a clear tendency for shallow template matching, it was considered likely that this behavior would also become more prevalent with more schematic, i.e. template-like reasoning chains. Likewise, Llama 3 exhibited practically no behavior indicating that the exemplars were applied or that their content was considered, making it unlikely that different exemplars would lead to performance improvement. Finally, even though Mixtral did seem to consider the exemplars without copying them in some cases, the majority of the time it did not seem to apply the exemplars either. Consequently, it was decided to pursue other additional prompting methods, namely Generated Knowledge Prompting (Chapter 7) and Multi-hop Reasoning (Chapter 8), instead, as they seemed more promising for both eliciting atypicality inferences and gaining further insights into the models' behavior.

Chapter 6
Exp. 3: Analysing the steps of reasoning process

As described in Section 2.1, Kravtchenko and Demberg (2022) and Ryzhova et al. (2023), the process of deriving atypicality inferences can be decomposed into the following sub-components:

1) identify the redundancy based on script knowledge;

2) realize that redundancy is infelicitous, as it violates conversational norms;

3) infer activity atypicality;

4) explicitly accommodate atypicality in situational context.

As the previous experiments have shown that the models do not consistently draw atypicality inferences, this chapter aims to test the performance of the models on each of the steps. The goal was to identify potential points of failure, i.e. an inability to perform any of the sub-components could of course explain the inability to draw the inferences at all. For this, the models were again prompted in a Zero-Shot manner, receiving an instructional system prompt and then each stimulus in the conv. habitual utterance condition, and one question at a time. The resulting responses were manually analyzed to assess the model's ability to perform each step. Below, only the question formulations that most successfully elicited what the intended behavior across models are presented[1]. The resulting responses were manually evaluated qualitatively to assess the model's ability to perform each step.

6.1 Prompt

The following system prompt was given to the model each time prior to showing a stimulus and a question. This wording was adapted from a prompt used by Han et al.

[1]For the alternative formulations and their results see Appendix A

(2024), as I considered it better suited to this open-ended question answering task:

> (6.1) *You are an expert on human behavior and communication who will be answering a question based on short contexts (C). There is no right or wrong answer to the questions you'll see, and you are willing to use your best judgement and commit to a concrete, specific response, even in cases where you can't be sure that you are correct.*
>
> *Please keep your answer as short and concise as possible. Use the following template for your output, where '<>' is a placeholder for content:*
>
> *Q: <Question>*
>
> *A: <Answer>*

6.2 Step 1: Identifying Redundancy

For identifying the informational redundancy, we report the results of the following two prompts:

- Q1: Does the direct speech contain any redundancies?
- Q2: The direct speech contains redundant information. Can you identify the redundancy and elaborate why it is one?

For Q1, where the presence of a redundancy is open-ended, GPT-3.5-turbo was not able to explicitly identify the informational redundancy. It did identify redundancies such as a grammatically unnecessary "that", and additionally sometimes stated the presence of a redundancy without elaborating on it. GPT-4 on the other hand, generally succeeded at identifying informational redundancies (18 times); it identified no redundancy and an unrelated redundancy twice each. Additionally, two times the reverse redundancy was identified, i.e. the model said that provided with a target activity such as *paying the cashier* it was unnecessary to also state the related event *grocery shopping*. Llama 3 identifies the informational redundancy 14 times, the reverse redundancy 3 times, and the remaining times the redundancy was a different one. Mixtral identified the desired redundancy 17 times, the reverse redundancy once, and either a different or an unspecified redundancy the remaining times (see Figure 6.1).

For Q2, where the presence of a redundancy was presupposed, GPT-4 identified the informational redundancy for all 24 stimuli, and Mixtral performed almost as well, identifying it for 22 stimuli (for the remaining two it once identified the reverse redundancy and once an unrelated one). The performance of GPT-3.5-turbo was also generally improved: it correctly reported the redundancy in 13 stories (4 times it reported the reverse redundancy and the remaining 7 times it was an unrelated redundancy). For Llama 3 there is no positive effect as it reported the expected redundancy 13 times for this prompt, as well as 4 times the reversed redundancy, 5 times an unrelated or incorrect one, and twice the answer provided was nonsensical (see Figure 6.1).

Overall, this finding can be taken as evidence that the model successfully draws on script knowledge and can in principle identify the informational redundancy.

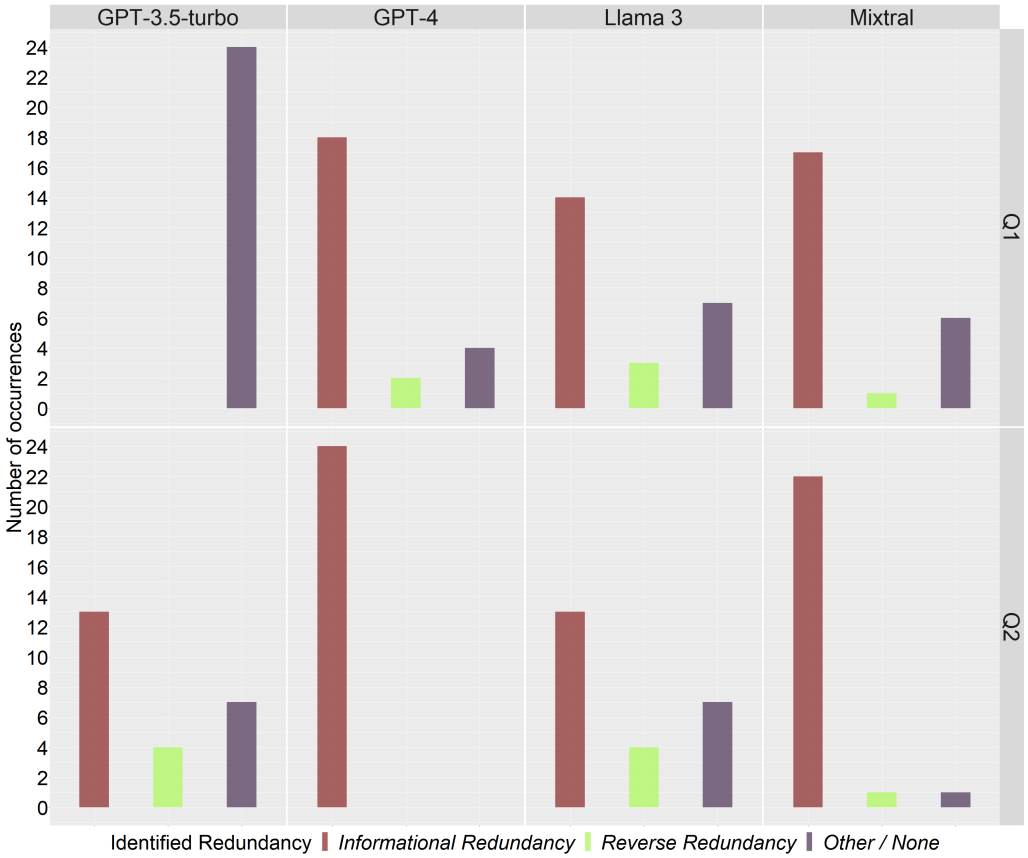

Figure 6.1: Distribution of what redundancies were identified by each model for Q1 and Q2

6.3 Step 2: Realizing that redundancy is infelicitous

The drawing of an atypicality inference is an accommodation process in which the comprehender 'repairs' an utterance that otherwise may be viewed as infelicitous due to the redundancy. Consequently, the question has to be considered, whether the conversational norm under which redundancies should be avoided (Maxim of Quantity) is known and accessible to the model. However, this aspect proved to be very difficult to assess via prompting, due to its subtlety (explicit reasoning about them would also be hard to elicit from humans, as pragmatic implicatures can always be denied – see e.g., Garmendia, 2023).

When asking the model whether the utterance including informationally redundant information was a good / effective way of communicating, GPT-3.5-turbo and GPT-4 tended to respond that redundancies could be a problem, but provided non-specific albeit reasonable examples of why redundancies can be acceptable, i.e. for emphasizing, clarification, attempts at humor etc. When asked to simply identify if the utterance acceptable regardless of the redundancy, GPT-4 found it acceptable for all 24 stimuli, and GPT-3.5-turbo for 21. The exact questions I used are reported in Appendix A. Mixtral agreed that expressing redundancies is not necessarily effective or good communication,

and also provided reasonable examples of why redundancies can be okay, often even applying its general assessment more directly to the actual story than the other two models. On general acceptability, it did consequently also find the utterance acceptable in every single stimulus. Llama 3 on the other hand, found redundancy to be inefficient / not good communication without providing examples of what could make them okay, and the model consequently identified the redundancy as problematic and unacceptable in 19 stimuli. With the questions used here Llama 3 did, however, exhibit an improved ability for correctly identifying the informational redundancy (identifying it correctly 1 times while accessing the acceptability, as opposed to 14 and 13 times with Q1 and Q2 respectively).

While this shows that the models' generally do have a concept of what communication should look like according to the conversational norms, and that redundancies could be considered a violation, all models but Llama 3 seemingly tend to regard the informational redundancies investigated here with a lot of nuance that allows favorable interpretation. Notably, this interpretation usually assumes that the redundancy serves a communicative or interpersonal purpose, rather than the purpose of relaying new information, which is what could give rise to the atypicality inferences. The investigation into step three presented below further solidifies this observation.

6.4 Step 3: Inferring Atypicality

Next, we tested whether the model can infer atypicality based on the mentioning of redundant information, using prompt Q3:

- Q3: The direct speech contains seemingly redundant information. Can you identify what I mean and explain why the speaker made the effort of conveying this information?

This wording improved the models' ability to identify the informational redundancy. GPT-4 correctly identified the redundancy for all stimuli, and provided lists of generic potential reasons (most commonly again including emphasis, occasionally pointing to some level of atypicality, i.e. forgetting, or past incidents, but also mentioning humor or the wish to establish a connection). Mixtral similarly identified all redundancies correctly again, but stated for the majority of the stimuli that the information wasn't truly redundant and likely served a purpose in the conversation, similar to the reasons GPT-4 provided.Additionally, it proposed that the redundancy could be a necessary clarification as, without confirmation, the other characters could have only assumed but not definitely known the information. GPT-3.5-turbo pointed towards undefined noteworthiness and attributed it to a desire to emphasize this information. Despite Llama 3 labelling redundancies as problematic in the previous step, the model provides reasonable and specific reasons for the redundancy. For the most part, the proposed reasons related to the conversational situation instead of the discussed activity.

Further prompt formulations were tested in order to elicit more specific explanations from the models. Best results were obtained when adjusting the question for each stimulus and detailing the specific redundancy, as shown in Q4:

- Q4: The second sentence in the direct speech conveys seemingly redundant information, because eating is a usual part of going to a restaurant. However, since it

was mentioned explicitly, it can be assumed that it is new or relevant information. Why could Mary eating be new or relevant information?

For this prompt, GPT-3.5-turbo mostly gave informative and creative explanations that were specific to the stimulus. For the majority of stimuli (13 times) that explanation did not point towards atypicality, 9 times atypicality was listed among the possible reasons, and twice atypicality was provided as the reason. However, those cases did not provide any further elaboration and no alternative behavior was provided.

GPT-4 identified atypicality 4 times, and provided reasonable accommodations for what is done instead. 16 times atypicality was mentioned on a list of potential reasons (though often times weakly in form of the person potentially forgetting sometimes), and the remaining 4 times the provided explanation did not point towards atypicality.

Llama 3 gave very specific and logical explanations of the noteworthiness for 22 stimuli, but only two of those could be classified as atypicality. Similarly, Mixtral provided specific and logical explanations of the noteworthiness for all stimuli, 5 times hinting towards varying degrees of atypicality / unusual behavior (see also Figure 6.2).

As had already been indicated by the more general results of step 2, the models are able to accommodate and accept the redundancy as 'good' and purposeful communication. This holds true even for Llama 3, which was more inclined to view it unfavorable during step 2. However, even when the question explicitly states that new information was conveyed, the models are likely to assign additional meaning to the utterance, rather than assuming that its face-value, i.e. Mary having eaten at a restaurant, is the new information. The drawing of atypicality inferences, however, is explained by encountering the informationally redundant utterance, taking it at its informational face-value, and then adjusting the common ground to accommodate the utterance, i.e. make the information no longer redundant. The fact that this seemingly not the route the models take is reflected in atypicality usually being only presented as one possible reason, if at all.

6.5 Step 4: Explicitly Accommodating Atypicality

Finally, at the last step the interest laid in assessing whether the model is in theory capable of 'completing' the picture that is caused by an atypicality by coming up with an alternative behavior or an explanation, i.e., whether the atypicality of the action can be accommodated if it is presupposed. The model was given the following prompt (again adjusted for each stimulus):

- Q5: The second sentence in the direct speech conveys seemingly redundant information, because eating is a usual part of going to a restaurant. However, since it was mentioned explicitly, it can be assumed that it is new or relevant information. That probably means that Mary doesn't typically eat. What does she do instead?

GPT-3.5-turbo managed to provide a specific and reasonable alternative behavior for 7 stimuli. An additional 7 times the alternative is mentioned but weak or unspecified (i.e. 'uses alternative method'). 8 times no alternative behavior is provided and once the premise that the person doesn't typically do that is actually rejected by the model. Finally, for one stimulus, the model provides an alternative that is not a valid in the context.

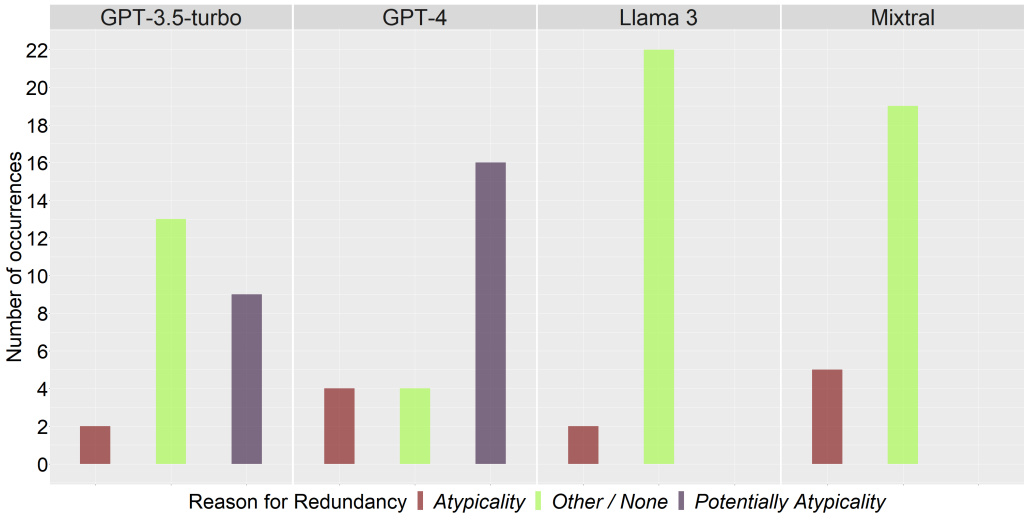

Figure 6.2: Distribution of what reasons for the redundancies were inferred by each model for Q4; potentially atypicality refers to cases where atypicality was listed among other reasons. The less effective Q3 was omitted for simplicity.

GPT-4 provides sensible alternative behaviors for 13 stimuli. Twice the proposed alternative is not a valid alternative, and 7 times the model says that an alternative cannot be inferred from the context and hence none is provided. In 2 cases it is simply stated that the person does "not that" which is not elaborative but reasonably specific in the context, i.e. if he doesn't close the fuel cap he leaves it open.

Llama 3 again committed to specific and reasonable alternative behavior for most stimuli, only twice offering a weakly specified alternative and once an illogical one.

Finally, Mixtral makes it very clear in almost all responses that a definitive alternative behavior can't be inferred. 15 times it still provides reasonable and specific alternative behaviors, while insisting for 3 stimuli that it cannot speculate. Interestingly, it rejects the notion of atypicality stated by the question 6 times, providing the previously discussed reasons for redundancy like emphasis instead. For an overview of the results, see Figure 6.3.

Generally, this has shown an ability to perform the final and fourth step of 'completing the picture' in all models to at least some degree. Notable is here Llama 3's excellent performance, especially since the model really commits to a reason instead of providing general lists. Mixtral also tends to be more specific than the GPT-models, though not to the same degree as Llama 3. What stands out is Mixtral actually rejecting the atypicality that is presupposed by the prompt. (GPT-3.5-turbo also does so once, the other two models do not reject the provided premise in this setup.)

6.6 Discussion

The different experiments performed in this chapter aimed at identifying how each of the models does on each of the steps for deriving at atypicality inferences in an isolated context. Notably, this method of prompting the model with questions that are aimed specifically at performing each of the steps does not reliably show whether or not a

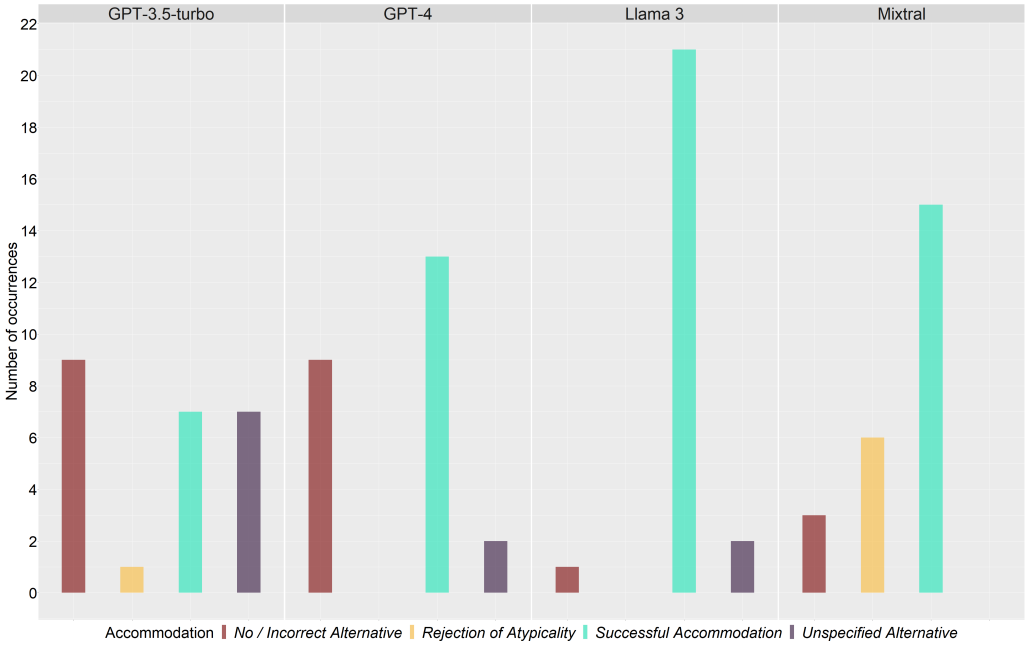

Figure 6.3: Distribution of explanations provided for atypicality by each model for Q5

given model is actually able to perform this step unprompted, or in a different context. I do however believe in the merits of assessing the models' abilities and behaviors in this controlled setting for providing initial insights into potential points of failure. Furthermore, a failure to perform at any of these steps in an isolated context and with a prompt specifically aimed at eliciting this behavior, can likely be taken as strong evidence that the model will also not successfully perform this step on its own, outside of this setting.

The findings of this experiment support the initial conclusion from Exp. 1 and Exp. 2 that the models do not lack the relevant script knowledge. A general awareness of the conversational norm also seems to be present, even though the models are very willing to accept redundancies in communication on the ground that they can serve conversational or interpersonal purposes. Notably, this behavior might be integral to not arriving at atypicality inferences, as they then don't need to accommodate the informational redundancy by adjusting their common ground.

Another notable finding from this experiment is the difference in behavior between Llama 3 and the GPT-models, with Mixtral's behavior landing somewhere between them. Llama 3 has shown much more willingness to commit to a specific answer, and also seems more suggestible, i.e. more willing to run with the information provided by the prompt, while the GPT-models tend to answer broadly, and more carefully, covering many possible options.

As this experiment strongly suggests that the LLMs' point of failure is at step 3, actually deriving atypicality, targeting this step, and generally guiding the models more strictly in their reasoning process became the main goal of further experiments. This is specifically done in Exp. 5 where the models are prompted with multi-hop reasoning. A precursor to this is Exp. 4, which test the waters with a two-step prompting method, while once again focusing on the aspect of script knowledge, and specifically the question whether

externalizing script knowledge and adding it to the context affects model behavior.

Chapter 7
Exp. 4: Generated Knowledge Prompting

Generated Knowledge Prompting (Liu et al., 2022) requires the generation of task-relevant knowledge prior to presenting the task. For this, the models were prompted in two steps. Firstly, the system prompt instructed the model to behave and respond a certain way, and then it was given the task of generating relevant knowledge. In this case, the relevant knowledge was a list or script detailing the necessary steps for the target activity (i.e. grocery shopping). In a second step, the model was then provided the instructions for the actual task (akin to the previously used system prompt), and then of course the respective stimulus and question. For Generated Knowledge Prompting, the models were again presented only Q1, and only the stimuli in the normal common ground condition.

The hypothesis was that making the model aware of the script knowledge, i.e. having it externalize it in form of a list that the model itself generates, could facilitate the models' ability to actually apply the knowledge when solving this task. As it was shown in Chapter 6, the models have the script knowledge, so the lack thereof is not the reason for failing at the task. While Ex. 3 has indicated that recognizing the redundancy based on script knowledge is also not the most likely point of failure for the models, this experiment can further solidify this point in a quantitative manner, while also providing more insights in what the script knowledge of the models actually looks like. Presented are the typicality ratings produced by the model, that are compared using a paired t-test. As before, the reasoning are annotated and the distribution is compared to the rating distribution. Additionally, the initial generation of a script is checked for coherence.

7.1 Prompts

The following system prompt was provided to the model initially:

(7.1) *You are an expert on human behavior and communication. In this conversation*

you will receive tasks and questions that you will answer as briefly and accurately as possible. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior.

Afterwards, the following generation prompt is provided to the model, adjusted for each of the stimuli (here the variation for the "going to a restaurant" stimulus is presented):

(7.2) *Generate a script or list that lays out all the steps that are part of going to a restaurant. Provide a brief but complete overview of the necessary steps without any further explanations.*

In the same conversation, meaning the model still follows the initial instructions and has access to the script it has just generated, the model is then presented with the known task and output template, followed by again the stimulus in each of the tested conditions and the question:

(7.3) *You will receive a context (C) and a question (Q1). Answer the questions by rating the frequency from 0% of the time to 100% of the time. Give a reasoning for your answer in no more than two sentences. Use the following template for your output, where '<>' is a placeholder for content:*

Q: <Q1>

A: <Answer (0%-100%)>

R: <Reasoning>

7.2 Results

Once again, there are no significant typicality rating changes between the baseline and the habitual utterance condition across the models (see Figure Figure 7.1). There was non-significant change for all four models in the opposite direction (GPT-3.5-turbo mean: 86.2→91.9;; t(23)=-1.48, $p > .05$; GPT-4 mean: 98.2→99.4; t(23)=-1.68, $p > .05$; Llama 3 mean: 80 →88.3; t(23)=-1.75, $p > .05$;Mixtral: 85→87.5; t(23)=-1, $p > .05$) i.e. activities are judged to be more frequent, when the utterance is seen. This is similar behavior to what was observed in Zero-Shot Prompting, where this trend was present in GPT-3.5-turbo and Llama 3.

As before, a sanity check comparing the typicality ratings of the habitual activity when the non-habitual utterance was present in the story was performed. Again, similarly to Zero-Shot prompting, the ratings in this condition not significantly different from the baseline for all models but Mixtral, where the ratings were significantly lower (mean: 85→76.5; t(23)= 2.19; $p < .05$). This is similar to the behavior that was previously observed in the Zero-Shot prompting (Section 4.2), and once again a manual inspection showed that this drop in mean typicality can be traced to a few outliers, e.g. the model claiming that it is unsure and has to speculate and then assigning a very low rating. The non-significant change in Llama 3 and GPT-3.5-turbo mirrored the behavior described above, i.e. an increased typicality rating from no utterance present (Llama 3 mean:80 →88.1; t(23)=-1.66, $p > .05$; GPT-3.5-turbo mean: 86.2→90; t(23)=-1.10, $p > .05$).

Once again, the explanations provided by the models' were in accordance with the high ratings – see Table Table 7.1. The vast majority of responses were classified as

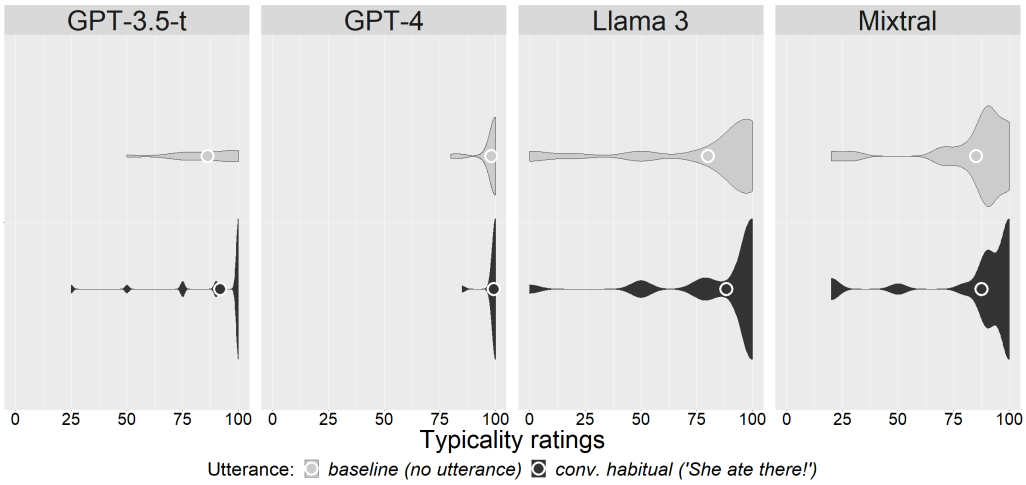

Figure 7.1: Generated Knowledge Prompting, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

no_atypicality, and *atypicality* rarely occurred in Llama 3 and GPT-3.5-turbo[1]. As seen previously, the majority of responses given by the models fell into the category of *reinforced utterance*, *script knowledge* or both categories. Notably, Mixtral exhibited a high number of explanations that couldn't be classified with in the scheme (*unclear_other*).

Additionally, it can be noted that all four models did generate valid and coherent scripts for all stimuli, thereby successfully externalizing their script knowledge. Notably, the generated scripts would occasionally be correct and sensible, without actually including the activity that was targeted by these experiments. For example, the Llama 3 script for "writing a letter" did not explicitly include mailing the letter once it is finished. However, the majority of scripts did include the target activities.

Table 7.1: Generated Knowledge Prompting Reasoning distribution (in%)

Annotation		GPT-3.5-turbo	GPT-4	Llama 3	Mixtral
atypicality	*noutt_concise*	0	0	4.17	0
	noutt_elab	0	0	4.17	0
	utt_concise	4.17	0	0	0
no-atypicality	*normal*	0	12.50	16.67	4.17
	reinforced-utt	16.67	20.83	37.50	8.33
	reinforced-utt+sk	16.67	29.17	16.67	4.17
	sk	58.33	37.50	16.67	20.83
unclear	*other*	4.17	0	4.17	16.67

[1]For both of those models, this label was in fact assigned to a stimulus that exhibits interesting behavior due to the conditional scope (visiting the doctor every few years means getting examined every few years, i.e. getting examined infrequently).

7.3 Discussion

In many ways, Generated Knowledge Prompting led to very similar results as Zero-Shot prompting. None of the models draw atypicality inferences, and instead, there is a non-significant tendency towards judging the habitual activity to be more frequent, even for GPT-4. It can therefore be concluded that having the model externalize script knowledge does not facilitate the drawing of inferences. This is also further proof that there is no lack of script knowledge, as all models were able to externalize an appropriate event description/list.

While there was no effect of simply appending the knowledge prior to the task, it seems worth considering that appending this knowledge without any further instructions is not "enough" to nudge the models towards the intended reasoning and inferences. Exp. 3 has already shown that the models exhibit a very low rate to identify over-informativeness based on script knowledge as actual redundancies, i.e. as maxim violations that need repairing. If the redundancy is presupposed, they are likely to seek a possibility to 'repair' the maxim violation by other means than adjusting common ground and inferring atypicality. Consequently, I decided to re-use the knowledge generation prompt that was used here for the Multi-hop Reasoning (Section 8.3, Section 8.4) to see if giving more context and instruction for applying the script knowledge could lead to success.

Chapter 8
Exp. 5: Prompting with Multi-hop Reasoning

For Multi-hop Reasoning, the concept of Three-hop Reasoning (Fei et al. (2023), see Subsection 2.2.4 was extended to include potentially more than three hops, and the hops were adjusted to this task. In the same manner as for generated knowledge prompting, the models are now prompted in multiple steps, with the final step again consisting of providing the actual experiment task and respective stimulus and question.

Exp. 3 (Chapter 6) has shown varying ability across the models to perform the individual steps of deriving atypicality inferences when prompted to do so. Consequently, the hypothesis was that a prompting method that would guide the models along these steps and hence make the previous steps explicitly available could facilitate the derivation of atypicality inferences. This chapter briefly discusses how Three-hop Reasoning was adapted for this work and then introduces three different prompt variants, i.e. three different step-by-step breakdowns of the task, and the results for all models, and finally a discussion tying the results of all three variants together.

8.1 Adapting Three-hop Reasoning for Atypicality Inferences

The concept of Three-hop Reasoning was devised for implicit sentiment analysis. It breaks the task down into sub-task that build on each other, increase in difficulty, and become more fine-grained. An adaptation of this concept for the task of deriving atypicality inferences considers the previously discussed steps for deriving atypicality inferences: 1) identify the redundancy based on script knowledge; 2) realize that redundancy is infelicitous, as it violates conversational norms; 3) infer activity atypicality; 4) explicitly accommodate atypicality in situational context, and the findings from Exp. 3, regarding the models' performance on these individual steps. Additionally, the actual generation of relevant scripts as discussed for Exp. 4 (Chapter 7) is incorporated.

As per the prompt engineering process described in Section 3.3, a multitude of related prompt, or rather hop, variants were tested using GPT-3.5-turbo, and only the best performing or most promising variant was pursued with the other models. Below, three variants are introduced. The content and structure of the variants build on previous findings and the results from the previous variants. Additionally, the degree to which the models are guided or 'nudged' towards deriving atypicality inferences increases in each variant.

The following system prompt, which has also been used for Exp. 3 (Chapter 6) and Exp. 4 (Chapter 7), was used for all three Multi-hop Reasoning variants:

> (8.1) *You are an expert on human behavior and communication. In this conversation you will receive tasks and questions that you will answer as briefly and accurately as possible. Always give a definitive answer, even if that means making assumptions and speculating based on common knowledge of human behavior.*

The presented results compare the obtained typicality ratings against the baseline that was obtained in the Zero-Shot setting (Exp. 1, Chapter 4), as this prompt method could not be applied for obtaining a distinct baseline due to being specifically tailored to the conv. habitual utterance condition. In addition to annotating the final model explanation, all model generations on the intermediate steps were reviewed manually, and a summary and relevant observations from this data are presented.

8.2 Multi-hop 1

The first variant, presented here, is a straight-forward adaptation of steps of deriving an atypicality inference and the findings from Exp. 3. In a first step the presence of unnecessary, or redundant, information in the direct speech is presupposed, and the models are instructed to identify this. For that, the model was the following prompt and then a singular stimulus at a time:

> (8.2) *Below you will see a story (C) containing direct speech. Identify the unnecessary information conveyed in that utterance.*

With this, all the models should relatively reliably identify the informational redundancy, and thereby perform the first step of the derivation process (identify the redundancy based on script knowledge). In the same conversation, the model is then given the following prompt:

> (8.3) *It is generally not considered "good communication" to express unnecessary information. Can you think of a reason that the speaker in context "C" chose to do so regardless?*

This step collapses the steps two (realize that redundancy is infelicitous, as it violates conversational norms), which is presupposed again, and three (infer activity atypicality), but is not explicitly nudging the model towards an atypicality reasoning. The fourth step (explicitly accommodate atypicality in situational context) is not included, as the atypicality is not presupposed.

Finally, the model is presented the actual task with an output template and of course the question:

(8.4) *Finally, considering the provided context "C" and your own observations, answer the following question (Q1). Answer the question by rating the frequency from 0% of the time to 100% of the time. Give a reasoning for your answer in no more than two sentences. Use the following template for your output, where '<>' is a placeholder for content:*

Q: <Q1>

A: <Answer (0%-100%)>

R: <Reasoning>

8.2.1 Results

As almost expected, this variant of Multi-hop Reasoning did not elicit atypicality inferences in any of the models. For GPT-4 (mean: 98.32 →98.54; t(23) = -0.253, $p > .05$), GPT-3.5-turbo (mean: 94.40 →95.00; t(23) = -0.253, $p > .05$) and Mixtral (mean: 89.19 →86.46; t(23) = 0.893, $p > .05$) no change from the baseline obtained in the Zero-Shot setting (Section 4.2) was observed. For Llama 3, the observed change was not statistically significant (mean: 87.83 →77.96; t(23) = 1.97, $p > .05$).

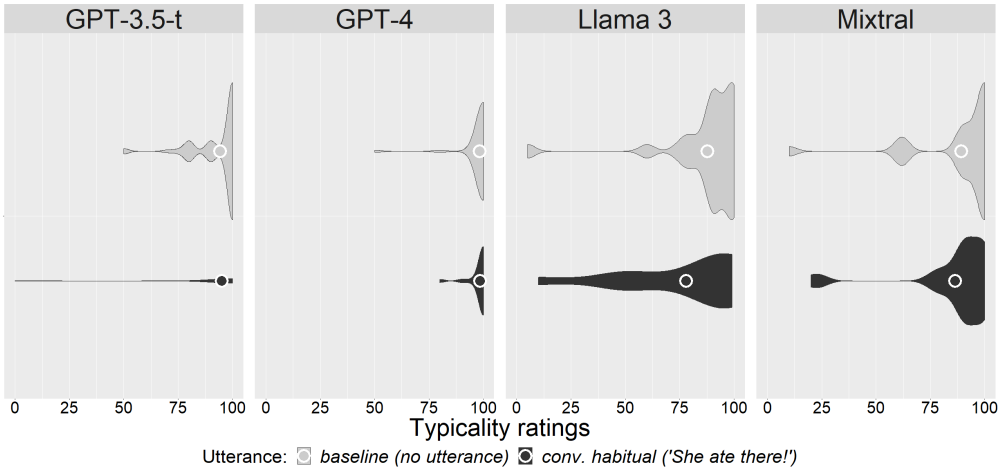

Figure 8.1: Multi-hop Reasoning, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

The analysis of the explanations (see Table 8.1) also matches this picture for GPT-3.5-turbo and GPT-4, with both models deriving no atypicality inferences. Additionally, script knowledge (*sk*), which has been already prevalent in these models' reasoning in previous experiments, now accounts for the vast majority of explanation, i.e., it appears that identifying redundancies based on script knowledge made the knowledge more prevalent in the models in later steps.

For Llama 3 it can be seen that atypicality was derived in more than 25% of all stories, which makes sense considering that the ratings showed a non-significant trend towards being lower in the presence of the conv. habitual utterance. The same is true for Mixtral, where this trend was even more minor, i.e., the mean rating change was very small. The remaining explanations are spread across the different *no_atyp* subcategories; additionally 12.5% of explanations provided by Llama 3 were classified as *unclear*, either due to

incorrect reasoning, or not being classifiable (*other*), e.g. by not clearly stating whether the character doing of the conv. habitual event or the forgetting of the event are what the model considers unusual.

Table 8.1: Multi-hop 1, Reasoning Distribution (in %)

Annotation		GPT-3.5-turbo	GPT-4	Llama 3	Mixtral
atyp	*noutt_concise*	0.00	0.00	25.00	0.00
	noutt_elab	0.00	0.00	4.17	0.00
	utt_elaborative	0.00	0.00	0.00	4.17
	utt_concise	0.00	0.00	0.00	25.00
no_atyp	*normal*	8.33	8.33	4.17	12.50
	reinforced-utt	8.33	0.00	8.33	37.50
	reinforced-utt+sk	0.00	12.50	8.33	12.50
	sk	83.33	75.00	37.50	29.17
unclear	*incorrect_reasoning*	0.00	0.00	4.17	0.00
unclear	*other*	0.00	0.00	8.33	0.00
	not-sure	0.00	0.00	0.00	4.17
	atyp-reject	0.00	4.17	0.00	0.00

In addition to analyzing the final explanations, this prompting method returns intermediate generations in response to the earlier steps, which can of course also be analyzed. The generated responses were manually reviewed in the same way responses were analyzed in Chapter 6. For the first hop (identifying the redundancy) it was checked if the identified redundancy was the informational redundancy.

For GPT-3.5-turbo, the redundancy was identified 22 times, for GPT-4 it was identified 19 times, and of the remaining messages, 4 identified the reverse redundancy (i.e. saying he went shopping is redundant when it's also mentioned that he paid a cashier). Llama 3 identified the redundancy only 12 times (and once the reverse redundancy), and Mixtral did so 14 times (and also once the reverse redundancy). Notably, out of those 12 correctly identified redundancies, only 4 occurred for stimuli where the model later derived atypicality. Consequently, the model also ended up deriving atypicality inferences (according to the annotated explanations) without first identifying the informational redundancies based on script knowledge.

Overall, it was surprising that Llama 3 and Mixtral seemed to struggle with identifying the redundancy, as they have previously shown to be able to do so. Furthermore, it's interesting that they derived the inferences despite struggling with identifying the redundancies, especially since there was no perfect overlap, i.e. atypicality was derived for stimuli where they did not identify the redundancy.

For the next hop (proposing a reason that the redundancy was expressed), each response was checked to identify whether it mentioned atypicality, and if the proposed reasons that weren't atypicality fit the pattern identified in Chapter 6, where the models assign conversational or interpersonal purposes.

A response proposing atypicality was only found once for GPT-4 and Llama 3, with the latter case mentioning it but not actually viewing it as the purpose, i.e. the character's failure to sometimes perform the conv. habitual event is hinted at to subtly characterize them. In this one instance, Llama 3 had, however, previously correctly identified the informational redundancy and does also derive atypicality in the final step. For all other cases where the model derives atypicality in the explanation, however, it did not consider atypicality in this intermediate step. As expected, the assigned purposes proposed by

all four models include emphasizing, making small-talk / conversation, establishing a connection, attempts at humor, or going into details for engaging storytelling. Interesting purposes proposed include Mixtral suggesting that re-affirming known details might actually be polite and 'correct' conversation in some cultural contexts, and GPT-3.5-turbo proposing that the speaker is actually emphasizing the correct common behavior to demonstrate positive traits (cleanliness, responsibility) and seeks validation for them.

8.2.2 Discussion

While the previous findings of this work have already indicated that giving the models a relatively open-ended question about the reason for redundancy would likely not lead to atypicality inferences, this experiment served to confirm that this holds true even when combined with the identifying of the redundancy and the actual task in one step.

In a similar manner, the next Multi-hop variant still does not strongly nudge towards atypicality. Instead, it opts to incorporate the previously discussed script generation and a more specific application of the conversational norms.

8.3 Multi-hop 2

As stated previously, this variant tries to incorporate the script generation process into the identification of the redundancy. Hence, the redundancy is no longer presupposed.

After the system prompt, this prompt combination provides the model with the script generation prompt that was used in Chapter 7, adjusted for each stimulus:

> (8.5) *Generate a script or list that lays out all the steps that are part of going to a restaurant. Provide a brief but complete overview of the necessary steps without any further explanations*

Next, the model is provided the following prompt and then the respective stimulus:

> (8.6) *Recall the gricean maxims of communication. Apply them to assess the communication that takes place in the story (C) below. Consider the knowledge about the typical process you have just laid out, and identify if the direct speech adheres to these maxims?*

With this prompt, the aim is now for the model to perform steps one and two of deriving atypicality inferences. Notably, the version presented here does not specify what maxim violation this is aimed at, and does not presuppose that a maxim is violated. This served to better asses the models' abilities of identifying the redundancies by just applying the maxims to the script.

In the third hop, the model is given the following prompt:

> (8.7) *Considering the story "C" and your previous assessment, can you identify potential reasons the speaker chose to communicate how and what they did?*

As stated previously, this still does not 'nudge' the models towards atypicality, hence still enabling all the general non-atypicality explanations that have been previously observed.

Finally, it is again provided the task, output template and question in the same manner we have seen above:

> (8.8) *Finally, considering the provided context "C" and your own observations, answer the following question (Q1). Answer the question by rating the frequency from 0% of the time to 100% of the time. Give a reasoning for your answer in no more than two sentences. Use the following template for your output, where '<>' is a placeholder for content:*
>
> *Q: <Q1>*
>
> *A: <Answer (0%-100%)>*
>
> *R: <Reasoning>*

8.3.1 Results

There is again no significant change in the typicality ratings for GPT-4 (mean: 98.32 →93.48; t(23) = 1.67, $p > .05$), Llama 3 (mean: 87.83 →83.13; t(23) = 0.966, $p > .05$) and Mixtral (mean :89.19 →84.79; t(23) = 1.07, $p < .01$)

For GPT-3.5-turbo the change is significant and occurs in the opposite direction, i.e. this Multi-hop variant has actually increased the typicality compared to the Zero-Shot baseline (mean: 94.40 →97.08; t(23) = -2.16, $p < .05$).

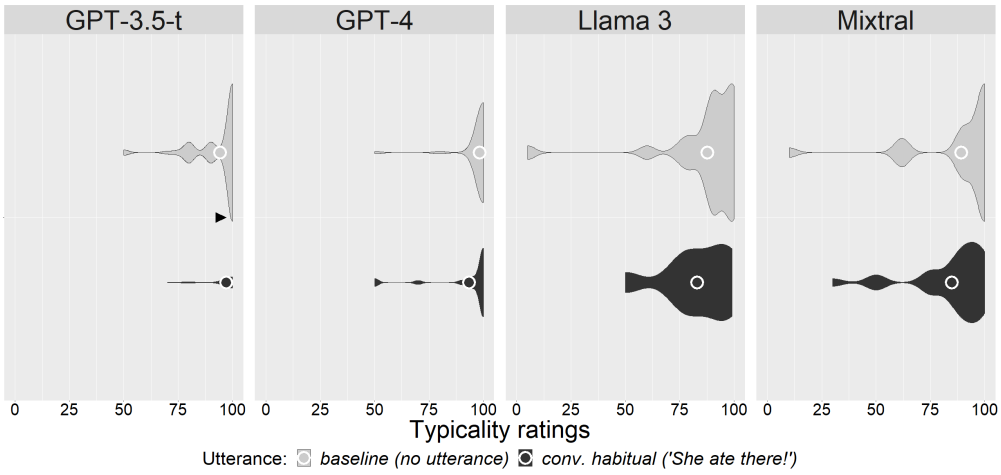

Figure 8.2: Multi-hop Reasoning, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

Once again, the annotation of the explanations reveals the same trend for GPT-3.5-turbo and GPT-4, with the former deriving no, and the latter almost no atypicality inferences. For Mixtral and Llama 3 the frequency of derivation has decreased. For all models but Mixtral, script knowledge is the main driver for seeing no atypicality. For Mixtral, the most common category is now *reinforced_utt*. For both Mixtral and Llama 3 an increase in the model being unsure *not-sure* can be observed.

The obtained responses in the earlier hops are again manually reviewed. The first hop generates scripts for the common activity, as previously done in Chapter 7, and the generated scripts are checked for correctness and completeness. And as before, the

Table 8.2: Multi-Hop 2 Reasoning Distribution (in %)

Annotation		GPT-3.5-turbo	GPT-4	Llama 3	Mixtrall
atyp	*noutt_concise*	0.00	0.00	4.17	4.17
	utt_concise	0.00	4.17	12.50	4.17
no_atyp	*normal*	4.17	0.00	16.67	8.33
	reinforced-utt	25.00	20.83	8.33	37.50
	reinforced-utt+sk	0.00	12.50	4.17	12.50
	sk	70.83	45.83	50.00	20.83
	sk+normal	0.00	4.17	0.00	0.00
unclear	*incorrect-reasoning*	0.00	0.00	4.17	0.00
	not-sure	0.00	8.33	0.00	12.50
	atyp-reject	0.00	4.17	0.00	0.00

models successfully generate valid and reasonable scripts, though on occasion these script do not use the same wording as the conv. habitual events (i.e. sending the letter vs. mailing the letter), or do not contain the relevant event (i.e. borrowing the book but no explicit mention of checking it out).

At the second hop, the models are instructed to recall and apply the gricean maxims, which are as follows (Grice, 1975):

1. Maxim of Quantity
    - Make your contribution as informative as required.
    - Do not make your contribution more informative than required.
2. Maxim of Quality
    - Do not say what you believe to be false.
    - Do not say that for which you lack adequate evidence.
3. Maxim of Relation
    - Be relevant.
4. Maxim of Manner
    - Avoid obscurity of expression.
    - Avoid ambiguity.
    - Be brief (avoid unnecessary prolixity).
    - Be orderly.

All models can correctly recall these maxims and apply them to the text. The identified violations, however, frequently do not refer to the informational redundancy and how it violates the maxim of quantity. More frequently, the models identify the maxim of relevance as being violated.

GPT-3.5-turbo sees the over-informativeness produced by the informational redundancy 4 times, but also describes the utterance as under-informative twice. GPT-4 identifies that over-informativeness 3 times, and considers the utterance under-informativeness 5 times; both types of violations are mostly classified only as potential or slight violations.

Llama 3 notably does not have a very consistent concept of these maxims, occasionally referring to them by different names, e.g. 'truthfulness', sometimes combining them, e.g. there are only three maxims and relevance and quantity are combined into 'cooperation', or introducing maxims such as 'Non-derogatory language'. Nonetheless, I was able to identify 5 violations of the maxim of quantity, or violations of an equivalent maxim, where the model assessed over-informativeness due to the informational redundancy, and similarly, 5 violations of the utterance being under-informative. Notably, the instances of identified over-informativeness do not align with the stimuli for which the model derived atypicality inferences in the explanations.

Finally, Mixtral actually does not identify any violations of the maxim of quantity, only once stating that the utterance might be slightly under-informative. Overall, Mixtral sees no maxim violations besides occasionally stating that the maxim of quality could be violated, of the statements weren't truthful.

For the third hop (identifying why the speaker chose to violate a gricean maxim), responses are again checked for mentioning of atypicality, and matching previously identified conversational and interpersonal purposes. Notably, the fact that the models frequently identified other maxim violations than intended affected the responses in this step, i.e. when no informational redundancy was identified the likelihood of inferring atypicality should be low. As expected, the reasons expressed by the model fall into the same categories that have been discussed ad nauseam. With GPT-3.5-turbo, not a single response can be identified as mentioning atypicality. For GPT-4, four responses hint more or less weakly towards atypicality as one of many potential reasons. Llama 3, much like in previous experiments, does not mention atypicality once, regardless of the fact that the model does derive some atypicality inferences in the final step. In this specific set-up, Llama 3 does however give long lists of potential reasons instead of committing to one. Finally, Mixtral also does not list atypicality among the proposed reasons. Again, these findings for all models are in line with the expectations, as the models didn't identify over-informativeness in the step before.

8.3.2 Discussion

As expected, with again not nudging more specifically towards atypicality, no atypicality inferences are derived in ratings. With this experiment, the main aim was to check how well the models identify the redundancy based on the explicit script and the maxims as well. As stated in Section 7.2, it was considered a possibility that combining the generated script with more guidance could lead to performance improvements. Further, it had to be checked whether the model would behave differently when it successfully identifies the redundancy on its own instead of it being presupposed. However, the identifying of the redundancy is not a given, and actually much less common now than in the previous multi-hop set-up (Section 8.2). It can be concluded that making the script knowledge explicit does not facilitate the recognizing of the redundancy, and that presupposing it is necessary to guide the models towards potentially deriving atypicality.

Consequently, with the next and final variant, this concept was adjusted to explicitly presuppose a violation of the maxim of quantity by being over-informative.

8.4 Multi-hop 3

With this final Multi-hop variant, the goal was to guide the models as closely as possible towards deriving atypicality inferences. Firstly, this is done by presupposing a maxim of quantity violation through over-informativeness. With this, the informational redundancy is explicitly marked as infelicitous for the model, basically leaving the model no choice but to follow the first two steps of deriving atypicality (see Section 8.1.

As a first step, this Multi-hop variant also provides the model with the script generation prompt, adjusted for each stimulus, again making the script knowledge explicit:

> (8.9) *Generate a script or list that lays out all the steps that are part of going to a restaurant. Provide a brief but complete overview of the necessary steps without any further explanations*

Next, this variant very explicitly states the violation of the maxim of quantity that occurs and asks the model to confirm this based on the provided stimulus and the generated script from the previous hop:

> (8.10)
>
> *Consider the gricean maxim of quantity for communication: The direct speech in the story (C) provided below violates this maxim by being over-informative. Can you confirm this assessment, specifically referring to the list of related common knowledge/steps that you have just generated?*

In the third step, the process of guiding the models closely towards the desired behavior is continued, using the following prompt:

> (8.11) *Do you think it is possible that the speaker wasn't being over-informative, and instead shared knowledge that was relevant and new in this specific situation, despite what common sense tells us about the usual required steps you listed earlier?*

The idea was that explicitly introducing the idea that the redundant information is new or relevant would lead the models to favor atypicality explanations over interpersonal or communicative purposes of communicating the redundant information. Finally, it is again provided the task, output template and question in the same manner as shown above:

> (8.12) *Finally, considering the provided context "C" and your own observations, answer the following question (Q1). Answer the question by rating the frequency from 0% of the time to 100% of the time. Give a reasoning for your answer in no more than two sentences. Use the following template for your output, where '<>' is a placeholder for content:*
>
> *Q: <Q1>*
>
> *A: <Answer (0%-100%)>*
>
> *R: <Reasoning>*

8.4.1 Results

Even with this Multi-hop variant, GPT-4 (mean: 98.32 →97.39; t(23) = 0.539, $p > .05$), GPT-3.5-turbo (mean: 94.40 →91.67; t(23) = 0.460, $p > .05$) and Mixtral (mean: 89.19 →93.75; t(23) = -1.02, $p > .05$) do not draw atypicality inferences. Llama 3 however, does draw atypicality inferences according to the rating change (mean: 87.83 →67.88; t(23) = 2.45, $p < .05$).

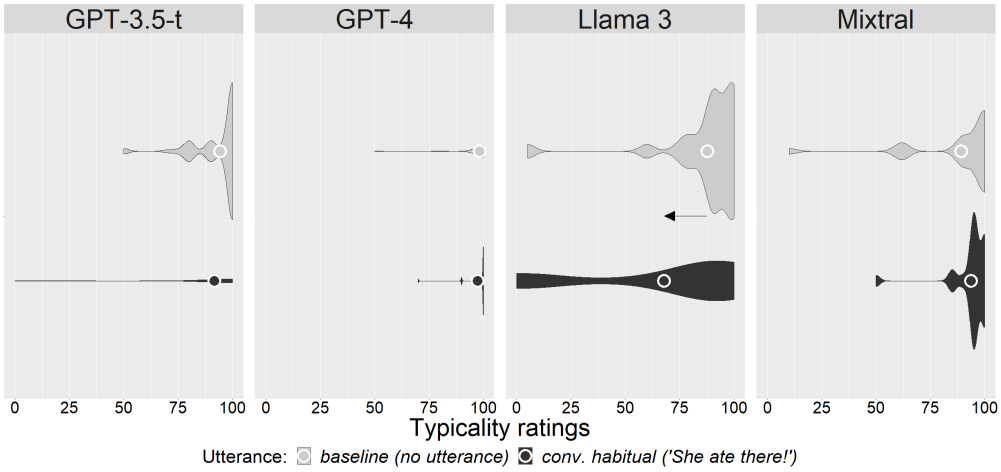

Figure 8.3: Multi-Hop-Reasoning, habitual activity analysis in the ordinary context. Boxplots are omitted, due to high skew in the data.

In line with the ratings, the explanations show (almost) no derivation of atypicality inferences for GPT-3.5-turbo, GPT-4 and Mixtral. Script knowledge as the basis for not deriving atypicality (*sk*) is again the most prevalent category; for Mixtral it accounts for over 80% of explanations, which is notable as Mixtral has previously exhibited less of a tendency to refer to script knowledge than the other models.

For Llama 3, atypicality is derived about 25% of the time. This is actually slightly lower than in Multi-hop 1 (Section 8.3), but still reasonably supports the finding that based on the ratings atypicality inferences are being derived. The largest group of explanations is here also classified as exhibiting script knowledge and inferring no atypicality. With 12.5% the number of responses that could not be classified is again high. The discussion of the performance on each of the intermediate steps provided below does show that for Llama 3 the derivation of atypicality inferences can be traced to the model exhibiting the intended behavior during the reasoning steps; however, a similar consistency is not observed in the other models.

The manual review of the first hop demanding script generation unsurprisingly showed similar results as discussed in Chapter 7 and Section 8.3, i.e., valid scripts that on occasion don't fully mention or match the conv. habitual activity.

At the second step, over-informativeness is presupposed, and the models are supposed to confirm this assessment. GPT-3.5-turbo agrees with the assessment for all 24 stories, successfully relating the over-informativeness to the previously generated script[1]. GPT-4

[1]Notably, one time the answer isn't fully as expected as it identifies the "list of steps for washing hair provided earlier [... as] much more concise and straightforward compared to the unnecessary details included in the dialogue".

Table 8.3: Multi-Hop 2 Reasoning Distribution (in %)

Annotation		gpt-3.5	gpt-4	llama	mixtral
atypicality	*noutt_concise*	0.00	0.00	4.17	0.00
	noutt_elaborative	0.00	0.00	8.33	0.00
	utt_elaborative	4.17	4.17	0.00	0.00
	utt_concise	0.00	0.00	12.50	0.00
no-atypicality	*normal*	0.00	0.00	8.33	0.00
	reinforced-utt	12.50	8.33	8.33	4.17
	reinforced-utt+sk	4.17	20.83	8.33	4.17
	sk	79.17	58.33	33.33	83.33
unclear	*hallucinated-facts*	0.00	0.00	4.17	0.00
	incorrect-reasoning	0.00	0.00	8.33	0.00
	not-sure	0.00	4.17	0.00	4.17
	atyp-reject	0.00	4.17	0.00	0.00
	other	0.00	0.00	4.17	4.17

only agrees with this assessment 9 times, the remaining 15 times arguing that the maxim is not violated, e.g. stating "[...] she is providing information that she believes is relevant. This reference [...] doesn't offer an overload of information based on their context and norms of conversation". Llama 3 also agrees with assessment for all 24 stories, even correctly identifying the redundancy based on common knowledge when the event was not part of the previously generated script, e.g. despite the script of *traveling by plane* not including *bringing your phone*, the model states "[the statement] is over-informative because it provides more information than what was necessary [...] and taking one's cell phone on board is not unusual". Finally, Mixtral also confirms the assessment for all 24 stories. At this point Llama 3, which ends up deriving atypicality inferences is performing as expected, following the intended reasoning steps. However, the same hold true for Mixtral, which does not derive atypicality inferences in this multi-hop-variant.

In the next step, the models are offered the possibility that the informational redundancy is actually communicating new and relevant information, and are supposed to assess whether they find this possibility likely. The idea was that introducing the idea of this being new information beyond commonly known steps, could nudge the models towards considering atypicality, as their usual purposes like emphasis, or small talk shouldn't be considered 'new information'.

However, GPT-3.5-turbo, while acknowledging the possibility that the information was new and relevant, still proposed very general reasons such as need for emphasis or reference to previous conversation. Another common suggestion is that the comprehender may not be aware of the steps, e.g. due to never having baked before. The model also frequently suggested that the utterance was over-informative regardless. For 4 stories, atypicality was proposed as the reason for the utterance.

After disagreeing with the presupposed violation of the maxim of quantity in the previous step, GPT-4 now provides reasons why this information is new and / or relevant for each story, and is surprisingly committed to providing context specific reasons, which has previously been a problem, i.e. the model previously preferred long general lists of potential reasons. For 14 of the stories, this proposed reason is actually some form of atypicality.

For this step Llama 3 also behaves similarly to GPT-4, but considers atypicality 6 times. Notably, 5 of those times occur for stimuli for which the model derives atypicality

inferences in the final step; for the remaining time, the model provides incorrect or nonsensical reasoning in the final step.

Finally, Mixtral does not consider it likely or even possible for most stimuli that the utterance was not over-informative. It acknowledges very generally that the context and relationships of involved character may have influence on assessing the maxim violation, but insists that the utterance is still over-informative as they are report very basic and commonly known information. For the few stimuli where it does consider an appropriate level of informativeness possible, atypicality is mentioned as the explanation three times.

8.4.2 Discussion

For GPT-3.5-turbo this Multi-hop variant also seems to not have any major positive effect towards deriving atypicality inferences. If anything, this was the most successful at having the model vehemently stick to the script knowledge and base its assessment solely on the commonly known steps (scripts). Interestingly, the performance on the intermediate steps is similar to that of Llama 3 – the model that derives atypicality inferences in this setting – without having that positive effect.

For GPT-4, the manual review of the responses to the intermediate hops did actually reveal that the third step successfully nudged the model towards considering atypicality. This is however not reflected in the ratings or the final explanations, as per the distribution of annotations. Consequently, I manually reviewed GPT-4's final explanations again, this time to see if they address the previous assessments in any way. Despite the prompt instructing the model to do so, there is no evidence that it considered the proposed reasons. Instead, the explanation, as indicated by the annotation results, very straightforwardly refer to script knowledge and / or the reinforcement through the utterance, hence not deriving atypicality inferences.

For Llama 3, the behavior in the intermediate responses is very similar to that of GPT-3.5-turbo; however, the model does end up deriving atypicality inferences in both ratings and explanations. Notably, the overlap between good performance in the intermediate step and deriving atypicality inferences in the final step is high, i.e. the model almost always derives atypicality inferences if the intermediate steps exhibited the expected / desired behavior. Regardless, Llama 3 only derives atypicality / performs the intended reasoning steps for a small number of stimuli.

Mixtral is not only not deriving atypicality inferences, it is no longer showing any tendency to do so. The analyzed intermediate responses show that this hop-variant has actually led the model to find the informational redundancy unacceptable, but to a degree where it is not really considering possible 'repairs' for the maxim violation. This had led to the model to becoming more uncertain in its typicality assessments, i.e. producing more explanations labeled *not-sure*.

8.5 Discussion

Overall, this experiment has further confirmed that the models do not derive atypicality inferences, even when prompts are designed to guide their reasoning in that direction. For GPT-3.5-turbo and GPT-4, this holds true for all three Multi-hop variants that were tested above, as the ratings are very similar across all variants. Similarly, they derive

practically no atypicality inferences in explanations either, with this behavior also being stable across all variants. GPT-3.5-turbo even maintains a similar reasoning distribution overall, exhibiting a very high tendency towards reasoning no-atypicality due to script knowledge. While this has always been present in GPT-3.5-turbo explanations (cf. Section 4.2), the reduction of *reinforced_utt* reasoning in favor of just script knowledge is notable. This indicates that the attempts at making script knowledge more prevalent and obvious throughout these variants were successful, but did not facilitate the derivation of atypicality.

For GPT-4, the prevalence of script knowledge is less obvious and consistent; overall, it seems like these different Multi-hop variants do not have a measurable influence on the model's behavior. Further, it should be noted that the third variant actually showed very promising behavior when analyzing the intermediate responses, i.e. the model did provide specific atypicality exhibiting explanations for over-informativeness in the second to last step. It did, however, not incorporate these into the final response. I have considered that this may stem from GPT-4 actually disagreeing with the presupposed over-informativeness in the step before, hence deciding to not incorporate the atypicality into its final response. This is, however, only speculation that cannot be further proven or disproved based on the collected data.

While Mixtral does also not derive atypicality inferences in ratings, there is an interesting trend observable in the distribution of explanations. For the first variant, the model actually exhibits atypicality in the explanation every fourth time. In the second variant, this number drops to only about 8% of the time, and in the final variant, not atypicality is observed in the explanations. This is counter-intuitive to the initial idea, that the degree to which the variants guide the models increase from variant to variant. In other words, the prompts that were aimed at pushing Mixtral towards identifying atypicality, actually had the opposite effect. More specifically, it seems that successfully pushing the model towards viewing the over-informativeness as infelicitous, has hindered the derivation of atypicality inferences.

Finally, Llama 3 actually did derive atypicality inferences in the third variant, and already exhibited non-significant tendencies to do so in the previous two variants. Interestingly, the number of derivations observed in the explanations remains stable across all variants, indicating that the effect of the different variants was not very strong, and instead just enough to push the rating change into significance. This becomes even more obvious, when the actual responses are reviewed and compared. Comparing the story about Lisa, who goes swimming and brings her swimsuit, across the three multi hop variants shows the following: In Multi-hop 1, the response assigns a rating of 95% and no atypicality is identified in the explanation (label: *sk*). In Multi-hop 2, this rating remains the same (95%), but the response is labeled as exhibiting atypicality:

"Since Harvey mentions that "She brought her swimsuit!", it implies that ***this is not a typical or usual behavior*** *for Lisa. If she always brought her swimsuit, there would be no need to mention it as an interesting fact, suggesting that this might have been an unusual day for her!"*

Finally, in Multi-hop 3, the rating drops to 5%, and the provided atypicality explanation falls into an extreme, where Llama 3 has not just modified its beliefs about Lisa, but completely overhauled its knowledge of how public pools work:

"Given that public pools typically offer swimwear rentals or have relaxed dress codes, it's unlikely for someone to consistently bring their own swimsuits. This suggests that Harvey's statement is providing new and relevant information about Lisa's specific situation on this particular day."

While this of course is an extreme example, it should be questioned if this actually is

the goal / desired model behavior. Furthermore, even this third variant led only to a relatively small and not very consistent adjustment, i.e. atypicality is still not derived most of the time and significant rating change is achieved by a small number of non-human-like over-corrections. Therefore ,I conclude that the use of an adapted Three-hop Reasoning was also not successful in facilitating the derivation of atypicality inferences in a human-like manner in the tested LLMs.

Chapter 9
Exp. 6: Faithfulness and Robustness of Model Reasoning

As has been stated in Subsection 2.2.5, LLMs are not necessarily faithful in externalizing their reasoning, i.e. there is no guarantee that the collected model explanation actually reflects the internal reasoning the model applied. Furthermore, the abilities of LLMs in regard to dealing with numbers are frequently questioned and under investigation (Subsection 2.2.5). This chapter firstly offers a qualitative discussion of relevant observations and considerations from the previous experiments. Then it presents a repeat of the Zero-Shot experiment (Chapter 4) using a 7-point Likert scale instead of the previously proposed 0% to 100% scale, and a Zero-Shot experiment that request self-calibration from the model. Finally, a brief and qualitative view into models' application of the 0% to 100% scale is provided.

9.1 Qualitative discussion of previous findings

For the most part, the work presented so far has however shown a relatively consistent picture, i.e. high rating were usually paired with explanations that did not show typicality. (This of course only holds true for the regular experiments (Chapter 4, Chapter 5, Chapter 7 and Chapter 8) and not the perturbation analysis presented in Section 5.3, as this intentionally provided the models with inconsistent exemplars.) It has previously been stated for Exp. 1 (see Section 4.2) that occasional derivations reflected in the explanations did not yield lower ratings. However, this is not necessarily viewed as inconsistent behavior, as these expressions of typicality were usually linguistically hedged or modified to be fairly weak ("might occasionally forget"), consequently justifying a rating that is relatively high and does not stand out among the typicality ratings. This is true especially for Llama 3 and Mixtral, as they generally assigned slightly lower typicality ratings even in the baseline condition with no utterance, i.e. in Chapter 8 Mixtral can be seen frequently deriving atypicality in the explanations which does not show up in ratings due to a low baseline.

Obviously, the need to include the annotation categories *incorrect reasoning* and *hallucinated facts* is a first indicator that the models did not always generate "good" explanations. However, it needs to be noted that Ryzhova (2024) also reports error explanations, where the reasoning provided by human subjects is strange or cannot be decoded. These error explanations are only produced by a very small number of subjects though. Generally, one would however hope for coherent and usable output from LLMs that find real world application.

Furthermore, such surface level consistency can of course not be enough to conclude that the models' explanations are faithful or that their scale application is consistent. For the former, the aforementioned perturbation experiments (Section 5.3) have shown that the model will not necessarily externalize its own reasoning and will at least under some circumstances adapt an in-congruent reasoning if it is provided in the exemplar. For the latter, annotating and generally checking and processing the data came at least with the anecdotal impression that very similar sounding explanations could lead to very different ratings. Of course, the way human subjects understand and pragmatically process certain expressions also depends on the individual and the specific circumstances (c.f. Pighin and Bonnefon (2011) on how qualifiers of certainty are perceived differently for positive and negative prospects). Regardless, it can be argued that the same model provided with same instructions should be consistent in their understanding and application of frequency expressions, especially for the same scenario, i.e. the actual frequency of doing something 'usually' may vary between using shampoo and paying the cashier, but should be consistent within application one event. Therefore, it was decided to design a probe for a potential analysis of the consistency of the frequency expression distribution, which is presented in Section 9.4.

A last point that I want to make in regard to robustness of the model behavior concerns the robustness of the models' world knowledge and their suggestibility. Specifically Llama 3 has not only generated some curios output throughout the experiment, but has shown to be easily influenced by the input provided, i.e. was much more willing to accept a presupposed atypicality in Exp. 3 and Exp. 5. This has led to very interesting adjustments of world knowledge where the model will then explain that obviously a character doesn't need to pay at the grocery store because you don't have to pay in European stores such as ALDI, or that of course a character never brings a swimsuit to the pool because all public pools have either swimsuit rentals or a lax dress code. While this behavior cannot be related directly to the behavior of human subjects and their way of dealing informational redundancies or the robustness of their world knowledge, it is an interesting point regarding the behavior of LLMs more broadly, and how reliably they can be employed for certain applications, especially as they are often viewed as more consistent and more knowledgeable than humans.

9.2 Exp. 6.1: Using a Likert Scale

As has been laid out in Section 3.3, part of the prompt engineering process was instructing the model on a scale or reference frame that the ratings needed to fall into, and that the model would adhere to. As the intention was to allow an as direct as possible comparison between the models and human participants of the experiments of Kravtchenko and Demberg (2022) and Ryzhova et al. (2023).

However, as laid out in Subsection 2.2.5, LLMs tend to struggle with tasks that involve concrete numbers. Consequently, this calls into question the reliability or lack thereof of

the 0% to 100% scale. Therefore, it was decided to test the same base experiment (Exp. 1: Zero-Shot prompting) with a Likert scale.

The below 7-point Likert scale was used to obtain rating from the model.

1. *Never*
2. *Rarely, less than 10% of the time*
3. *Occasionally, 30% of the time*
4. *Sometimes, about 50% of the time*
5. *Frequently, about 70% of the time*
6. *Usually, about 90% of the time*
7. *Every time*

Having the models generate ratings on the above Likert scale using the zero shot prompting method did not yield significant rating change between the baseline and the conv. habitual utterance condition for GPT-3.5-turbo and Mixtral, when comparing the obtained ratings using a paired t-test. For Llama 3 and GPT-4 the rating change is significant and occurs, as previously seen, in the opposite direction, i.e. the activity is judged to be more frequent when the utterance is seen (GPT-4 6.58 →6.75; t(23) = -2.14, $p < .05$; Llama 3 5.62 →6.32; t(23) = -3.39, $p < .005$)

The same sanity check as performed above did show that for all four models there is no significant rating change for the conv. habitual utterance when the non-habitual utterance is present. Furthermore, the results of using a wonky context, i.e. overwriting the script knowledge, and the typicality rating of the non-habitual activity are in line with the previously reported results, i.e. the overwriting is successful, and the non-habitual activity is rated to be atypical[1].

9.3 Exp. 6.2: Requesting Calibration

This experiment followed an approach that asks the models to self-calibrate its responses that was introduced by Tian et al. (2023). They have taken inspiration from human psychology showing that considering multiple possible answers can mitigate overconfidence, and consequently ask the models to provide multiple responses that they had to assign likelihood to. The variations of their calibration method were applied to question-answering datasets that assessed factual knowledge, i.e. datasets with question that have clear and brief answers and also a ground truth answer that they used to evaluate performance improvements. The variations they tried differed in the number of responses they requested, the way the request was worded, and by having the likelihood expressed either as a number or a natural language expression. This work applies their most consistently successful approach of considering 4 responses and assigning a probability *p* between 0 and 1.

However, adjusting the proposed calibration method to the task of providing typicality ratings as presented in this work was not successful for Llama 3, i.e. the model could not consistently follow the instructions, despite the Tian et al. (2023) using Llama-2-70b-chat for their experiments. I.e., despite asking for several answers, Llama 3 would generate

[1]The full data and scripts for the necessary analysis are available at:`https://github.com/Lotta-K/Thesis_Atypicality`

just one response, or an empty output template, or multiple responses but without assigning them probabilities. We attribute this to our more complex task and output format, and consequently cannot report results for Llama 3. Mixtral also struggled with the task of self-calibrating, only providing relevant and usable output less than half of the time, and consequently no results can be reported either. For GPT-3.5-turbo and GPT-4 the results were indiscernible from the regular Zero-Shot prompting presented in Chapter 4. For the conv. habitual activity the non-significant rating change for GPT-3.5-turbo is in the opposite direction (88.5 $\rightarrow$94.0; t(23) = -0.955, $p > .05$). For GPT-4 we see very high typicality ratings and almost no rating change in the presence of the conv. habitual utterance (97.9 $\rightarrow$98.8; t(23)=-0.582, $p > .05$).

9.4 Exp. 6.3: Frequency expressions and their associated ratings

For this final analysis, the goal was to investigate how consistent and robust the internal scale the models apply is, i.e. are similar explanations with identical frequency expressions associated with the same type of ratings. While this should generally be considered worthy of investigation, this analysis was, as previously stated, strongly inspired by the anecdotal experience of inconsistency when annotating and checking data for any of the previous experiments.

Specifically, the two objectives were a) confirming that clearly defined frequency expressions such as 'always' were assigned an adequate rating, and b) assessing how consistent the ratings associated with more subjective expressions such as 'frequently' and 'usually' are. For this analysis, all explanations collected in Exp. 1, Exp. 2 (regular analysis but not perturbation), Exp. 4 and Exp. 5 were considered. In total, that means there were 8390 data points for GPT-3.5-turbo, 1265 data points for GPT-4, 924 data points for Llama 3, and 637 data points for Mixtral[2]

Below, the process for data extraction and manual review is described, and how the intricacies of the task and the process do not allow for this analysis to be of quantitative nature. Furthermore, this cannot be considered a comprehensive qualitative analysis of how each of the models applies the scale, and instead a probe for such an analysis that should be extended in future work to have a solid qualitative basis.

9.4.1 Data extraction and methodology

For this analysis, the straight forward approach of writing a very basic script with hard-coded rules that extract all explanations containing certain frequency expressions was employed. After some initial tests and manual review, the following frequency expressions were chosen for objective (a): 'always' and 'every time'; and objective (b): 'usually', and 'frequently'.'[3]. Occurrences of 'occasionally' were also extracted, but results can only be reported for GPT-3.5-turbo. Generally, expression that would be associated with a lower rating rarely occurred, due to the models assigning very high ratings. The

[2]The number of data points available for each model varies due to several factors such as incomplete generations and repeated experiments (the latter specifically with GPT-3.5-turbo). However, I decided to include all available, correctly collected data points to maximize the data I can work with.

[3]Additionally, 'typically' and 'often' were considered, but for both an initial manual review revealed that they were not commonly employed by the model to describe the event frequency, and instead referred to the activity frequency, i.e. *often goes swimming* but not *often brings a swimsuit*

script also used a list of negation expressions[4] and noted whenever they co-occurred with the frequency expression within a phrase[5]. Following the same scheme, conditional phrases marked by an 'if' were not considered. While this method cannot claim to correctly extract only explanations that definitely employ the considered frequency expressions, this offered a solid basis for further review.

In the next step the data was manually checked to ensure that: 1) the extracted explanations uses the expression as expected, i.e. the expression is not negated; 2) refers to the correct event, i.e. refers to the frequency of the conv. habitual activity; and 3) was not overwritten by a different expression, e.g. an expression such as 'only occasionally' is followed up with something like 'but he has to always do this' as is the case when the atypicality is rejected; the expression isn't modified or linguistically hedged, e.g. 'frequently' occurring as 'less frequently than others' or 'frequently enough', which cannot be considered equivalent.

This process turned out rather complex, with many (pragmatic) nuances needing to be considered in order to identify explanations that are similar enough to reasonably compare them. The necessary steps for correctly and confidently pairing similar expressions and counting their absolute occurrences and associated ratings would be outside the scope of this work, hence this is not a quantitative analysis, and furthermore not a comprehensive qualitative analysis either.

Instead, this analysis simply aims to identify the range of ratings that credibly occurred with each of the analyzed frequency expressions, i.e. for every number rating that was assigned with an explanation that contained one of the expressions, it was checked that at least one occurrence of this expression-rating pair adhered to all four rules for 'correct' data points laid out above. Consequently, it was no longer deemed necessary to manually check every expression-rating pair that was extracted, and the correctness and appropriateness of a rating falling in the range was on occasion confirmed by checking a random sample of expression-rating pairs. I.e. according to the initial extracted data, GPT-3.5-turbo assigned a rating of 100 at the presence of 'usually' 300 times (out of 434 total occurrences); instead of checking all 300 explanations, a subset was viewed manually to confirm that this is a valid expression-rating pair. Analyzing 10 of the explanations at random revealed that all of them adhered to the four rules, obviously making 100 a valid part of the range, but not guaranteeing that all 300 data points are actually 'correct'.

As this analysis did not confirm every single data point, only the range, i.e. the minimum and the maximum of the ratings are reported. Further it is noted, whether one end of the range was underrepresented, i.e. while there were 300 potential data points for **usually : 100%**, there were only 8 data points for **usually : 70%**. 7 of these data points held up during manual inspection, with 'Going clothes shopping usually involves trying on clothes, so it is likely that Esther tries on clothes most of the time.' being excluded as it was determined that the quantifier associated with this rating is 'most of the time'. Therefore, the reported range 'usually' for GPT-3.5-turbo is 70 to 100 with a tendency for higher ratings.

[4] "not", "don't", "doesn't", "won't", "wouldn't"; negations in the past tense were excluded as a manual review found them to only occur in conditional phrases

[5] As this analysis aimed for simplicity, a phrase was simply defined as a subunit of a sentence constrained by punctuation marks, and a manual reviewed that this was sufficient for correctly capturing co-occurrence of the negation with the frequency expression.

9.4.2 Results

First of all, as expected, 'always' and 'every time' were associated with very high ratings of 100 or nearly 100 for all models except Llama 3. For Llama 3, there were some 80% ratings for always, with no further modification or hedging expression in the explanation.

For 'usually', we see a relatively consistent picture for Mixtral and GPT-4 with the following ranges: Mixtral - 85 to 100, with the majority of ratings falling in the middle of the range; and GPT-4 - 85 to 100, with 85 being only a single data point and the remaining points evenly spread across the range. The ranges for Llama 3 and GPT-3.5-turbo are relatively large, and skewed towards one end: GPT-3.5-turbo[6] - 70 to 100, with more ratings at the high end of the spectrum; Llama 3 - 80 to 100, with more ratings on the low end. Of course, it already seems less than optimal that for GPT-3.5-turbo 'usually' means 100% most of the time, but can also mean about 70%. This discrepancy becomes even more obvious when we see that almost the full range is reflected in responses for a singular stimulus:

"[...] he usually carries his cell phone on board with him while flying, as it is a common practice for many travelers.":75%

"[...] it can be assumed that he usually carries his cell phone on board with him when flying.":100%

For 'frequently', the picture is similarly clear for Mixtral (80 to 95, majority in the middle of range) and GPT-4 (80-95, evenly spread across the range). For GPT-3.5-turbo the range is again 70 to 100, with the majority falling in the middle of the range. Now, as expected based on anecdotal observations, Llama 3 also shows a large range, with most data points falling at 80 and up to 100, but also with valid ratings of 20, 60, 75 (though all single data points). And again we see the main range reflected in a singular stimulus.

"[...] she frequently uses scissors when cutting her hair.":80%

"[...] it's likely she frequently uses this method to cut her own hair.":100%

As a final expression, I want to briefly discuss 'occasionally'. For Llama 3 and Mixtral it only accounted for four and three data points respectively, hence not being quite indicative even for a qualitative analysis (the ranges were, however, 50 to 99, and 20 to 30, respectively). For GPT-4 there we no data points for 'occasionally' that held up during the manual inspection. For GPT-3.5-turbo, however, the observed range was 20 to 75, with an even distribution across the entire spectrum, making it the least consistently interpreted expression.

9.5 Discussion

The experiment using a Likert scale (Section 9.2) and the attempt at having the models' self calibrate (Section 9.3) have indicated a certain degree of robustness. Using a different and potentially better-suited scale has not resulted in the deriving of atypicality inferences, and has not shown strongly differing trends from previous experiments. This cannot only be seen as validation for the scale that was sued, but also as further support for the claim that LLMs do not draw atypicality inferences. Similarly, requesting self-calibration did not yield different results, further showing that the models reliably

[6]It should be noted that for GPT-3.5-turbo there is much larger amount of data points available as more experiments were performed, i.e. the same experiments were performed multiple times at the beginning of this work and only with this model.

do not derive atypicality inferences.

This probe analysis in Section 9.4 has, however, indicated that the scales are not applied entirely consistently, at least not by GPT-3.5-turbo and Llama 3. For Mixtral and GPT-4, however, the consistency was actually quite decent with smaller ranges and the majority of data points in the middle of the range, or evenly distributed across the range.

Now generally speaking, one could argue that humans could similarly struggle with adhering to a consistent scale, or rather that their understanding or interpretation of certain non-fixed frequency expressions may also not be consistent and vary depending on the individual or the context (Stump (1981), Pighin and Bonnefon (2011), also on relativity of quantity c.f. the popular meme: "Is four a lot?" "Depends on the context. Dollars? No. Murders? Yes.[7]")

Based on this, it is reasonably to assume that the scales are relatively consistent, or consistent enough, at least with Mixtral and GPT-4. For GPT-3.5-turbo and Llama 3 the consistency is not as promising, which is also not too surprising given that these two models generally also had higher rates of weird or nonsensical generations. In addition to the previously discussed limitations of this analysis, it should also be noted that this analysis has been limited by the type of responses that the models give. As they tend to assign very high typicality ratings and provide explanations that do not show the derivation of atypicality inferences, quantifiers that would typically be associated with lower ratings and the derivation of atypicality inferences could not be assessed.

Overall it can be concluded that the results, especially those more promising ones obtained with GPT-4 (and maybe Mixtral) are robustness enough to reasonably assess the models' abilities in regard to drawing atypicality inferences.

[7]Source: https://knowyourmeme.com/photos/1504468-depends-on-the-context

Chapter 10
Conclusions

The work of Kravtchenko and Demberg (2022) and Ryzhova et al. (2023) has shown that humans can deal with informational redundancies based on script knowledge by inferring the atypicality of a conv. habitual activity in the given context. In humans, this derivation of the pragmatic inferences known as atypicality inferences cannot be observed uniformly – variability exists at the level of activities (some activities exhibit a larger rate of atypicality inferences than others) and at the level of participants. For the latter, Ryzhova et al. (2023) showed that the ability to draw atypicality inferences is correlated with reasoning ability.

Recently, LLMs have been exhibiting reasoning and pragmatic abilities that are sometimes considered to be on par with or exceeding human abilities. Consequently, this work considered that LLMs should potentially be able to derive atypicality inferences in a human-like manner. The most straightforward adaptation of the human experiments for testing atypicality inferences for LLMs, a Zero-Shot approach presented in Chapter 4, did however not yield any success. Without further guidance, GPT-3.5-turbo, GPT-4, Llama 3 and Mixtral did not only not exhibit atypicality inferences, they also frequently reasoned the consequent typicality in a non-human-like manner, i.e. by arguing that the mentioning of the conv. habitual activity in the provided context actually reinforces the fact that the activity always takes place.

While these initial findings were not promising, it had to be considered that good performance on reasoning tasks exhibited by LLMs does not usually occur in a Zero-Shot setting, and that a different prompting methods might yield a higher success rate. Consequently, this work employed further prompting methods in hopes of facilitating the derivation of atypicality inferences in LLMs. The first method, Few-Shot prompting (Chapter 5), provided the models with exemplars that showed the derivation of atypicality inferences, in hopes that the models could not only emulate but successfully adapt and apply the reasoning process these exemplars showed. However, Llama 3 and Mixtral did not show an improved performance in this setting, apparently not even shallowly copying the provided exemplars. GPT-3.5-turbo did show the desired behavior but failed a sanity test, showing that it did not differentiate between the informationally redundant conv. habitual utterance and the non-habitual utterance, i.e. an utterance about *getting apples* led the model to decrease the typicality rating of *paying the cashier*. This clearly

indicated that the model did not understand and apply the necessary reasoning for deriving atypicality inferences, and instead just matched the pattern from the exemplar. While GPT-4's performance seemed initially more promising, a consequent perturbation analysis led the model to also fail the same sanity check. This showed that, while more advanced than GPT-3.5-turbo, GPT-4 also only performed shallow pattern matching.

Before delving into additional prompting methods, this thesis performed an analysis (Chapter 6) into the potential points of failure that followed the proposed steps for deriving atypicality inferences: 1) identify the redundancy based on script knowledge; 2) realize that redundancy is infelicitous, as it violates conversational norms; 3) infer activity atypicality; 4) explicitly accommodate atypicality in situational context (Ryzhova et al., 2023). This analysis did not only offer valuable insights into the models' behaviors, but also served in determining how to proceed with the additional prompting methods. Firstly, the tested LLMs were found to have access to the necessary script knowledge, and are able to use this knowledge to identify the redundancy. However, while they *can* identify the redundancy without further input, it was found that this identification only occurs reliably and consistently across all models when the existence of a redundancy is presupposed. Hence, while not necessary 'failing' at identifying the redundancies, presupposing them could facilitate their recognition and, consequently, also the derivation of inferences. Further, the analysis found the models capable of performing the fourth step of accommodating the atypicality, with Llama 3 showing the best performance, hence indicating that this is also not where the models fail. For the second and third step, however, the analysis found that the models did not usually identify the redundancy as infelicitous or problematic, and instead always assumed that it served a purpose. More specifically, they would often propose conversational or interpersonal purposes for uttering the redundancy, ranging from simple emphasis or small-talk to attempts at humor. In other words, even when presupposing that the redundancy was infelicitous and consequently should be 'repaired' by making a pragmatic inference, the models rarely considered atypicality for that purpose. The fact that targeting these steps in isolation, and presupposing the necessary inferences / assumptions from the previous steps was not sufficient for getting the models to perform these steps, strongly indicates that an inability to react to this subtle maxim violation introduced by the informational redundancy in a human-like manner is what keeps the models from deriving atypicality inferences. However, it has to be remarked that this analysis, that was aimed specifically at performing each of the steps, does not reliably show whether or not a given model is actually able to perform this step unprompted, or in a different context.

Consequently, this work also presented an adaption of Liu et al. (2022)'s Generated Knowledge Prompting, where the models externalized their script knowledge prior to performing the task, in hopes that this externalization could facilitate the necessary reasoning. However, this experiment only further confirmed the previous findings and assumptions: 1) The models do have the necessary script knowledge and can appropriately externalize it, and 2) accessing this script knowledge is not enough to facilitate the reasoning process for deriving atypicality inferences. In fact these results barely differed from the results obtained from Zero-Shot prompting; it is however unclear if that is the case because the models' applied the script knowledge in the same way regardless if it was externalized or not, or if this stems from the script knowledge being not a deciding factor in the model behavior, as they fail at a later point in the derivation process.

The follow-up experiment in Chapter 8 based on Three-hop Reasoning (Fei et al., 2023) then not only further attempted to elicit atypicality inferences in the models, but also offered more insights into the behavior of each model. In this experiment, the models were

guided along the reasoning process, which was broken into sub-tasks that the models then performed step by step. A total of three variants of step-by-step instructions, i.e. multi-hop variants, were tested, with each increasing the degree to which it presupposed necessary information and consequently pushed the models towards deriving atypicality inferences. By solely looking at the final results, this approach was only a partial success, as only the final variant (Section 8.4) with the most guidance was able to elicit atypicality inferences and only in Llama 3. Further, while significant, these inferences were only derived for a relatively small number of stimuli. However, beyond the final results, this experiment provided model responses for the intermediate reasoning steps, consequently to some degree removing the model performance on these steps from the isolation under which they were viewed in Chapter 6. And notably, in almost all cases when Llama 3 derived the atypicality inferences, it also performed 'well' or 'correctly' in the intermediate steps. But while this can be taken as evidence that for Llama 3, these reasoning steps actually lead to the derivation of atypicality inferences, these finding come with three major caveats: 1) actual derivation and the desired performance rarely happened and can therefore not be considered consistent behavior; 2) the derivation in ratings stemmed from extreme typicality drops that cannot be considered human-like, i.e. the model assessing that no-one ever brings swimsuits to public pools; and 3) the derivation according to explanations actually happened at a similar rate in previous hop variants, and in those previous variants they did not align with the desired behavior in intermediate steps, e.g. in the second variant (Section 8.3) derivations did not occur in cases where the non-presupposed over-informativeness was successfully identified.

Notably, these finding also do not generalize across models; from the collected data, it cannot be concluded that the other models do not derive atypicality inferences because they do not perform well on the intermediate reasoning steps. Recall that GPT-3.5-turbo actually did perform quite similar to Llama 3 on the intermediate steps of the third variant without arriving at atypicality inferences. For Mixtral introducing these intermediate reasoning steps actually led to a really good performance on the earlier defined point of failure, i.e. it led Mixtral to consistently view the informational redundancy as a maxim violation. However, instead of facilitating the further derivation process, the model actually no longer exhibited any tendency for accommodating or repairing this over-informativeness, consequently not considering atypicality (or even conversational or interpersonal purposes) and not deriving any atypicality inferences at all. Finally, as previously discussed, GPT-4 also exhibited a somewhat strange behavior where it would not do well on accepting the maxim violation, then would nonetheless successfully accommodate it, but ultimately did not take this accommodation and potential atypicality into consideration when providing a final assessment.

So what does this leave us with? Overall, this thesis has provided a solid foundation for assessing that the tested LLMs do not derive atypicality inferences in a human-like manner, even when they are guided through the reasoning process. Further, occasional derivations of atypicality inferences in ratings or reasoning do not occur reliably or consistently, consequently serving as additional evidence that, at this time, the derivation of this specific pragmatic inference is outside the scope of the LLMs' pragmatic and reasoning abilities. This thesis has provided strong evidence that this failure does stem from problems with identifying over-informativeness as a maxim violation that needs to be accommodated, and an unwillingness to consider atypicality as the reason that an informational redundancy was produced. This 'problem' is so ingrained in the GPT-Models, that even presupposing the over-informativeness and maxim violation is not enough to influence their performance. Their learned world knowledge and understanding of common human behavior appears solid and immutable to a degree where it leaves no

room for this subtle pragmatic nuance. While I can only speculate about the reasons for this (especially with OpenAI offering only limited insights into the details of their models and training procedures) it seems likely that this is an intentionally enforced behavior, for example facilitated by Reinforcement-learning from human feedback, to limit the models' generation of hallucinated facts and misleading / wrong information. In other words, these commercial models may contain safeguards against suddenly claiming that public pool usually offer swimsuit rentals and lax dress-codes, as stated by Llama 3. Similarly, the fact that the GPT-models tend to provide very general lists of potential explanations when asked to explain a specific behavior or situation may be another result of measures aimed at preventing the models from committing to the 'wrong' response.

On the other hand, Llama 3 does not exhibit any of these tendencies, consequently committing to more specific, less careful explanations, which in turn reads more like human-like behavior. But Llama 3 also ends up generating the highest rate of nonsense and 'incorrect' or misleading assessments. Comparing the behavior of Mixtral, it almost seems to offer a middle ground, where the generated responses seem less careful and restrained, and consequently more clear and human-like than those of the GPT-models, but with a more firmly defined world knowledge and less of a tendency to generate nonsense than Llama 3. As a final interesting difference between the four models, I also want to return to the apparent failure of Llama 3 and Mixtral when emulating exemplars in the Few-Shot setting (Chapter 5). While it was ultimately found that, for the specific task presented in this work, the GPT-model also only perform template matching, it was notable that the other two models did not even do that. Especially interesting though is the fact that Mixtral did consider the provided exemplars to some degree when providing their typicality judgements, as it rejected the modeled atypicality for more than a quarter of the stimuli, and continued to do so even during the perturbation analysis. Without exemplars modeling atypicality, i.e. in the Zero-Shot and Generated Knowledge Prompting experiments (Section 4.2, Section 7.2), this rejection of atypicality did not happen, consequently making this behavior a direct reaction to the modeled exemplars. Interestingly, GPT-4 also did reject atypicality at the same rate as Mixtral – in the regular Few-Shot experiment (Section 5.2) and despite overall deriving atypicality inferences – indicating that GPT-4 also is capable of considering the content of the exemplars without copying them.

Despite these differences, the models do share a lot of common behavior in regard to not deriving atypicality inferences, namely in finding the informational redundancy fairly acceptable and leaning towards interpersonal and conversational purposes for expressing such a redundancy. Crucially, this assessment of the model behavior is in no way comprehensive, and does not wish to claim superiority of performance for any of these models. While I have discussed the trends and tendencies that were observed, it also has to be noted that at times the picture remains quite unclear, and the models' responses either do not follow a pattern or suddenly deviate from a previously observed pattern / behavior. In regard to the models' performances, I do also want to briefly return to the question of faithfulness and robustness that was posed in the previous Chapter 9. The first two experiments in this chapter – using a Likert scale and a calibration method – did confirm that the findings of this work can be considered a robust representation of the models' abilities at this point in time. Assessing the faithfulness of the explanations of the model did however prove to be quite difficult, and ultimately this work could only offer a probe analysis into the internal consistency of these models. This probe analysis did show that GPT-4 and Mixtral exhibited a fairly promising internal consistency in regard to judging typicality / frequency, while GPT-3.5-turbo and Llama 3 seemed less consistent.

Ultimately, it can be said that this thesis offers a solid initial picture of the tested LLMs' inability to derive atypicality inferences and insights into underlying behavior. It is however only a small contribution towards understanding the pragmatic and reasoning abilities of LLMs and, with its strengths and limitations, offers up many avenues for further research that are discussed in the following Chapter 11.

Chapter 11
Limitations and Future Work

One limitation from the NLP perspective of this study is that the size of the dataset is small (only 24 stories) and only in English. This is a common limitation of psycholinguistic studies due to the costs of human experiments, and it is furthermore a limitation of investigating atypicality inferences, where the number of suitable event sequences and target activities from script knowledge is limited, i.e. humans (and by extension LLMs) don't have unlimited script knowledge, and it may also vary based on a variety of factors such as cultural differences.

While, during the initial Zero-Shot experiments, it was considered to collect a distribution of data points from each model, it was ultimately decided to treat each model as one participant that provides responses once, as it has been previously done in other research with LLMS (cf. Han et al. (2024)). Consequently, there is a low number of total data points for each model, which of course can be a problem in regard to the statistical power.

This work has aimed to investigate promising prompting methods aimed at reasoning abilities in addition to the more classic Zero-, and Few-Shot prompting, but it did of course not cover the whole spectrum of potential prompting methods. While it has been shown that the inferences are not derived in a human like manner without further input, and not with Generated Knowledge Prompting, or the adaptation of Three-hop Reasoning applied here, it is therefore possible that the models could perform this task when prompted in a different way. Hence, it should be considered to repeat these experiments with additional, and potentially new prompting methods.

Another limitation is of course in the selection of models, as it does not cover the full range of different available architectures, due to not only the amount of different models, but also the frequency at which they are released. For that reason, I also did not include the newest OpenAI model GPT-4o.

A major limitation stems from only analyzing the generated tokens and not their probabilities, as this is not supported by the OpenAI API. Furthermore, our efforts at testing a Likert scale in addition to 0% to 100% scale and requesting self-calibration (Section 9.2, Section 9.3) from the model through considering multiple answers cannot fully mitigate the potential problems of having the models output concrete values, and within the limited data it was not possible to satisfyingly assess how consistently the model can

actually adhere to any given scale. In that same vein, the faithfulness of externalized model reasoning has been previously questioned, and again the degree of faithfulness exhibited in these experiments cannot be reliably assessed. Instead, this work provided a probe analysis into the faithfulness and internal-consistency of the models (Section 9.4) that should be extended in future work, i.e. by developing a more detailed and complete pragmatic framework for assessing the models' use of quantifiers and frequency expression, or by collecting more data that also represents a broader range of potential model behavior. But while the potential for further research is large, I believe that the combination of concrete values and explanations obtained, paired with our qualitative analysis of the performance on different steps provide a solid initial picture of the models abilities in terms of deriving atypicality inferences.

Finally, this work has treated each model as a black box, only assessing their abilities through prompting, and only with a limited number of manually engineered prompts. Further research aimed more at the models' internal mechanisms, i.e. by probing and investigating the layer-wise capabilities, would be recommendable.

Bibliography

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 4 2022. URL `https://arxiv.org/abs/2204.05862v1`.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400, 2023.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Gordon H. Bower, John B. Black, and Terrence J. Turner. Scripts in memory for text. *Cognitive Psychology*, 11:177–220, 4 1979. ISSN 0010-0285. doi: 10.1016/0010-0285(79)90009-4.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Jiuhai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*, 2023.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Catherine Davies and Napoleon Katsos. Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, 120(8):1956–1972, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Association for Computational Linguistics (NAACL)*, 1, 2019. URL https://github.com/tensorflow/tensor2tensor.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2:1171–1182, 2023. ISSN 0736587X. doi: 10.18653/v1/2023.acl-short.101. URL https://beta.openai.com/docs/models/gpt-3.

Joana Garmendia. Lies we don't say: Figurative language, commitment, and deniability. *Journal of Pragmatics*, 218:183–194, 2023. doi: 10.1016/j.pragma.2023.11.003.

Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155, 1 2024.

Pengfei Hong, Deepanway Ghosal, Navonil Majumder, Somak Aditya, Rada Mihalcea, and Soujanya Poria. Stuck in the quicksand of numeracy, far from agi summit: Evaluating llms' mathematical competency through ontology-guided perturbations. *arXiv preprint arXiv:2401.09395*, 2024.

Xudong Hong, Margarita Ryzhova, Daniel Adrian Biondi, and Vera Demberg. Do large language models and humans have similar behaviors in causal inference with script knowledge? *arXiv preprint arXiv:2311.07311*, 2023.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*, 2022.

Jie Huang, Kevin Chen, and Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022a.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. pages 9118–9147, 2022b.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. Promptbert: Improving bert sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid Google Research, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 12 2022.

Ekaterina Kravtchenko and Vera Demberg. Informationally redundant utterances elicit pragmatic inferences. *Cognition*, 225:105159, 8 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam Mccandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

Fangru Lin, Daniel Altshuler, and Janet B Pierrehumbert. Probing large language models for scalar adjective lexical semantics and scalar diversity pragmatics. *arXiv preprint arXiv:2404.03301*, 2024.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1: 3154–3169, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL `https://github.com/pytorch/fairseq`.

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

Emmanuele La Malfa, Matthew Wicker, and Marta Kwiatkowska. Emergent linguistic structures in neural networks are fragile. *arXiv preprint arXiv:2210.17406*, 2023. URL `https://github.com/EmanueleLM/emergent-linguistic-structures`.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint arXiv:2109.12951*, 2021.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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 2463–2473, 2019.

Stefania Pighin and Jean-François Bonnefon. Facework and uncertain reasoning in health communication. *Patient Education and Counseling*, 85(2):169–172, 2011.

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. Pragmatic implicature processing in chatgpt. *PsyArXiv*, 2023. URL `https://doi.org/10.31234/osf.io/qtbh9`.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*. Association for Computational Linguistics, 2019.

Laria Reynolds and Kyle Mcdonell. Prompt programming for large language models: Beyond the few-shot paradigm. *Conference on Human Factors in Computing Systems - Proceedings*, 5 2021. doi: 10.1145/3411763.3451760. URL https://dl.acm.org/doi/10.1145/3411763.3451760.

Margarita Ryzhova. Processing cost and individual differences in the derivation of atypicality inferences. Manuscript in preparation, 2024.

Margarita Ryzhova, Alexandra Mayn, and Vera Demberg. What inferences do people actually make upon encountering informationally redundant utterances? an individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45, 2023.

Abhilasha Sancheti and Rachel Rudinger. What do large language models learn about scripts? **SEM 2022 - 11th Joint Conference on Lexical and Computational Semantics, Proceedings of the Conference*, pages 1–11, 2022.

R Schaeffer, K Pistunova, and S Khanna. Invalid logic, equivalent gains: The bizarreness of reasoning in language model prompting. *arXiv preprint arXiv:2307.10573*, 2023. URL https://arxiv.org/abs/2307.10573.

Roger C. Schank and Robert P. Abelson. Scripts, plans and knowledge. *Proceedings of the 4th International Joint Conference on Artificial Intelligence. IJCAI*, 1, 1975.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*, 2024.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.

Gregory T Stump. The interpretation of frequency adjectives. *Linguistics and Philosophy*, 4:221–257, 1981.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, Ellie Pavlick, and Google AI Language. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019. URL https://github.com/jsalt18-sentence-repl/jiant.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330.

Miles Turpin, Julian Michael, and Ethan Perez. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388.*, 2023. URL https://arxiv.org/abs/2305.04388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:2717–2739, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3494–3501, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1556.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour 2023 7:9*, 7:1526–1541, 7 2023. URL https://www.nature.com/articles/s41562-023-01659-w.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2 2023. URL https://arxiv.org/abs/2302.11382v1.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Deqing Yang, and Yanghua Xiao. Distilling script knowledge from large language models for constrained language planning. *arXiv preprint arXiv:2305.05252*, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Alex Smola, † Shanghai, and Jiao Tong. Automatic chain of thought prompting in large language models. *arxiv.orgZ Zhang, A Zhang, M Li, A SmolaarXiv preprint arXiv:2210.03493, 2022•arxiv.org*, 2022. URL https://arxiv.org/abs/2210.03493.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

Rolf A Zwaan, Joseph P Magliano, and Arthur C Graesser. Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition*, 21(2):386, 1995.

Appendix A
Additional Prompts: Exp. 3

Below the alternative questions/wordings for the steps are presented and a brief discussion of why they were less efficient than their counterparts is provided.

A.1 Step 1:

For this step, the following alternative questions were tested:

- A_Q1: Is any part of the direct speech superfluous or unnecessary?
- A_Q2: Does the context (C) contain any redundancies?

With A_Q1 the word redundancy was replaced, as it was considered that it might be too specialized, i.e. not be the word a layman would choose to describe the phenomenon. For the most part, this did however perform on par with or slightly worse than Q1 reported in Chapter 6, ultimately showing that this distinction did not matter to the models. With A_Q2 the goal was trying an even more open-ended approach by not restricting the potential redundancies to the direct speech. This did, however, unsurprisingly yield even fewer identifications of the desired informational redundancy.

A.2 Step 2:

As explained before, identifying the models' ability to perform this step was challenging through a set of questions was challenging due to its subtlety, and since humans might also not verbalize their implicit understanding of the conversational norm that is violated by redundancies (Maxim of Quantity). Ultimately, we used the following questions to gauge a more general understanding of the models' awareness of conversational norms:

- A_Q3: The second sentence in the direct speech provides redundant information, since the action it talks about is already implied in the first sentence. Do you think this was an acceptable utterance?

- A_Q4: The direct speech contains redundant information. Is providing redundant information a good and efficient way of communication?
- A_Q5: The direct speech contains redundant information. Do you see any issue with that?

For A_Q3 the utterance was mostly deemed acceptable by GPT-3.5-turbo, GPT-4 and Mixtral. When reasoning was provided in the model response, for GPT-3.5-turbo and GPT-4 it would be very general, usually suggesting that the redundancy served the purpose of emphasizing or expressed general noteworthiness. Mixtral also frequently suggested that the purpose might be emphasis or clarification, while relating the potential purposes more closely to the content of the story. Llama 3, on the other hand, found the utterance mostly not acceptable, sometimes reasoning that it may serve as emphasis or to provide nuance, but mostly classifying them as unnecessary or even awkward. For A_Q4 GPT-3.5-turbo answered no 24 times without elaborating further. GPT-4, Mixtral and Llama 3 also agreed that it is not acceptable and elaborated why (i.e. confusing, waste of time), but the majority of time it was then also stated that there still might be good reasons (i.e. emphasizing, clarification). For A_Q5, GPT-3.5-turbo saw no issue for most items, and the remaining times it said there was an issue with redundancy, though usually not the informational redundancy we were investigating but one from the broader context (i.e. "Don mentions that he took a train with Jane, which is already implied by the fact that he saw Jane at the subway station and they took the train together", which is arguably not a redundancy because the character he tells this to does not know that he saw her and that they took a train). GPT-4 generally saw no issue, occasionally stating the (correct informational) redundancy and for each item elaborating reasons the redundancy occurred. These reasons are however mostly very general and broad (i.e. emphasis, enthusiasm, creating a relaxed atmosphere, establishing a connection). Finally, with A_Q5 Llama 3 did actually identify the informational redundancy this analysis aimed at for the majority of the stimuli, and did claim that it was an issue.

A.3 Step 3:

Find below additional questions we tested for this step:

- A_Q6: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can be unnecessary and inefficient for communication. Why was the redundant utterance made?
- A_Q7: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can be unnecessary and inefficient for communication. Consider only what you can tell about the people from the provided context (C) and tell me definitively: Why did the speaker still choose to express the redundant information in this specific situation?
- A_Q8: The second sentence in the direct speech conveys seemingly redundant information. Providing redundant information can be unnecessary and inefficient for communication. However, the speaker made the effort of conveying this information. Since they have no reason to be inefficient, this information must actually be new or important. What new or relevant information can you infer from the second sentence?

A_Q6 and A_Q7 resulted in very general responses from GPT-3.5-turbo that covered the same potential reasons for redundancy that have been stated in previous steps. GPT-4 ,Llama 3 and Mixtral also provided similar reason but did a better job of applying them to the specific scenario rather than keeping them general. Notably, atypicality was not among the reasons that Llama 3 and Mixtral came up with. For GPT-4, atypicality was sometimes stated as a possibility, but more frequently the model proposed attempts at humor or conversational purposes. For A_Q8, GPT-3.5-turbo defaulted to just stating the exact contents of the sentence, while GPT-4 performed slightly worse than with the for each item adjusted Q4 reported in the paper (i.e. it gave appropriate reasons, but not as specific to the item content, and fewer explanations pointing towards atypicality). Llama 3 and Mixtralunsurprisingly showed a similar performance to the other questions as the model has less of a tendency to generalize.

A.4 Step 4:

For step 4 there were no further prompt formulations tested, as the best performing question from step 3 was directly adapted by inserting the desired atypicality answer and then adding a simple question to elicit alternative behavior.

Appendix B
GPT-4-turbo Results

Here the main findings from experiments conducted with GPT-4-turbo are briefly presented[1]. After these experiments it was concluded that GPT-4-turbo, which was introduced as a faster and slightly eloquent version of GPT-4, indeed only outperforms GPT-4 on these measures, but with reasoning abilities that do not exceed GPT-4, and overall performance closer to that of GPT-3.5-turbo. In that light, the model was not further pursued.

B.1 Exp. 1: Zero-Shot Prompting

As with the other models, there was almost no belief change observed for the conv. habitual utterance from the baseline to the conv. habitual utterance condition (95.16 →96.30; t(23) = -1.64, $p > .05$). The sanity check comparing the baseline against the non-habitual utterance condition revealed no rating change in line with human behavior and the other models (95.16 →95.56; t(23) = -0.637, $p > .05$)).

The manipulation of the context to state atypical behavior reduced the baseline typicality ratings (again with a very high standard deviation, i.e. the effect varying greatly across stimuli). The belief change after encountering redundancies was minimal GPT-4-turbo (36.85 →37.29).

Of the explanations provided by the model, ~48% were classified as *reinforced utterance*, and ~58% of responses as *script knowledge*. The categories of *hallucinated facts* and *incorrect reasoning* no longer played a role.

[1] As with the other GPT-Models default parameter temperature t=1, presence_penalty = 0, and top_p = 1 were used.

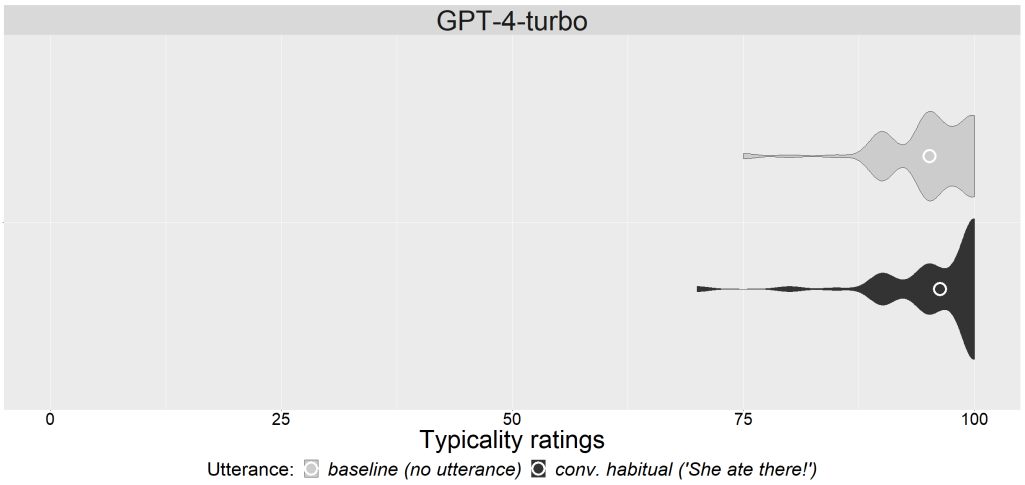

Figure B.1: Zero-Shot, habitual activity analysis for GPT-4-turbo

B.2 Exp. 3: Analysing the steps of reasoning process

B.2.1 Step 1

For Q1, GPT-4-turbo identified the informational redundancy 16 times and identified the reverse redundancy twice. Additionally, it saw no redundancy once, and gave unclear, inconclusive or not relevant redundancies the remaining 5 times. For Q2, GPT-4-turbo correctly identified the informational redundancy 21 times, the reverse once, and the remaining 2 times the identified redundancy was grammatical. Overall, the model therefore performed very similarly but slightly worse than its predecessor GPT-4

B.2.2 Step 2

For this step, GPT-4-turbo again behaved similarly to GPT-4, being aware that redundancies were "not typically efficient" in communication, and actually most of the time elaborating why. Regardless, it was also aware of reasons why it could be acceptable (again, mostly emphasizing or clarification). Consequently, the model saw no issue with redundancies, occasionally stating the (correct informational) redundancy and for each item elaborating reasons the redundancy occurred. These reasons were, in-line with previous and GPT-4 behavior, mostly very general and broad.

B.2.3 Step 3

For Q3, GPT-4-turbo correctly identified the redundancy for all stimuli, and, similarly to GPT-4, viewed emphasis as the most common explanation (generally included in a broad lists of potential reasons).

With the stimulus-specific Q4, GPT-4-turbo included atypicality 11 times, sometimes providing a reasonable alternative of what is done instead. In all other cases, again much like the other models, it gave reasonable explanations regarding the purpose of redundancy which, however, did not indicate an atypicality inference.

B.2.4 Step 4

In this step with the stimulus-specific Q5, GPT-4-turbo comes up with a sensible alternative explanation for 22 stimuli, and for the same stimuli as GPT-4 it only provides the non-elaborative "not that" answer.

B.3 Discussion

As seen here,GPT-4-turbo, much like the models discussed in the main body of this work, did not draw atypicality inferences. The decomposing of the steps showed that the model also has script knowledge, and it actually outperformed GPT-4 in specifically accommodating atypicality inferences, which did however not transfer to a better performance in the inferences task (as has also been observed for Llama 3).