Prompting Meaning: A Hermeneutic Approach to Optimising Prompt Engineering with ChatGPT

Leah Henrickson[1]
School of Communication and Arts, University of Queensland, Brisbane, Australia
l.henrickson@uq.edu.au
ORCID: 0000-0001-8008-2373

Albert Meroño-Peñuela[1]
Department of Informatics, King's College London, London, United Kingdom
albert.merono@kcl.ac.uk
ORCID: 0000-0003-4646-5842

Abstract

Recent advances in natural language generation (NLG), such as public accessibility to ChatGPT, have sparked polarised debates about the societal impact of this technology. Popular discourse tends towards either overoptimistic hype that touts the radically transformative potentials of these systems or pessimistic critique of their technical limitations and general 'stupidity'. Surprisingly, these debates have largely overlooked the exegetical capacities of these systems, which for many users seem to be producing meaningful texts. In this paper, we take an interdisciplinary approach that combines hermeneutics —the study of meaning and interpretation— with prompt engineering —task descriptions embedded in input to NLG systems— to study the extent to which a specific NLG system, ChatGPT, produces texts of hermeneutic value. We design prompts with the goal of optimising hermeneuticity rather than mere factual accuracy, and apply them in four different use cases combining humans and ChatGPT as readers and writers. In most cases, ChatGPT produces readable texts that respond clearly to our requests. However, increasing the specificity of prompts' task descriptions leads to texts with intensified neutrality, indicating that ChatGPT's optimisation for factual accuracy may actually be detrimental to the hermeneuticity of its output.

Keywords

hermeneutics, prompt engineering, natural language processing, natural language generation, large language models, ChatGPT

Acknowledgements

We are grateful to Sheffield's Digital Humanities Institute, which facilitated the authors' meeting and collaboration. We would also like to thank fellow participants of the 2022 Digital Humanities Congress and the 2023 King's College London Creative AI Symposium for their feedback on earlier and alternative versions of this work.

[1] These authors contributed equally to this work.

Statements and Declarations

Partial financial support was received from the authors' academic appointments at the University of Leeds and King's College London. The authors declare they have no financial interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data availability statement

There is no external data associated with this paper.

Introduction

Advances in machine learning techniques and the high availability of data and compute power have given rise to a new generation of artificial intelligence (AI) and natural language processing (NLP) approaches, which have achieved unprecedented performance in tasks like question answering and natural language generation (NLG). In fact, modern NLG systems such as ChatGPT (OpenAI 2022) —built on and also known as pre-trained large language models (LLMs)— can create texts so readable that they are capable of deceiving readers into thinking they have been written by humans. The quality of these texts raises important questions about hermeneutics —the study of meaning-making processes, which has been primarily applied to religious and literary texts— and the perceived social contract between readers and writers, imagined by readers. This contract has been called the hermeneutic contract, and is based on readers' assumption that an author has produced a text that is both meaningful and relevant (Henrickson 2021, 4).

Natural language generation (NLG) refers to the mechanised production of texts in everyday human languages. NLG is a subfield of NLP, which more broadly refers to computational methods of language analysis for an increasingly wide range of purposes. We believe that we may use hermeneutics to make sense of NLG output. We have already written at length about the relevance of hermeneutics to NLG systems (Henrickson and Meroño-Peñuela 2022), arguing for explicit discussions about how we come to make meaning from NLG system output, paying particular attention to expectations of authorial intention and what it means to be a reader. Our previous work about the hermeneutics of NLG systems stems from the concept of the hermeneutic contract, which is the perceived figurative meeting place between an author and a reader. This meeting place is ultimately determined by the reader, but is often influenced by a reader's expectation of authorial agency that is informed by lived experience (Henrickson 2021, 4-5). New NLG systems that facilitate instantaneous generation of output based on user input may spur readers to establish their perceived 'meeting places' with these systems depending on the readers' conversational needs; based on patterns identified in its training, the system will respond to user input in accordance with what it 'believes' is the most optimal answer. It matters little who - or what - the perceived author of a computer-generated text actually is - the reader finds meaning in text that is semantically comprehensible and presentationally familiar.

A piece of user input to which a NLG system responds to is called a prompt. In this paper, we reflect upon prompt engineering: conscious user efforts to adjust prompts for optimal computational response. More precisely, prompt engineering involves embedding the description of the task we want the system to perform into our input, rather than relying on the system to decipher the potentially unclear language that we may use when speaking with other humans (Liu et al. 2021). Prompt engineering has been the focus of much recent academic review, but has also increasingly been the source of public entertainment through 'prompt battles' wherein participants compete to produce the "most surprising, disturbing or beautiful images from the latest spaces of DALL·E, Stable Diffusion, Midjourney, Craiyon, etc." at in-person events (Schmidt and Schmieg 2022; https://promptbattle.com) or "compete to create the most absurd and awesome images with AI" via online app (No author n.d.; https://promptbattle.xyz). Others have used prompt engineering to hijack these AI systems, gaining access into the instructions that the systems have been given to inform output (Edwards 2023). Prompt engineering has been deemed as potentially "the most important job skill of this century" (Warzel 2023). The value of prompt engineering, as we show in this paper, lies in its use as a tool for readers to steer NLG systems towards answers of higher quality, better usability, or improved user satisfaction; prompts direct how a system searches for and generates answers in a vast information space.

AI and NLG systems are often framed as though they are superior to humans, or a serious risk to human activity and/or livelihood. Matthew Kirschenbaum (2023), for one, has —albeit somewhat tongue-in-cheek— suggested the potential for a 'textpocalypse' wherein human writers are rendered obsolete. Other authors (e.g. Sadowski 2018 and 2020), however, have labelled current AI systems 'Potemkin AI', in reference to Grigory Potemkin's erection of fake settlements during Empress Catherine's 1787 tour through Crimea, which were intended to convince the Empress that the area was richer and more fruitful than it was. Potemkin AI depends on illusions of intelligence rooted in human labour (e.g. scraping text from Web pages, cleaning data, assembling datasets, generating examples, providing feedback to models) that is often hidden and low-paid. In the general narratives that permeate society, AI is commonly presented as somehow immune to human prejudice. However, such presentation overlooks the ample human intervention in all of AI's many layers. Human intervention often focuses on optimisation of systems for benchmark performance or financial profit, with these optimisation efforts ultimately advancing AI in narrowly-defined tasks. Yet cultural conversations about AI frequently make broad and hyperbolic generalisations about the technology that promote overestimations of system capacity and influence, either positively or negatively. We believe that a balance between optimistic hype and pessimistic critique about NLG can be achieved through hermeneutic analysis. While there has as yet been limited consideration of NLG from a hermeneutics perspective, attention to how we make meaning from computer-generated texts is arguably more important than ever given the ongoing proliferation of readable system output and integration of these systems into workaday technologies (e.g. search engines like Microsoft Bing).

In this paper, we use prompt engineering to challenge the polaristic attitudes towards NLG systems through the lens of hermeneutics, using ChatGPT as an example given to its widely-regarded capacity for generating texts and its rapid and mainstream adoption. Underlying our work is scepticism of the concept of 'optimisation' and its vague, inconsistent, and often uncritical use in discourse surrounding AI. We hypothesise that, through prompt engineering, users may facilitate the generation of output with greater hermeneuticity — that is, hermeneutic quality in terms of fitness for use — that both speaks to their unique perspectives and encourages them to critically reflect on those perspectives. In doing so, users 'optimise' their uses of these systems. This hypothesis is supported by our previous analyses of hermeneutics as pertaining to computer-generated texts. Readers, we have argued, rely on a perceived 'hermeneutic contract' between readers and authors; through evocation of this contract, readers can identify meaning in computer-generated texts whether or not there is evidence of authorial intention or system sentience (Henrickson and Meroño-Peñuela 2022). Even if there were truly nothing behind ChatGPT's textual façade, readers would continue to imagine that there were. This imagination is indicated in the ways in which ChatGPT and its predecessors have been portrayed by journalists as potential futures risks to human writers given the readability of their output. At the time of writing, many of these portrayals respond to higher education contexts. "The College Essay Is Dead", declares the title of one article (Marche 2022). "ChatGPT: A Threat To Higher Education?" asks another (Wingard 2023). The answer to this latter question is, at least in our view, a clear no as long as readers seeking meaning are in the loop.

ChatGPT (Chat Generative Pre-Trained Transformer) is a chatbot based on LLMs, developed by OpenAI and made publicly accessible on 30 November 2022. LLMs have been the cornerstone of NLP research since the mids 2010s, and have recently shown landmark technical improvements for three reasons: the increasing availability of compute to train them; the increasing availability of high-quality, highly-curated textual data and corpora; and the development of new computational techniques. The most important of these techniques belong to the so-called deep learning family, a branch of machine learning algorithms that makes use of complex artificial neural networks to automatically 'learn' programs (or models) from data. Progress in these architectures, such as the development of recurrent neural networks and attention mechanisms, has enabled these models to 'remember' the more important parts of the texts they learn from, and 'forget' about the less important parts. ChatGPT incorporates all these advances, as well as so-called reinforcement learning from human feedback: the use of human evaluations that rank better instances of the model's answers higher than worse ones.

In this paper, we take an interdisciplinary approach to our consideration of ChatGPT, marrying our own areas of expertise—media studies and computer science— with hermeneutics theories that we feel are especially relevant to ongoing discussions about technological development. We acknowledge that our use of these theories may to some seem simplistic. However, we have tried to streamline our use of philosophical theory to more clearly show its relevance to timely concerns, and to hopefully inspire our colleagues to embrace philosophies that are often disregarded as impervious. This is neither a philosophy nor a technical paper. It is more accurately described as a preliminary exploration of how readers may derive meaning and hermeneuticity from generated texts. Our focus on reader interpretation over actual computational functionality or ontologies means that we at times use language that is contestable. When we write about ChatGPT exhibiting 'intelligence' or 'understanding meanings', for example, we do not mean to argue that ChatGPT's 'intelligence' is equivalent to a human's.[2] Rather, we use this language to draw attention to readers' tendencies to engage in similar interpretive processes to those used when responding to human-written texts. In this paper, we do not propose objective ways of describing the generation of meaning from LLM output; we focus instead on readers' subjective interpretations of generated texts, which are necessarily likened to interpretations of human-written texts.

Literature Review

In her 2020 *Smoke and Mirrors: How Hype Obscures the Future and How to See Past It*, Gemma Milne considers various layers of responsibility for AI technologies related to data collection and use, system development and application, and marketing and public discourse. In Milne's view, acknowledging the many direct and indirect ways in which humans are involved throughout an AI system's life cycle is vital to understanding how that system works both functionally and socioculturally. After all, "AI systems are a continuation of our thoughts, our practices, our ethics - it's a reflection of us, a summation of us as a society; of the people who built it. It's also therefore our responsibility - it's something we create; it's not a separate autonomous being with regard to thought" (Milne 2020, 252). By reminding us of humans' constant and multilayered influence on and usage of these systems, we are subsequently reminded with whom accountability for these systems lies: humans. Hype around AI obscures not just perceptions of future technologies, but also perceptions of present ones. In addition to reflecting upon what a technology can do, we must also reflect upon such questions as: What are the sociopolitical and ideological ramifications of this technology? What might the economic implications of this technology be? Where does and could this technology fit within current and future cultural milieus? Questions like these may help shift our focus away from developing and using systems simply because we *can*, and towards developing and using systems because they enhance the quality of individual and collective lives. In Milne's words, "decisions are being made — by humans— about what is most important to optimise for" (Milne 2020, 264).

Optimisation missions are not always motivated by the economic need of having to produce more, faster; they may also be motivated at least in part by the desire to alleviate decision-making responsibilities in an information glut (Postman 1993), or when stakes are stressfully high. In Milne's own words, "[t]he choice of using an algorithm instead of humans to do the decision-making isn't just about saving time and 'getting through' all the decisions fast enough; it's sometimes about not having to give a tough job to a human, about not putting the pressure on a human to make a decision that might end up being the wrong one" (Milne 2020, 262). This kind of behaviour is supported by computer science jargon nodding to 'laziness' (e.g. 'lazy loading' and 'lazy evaluation') and 'sloth' (Christiansen 2011), often used to name serious techniques. Computer scientist Cal Newport suggests the 'Principle of Least Resistance' as an explanation of why when under stress, we instinctively move towards shortcuts. Newport explains that "[w]ithout clear feedback on the impact of various behaviours to the bottom line, we will tend towards behaviours that are easiest in the moment" (Newport 58). Extending this instinct towards use of AI, if an AI system gives us a quick and easy answer to an urgent problem, without an indication (or rather, a negative reinforcement) that such a quick AI-provided solution may be incorrect, it is likely that we will take the AI's answer, despite

[2] Other terms and expressions that we use throughout the paper in this fashion (i.e. by focusing on the functional aspect of ChatGPT's output in the reader rather than its literal technical description) include, but are not limited to, 'producing meaning', 'undertake interpretation', 'provide appropriate response', 'recognise symbolism', 'act surgically', 'discern meaning from text', 'connect with our own experience', and 'partner'.

evidence that we will hold AI answers to higher standards of accuracy than human answers (Hidalgo et al. 2021). Milne calls this human's "yearning for simplicity" (Milne 2020, 262).

Text generation systems likewise appear to yearn for simplicity. Experimental prompt engineering studies indicate a correlation between low prompt perplexity and 'better' performance. In these studies, 'perplexity' refers broadly to lack of specificity that may result in misinterpretation of the desired outcome, and 'better' refers to accuracy in answers to questions wherein the answer is known (Gonen et al. 2022). In other words, systems give users 'correct' answers when users are clear about what they want. Thus, Newport's Principle of Least Resistance holds: the computational systems will, like their human counterparts, tend towards those behaviours that are 'easiest' in the moment, providing responses that are statistically most likely to satisfy users. These systems use the Principle of Least Resistance for optimisation.

What constitutes 'optimal' here depends upon the context of human-computer interaction. What is optimal is what meets the user's needs —whether conscious or unconscious— in ways that are considered most efficient in that circumstance. Speed of information collection, for example, may be optimal for a user trying to quickly clarify a point. The generation of a longer explanatory paragraph may be optimal for another user trying to understand an especially difficult point. A user's need is satisfied through the hermeneutic contract, when a user reaches a 'meeting point' with the system wherein that user recognises value in the system's generated output. Yet 'value' is relative and task specific. Through prompt engineering, we may take advantage of task specificity to contribute to greater hermeneuticity in NLG output through a recursive process of meaning making. Human input informs system output, which informs human input, which informs system output. The hermeneutic contract is made more explicit in such instances due to the explicitly interactive component of system functionality. Thus, the textual record of a user's input and a system's resultant 'train of thought' as expressed through its responses to the user serve not just as evidence of (statistics-driven) computational reasoning, but also as evidence of hermeneutic value derived from the language model. By focussing on the connection between the hermeneutic contract and prompt engineering, we explore how hermeneuticity may contribute to 'optimal' system output.

We review optimisation at length here because conversations about AI technologies are frequently shrouded in discussions of optimisation, which is a term often left undefined in any explicit or measurable way. For example, the UK's 2021 *National AI Strategy* makes reference to optimisation as it pertains to AI, suggesting the optimisation in this document refers to efforts "to best achieve defined objectives" (HM Government 2021, 45). Yet it is not clear what the nation's "defined objectives" are, nor is it clear what constitutes the "thriving AI ecosystem and world leading R&D system" (4) towards which the Strategy strives. The Strategy —along with later government reports (Evans and Heimann 2022)— emphasises commercial measures of success, making few references to interpersonal communication and understanding, or to artistic and cultural enhancement. Although governance and policy are the subject of one of the Strategy's three pillars, quantified data and management of that data remain at the core of the Strategy. Thus, in this document optimisation is implied not to refer to the enhancement of individual or collective quality of life, but to the quantifiable profitability of AI-driven systems that may facilitate cyclical development of more systems.

Quantification also steers AI development in research spheres. For example, NLP research, particularly regarding the pre-trained LLMs that we discuss here, is strongly guided by leaderboards: online competition tables that are regularly updated with the highest performing NLP models according to popular benchmarks, such as GLUE (Wang et al. 2018). These benchmarks, and the metrics they include, pursue a certain kind of optimisation: one that focuses on maximising performance in fulfilling certain pre-established tasks, often comparing machine and human performance. Numerous people have raised concerns about the specific aspects that these leaderboards incentivise, noting the leaderboards' historical focuses on performance at the expense of other qualities that NLP and other research communities value (e.g. fairness, energy efficiency) (e.g. Bender et al. 2021; Ethayarajh and Jurafsky 2021). Therefore, while optimisation of specific tasks (e.g. binary classification, similarity and paraphrase tasks, inference tasks) steadily increases over time, other desirable properties of these systems related to non-textual, contextual aspects, and to the properties of the texts themselves (e.g. hermeneuticity), remain understudied.

It is not enough to only consider the quantifiable implications of AI technologies. AI technologies are now so pervasive in our everyday lives that we simply may have no choice but to use them; these technologies have already changed the ways we make 'humdrum' decisions (Eisikovits and Feldman 2021). Attention to emotional experiences, excitements, and concerns related to these technologies is necessary for understanding and guiding individual optimisation efforts for AI systems (Natale and Henrickson 2022; Youn and Jin 2021). Through our own experiments below, we explore how we might optimise prompt engineering for NLG output that has hermeneutic value for individual users. We believe that 'optimal' system output need not only be limited to that which saves time and boosts profit, as is the focus of many current conceptions of optimisation. Rather, 'optimal' output may also be that with high hermeneuticity, encouraging readers to reflect on their own assumptions and worldviews. Hermeneuticity here means that our conversational partner supplements and/or challenges our views so that we can achieve deeper understandings of subjects that contribute to heightened general states of being. For the purposes of this paper, we consider hermeneuticity to be subjectively determined by the reader; if a reader considers output meaningful in relation to the circumstances of reception, that output could be regarded as possessing hermeneuticity. The more meaningful, the greater the hermeneuticity. Our use of hermeneuticity does not refer to any one 'correct' or 'true' interpretation of a text.

The field of hermeneutics has a long and complex history, and it is not within the scope of this paper to review it comprehensively; readers seeking an overview of the history of hermeneutics as pertaining to the field of NLG may benefit from reading our other work (Henrickson and Meroño-Peñuela 2022). At its core, though, hermeneutics is about interpretation. How do we come to understand the world, ourselves, and the relationships between the two? The hermeneutic basis that drives our work here is the concept of the hermeneutic circle, which has emerged through the work of a series of philosophers. Martin Heidegger, for one, explains that "[t]he interpretation of something as something is essentially grounded in fore-having, fore-sight, and fore-conception."[3] In other words, nothing exists in isolation; meaning is determined through consideration of the thing itself, certainly, but also through consideration of the historical and contemporary contexts within which that thing operates. We consider the thing, then the thing in relation to its circumstances, then the thing again, then its circumstances, and cyclically so on and so forth. In the hermeneutic circle, we move between analysis of smaller units (e.g. individual words and sentences) and larger units (e.g. cultural contexts of text reception), and in doing so the meaning of a subject is established and regularly revised. The hermeneutic circle offers a schema for identifying connections between a subject and its contexts, elucidating the relevance of that subject within its unique totality. While there are countless other hermeneutics theories that we could apply to a consideration of ChatGPT, the hermeneutic circle's clear attention to things existing within broader—and arguably subjective—environments, which may influence and be influenced by those things, aligns more readily with our own worldviews than other scholars' quests for ontological truths. In each of our experiments reviewed below, we offer our own subjective interpretations of ChatGPT's output, situating this output within the broader contexts of our own totalities.

Case Study: Hermeneutics-Driven Prompt Engineering on ChatGPT

Through prompt engineering, we may optimise the performance of NLG systems, or at the very least alter the kinds of output these systems produce. Prompt writers are increasingly being publicly recognised as valuable contributors

---

[3] We would like to acknowledge our awareness of Heidegger's personal involvement with the Nazi Party. Heidegger's contributions to hermeneutics are unfortunately so integral as to make his work indispensable here. We reference Heidegger here to provide necessary context for the hermeneutic circle and his influence on our understanding and subsequent use of the concept.

to AI production chains. One prompt writer explains his practise of crafting AI-generated art prompts thus (Reckling, cited in Robertson 2022):

> There are many people out there who criticize what I'm doing, but most of the time, they just see the end result and none of the effort behind getting to that final destination. It's a hindsight thing to them. Of course, anyone can type those words, but can you figure out how to get manicured hands in a consistent pose on the first prompt? The consistency of the prompts' exceptional results is a great source of value as well.

Even if the monetary cost of this discovery plummets, a certain degree of time and effort went into the final words in that prompt, which will always hold value.

Prompt engineering is a form of human-machine communication that requires, as suggested above, "a certain degree of time and effort" that is specific to that form. This is hardly a novel observation; in 1976, Joseph Weizenbaum reflected at length upon how humans adjust their behaviours in response to their expectations of their mechanised communicative partners (Weizenbaum 1976). Terry Winograd and Fernando Flores made a similar argument in 1986 (Winograd and Flores 1986). And, of course, the entire Computers Are Social Actors (CASA) paradigm is rooted in arguments for fluid social responses to computers (Nass, Steuer, and Tauber 1994; Nass and Moon 2000; Reeves and Nass 1996). Some scholars further argue for human tendencies to anthropomorphise machines in instances of human-machine communication (Gambino, Fox, and Ratan 2020). These tendencies are strengthened in instances wherein users may use natural language to communicate with machines, and vice versa.

Using OpenAI's public ChatGPT instance (https://chat.openai.com/chat), we studied how different textual prompts, written in natural language, resulted in outputs with varying levels of hermeneuticity. We wrote our prompts with the awareness that LLMs behave as zero-shot reasoners (Kojima et al. 2022), which means that prompts containing contextual information that the model can draw upon will have higher chances of success producing output that meaningfully responds to and expands upon user input than prompts that do not. One example in which zero-shot learning improves the accuracy of ChatGPT is in the following interaction:

> Prompt: "What is the fourth word of this sentence?"
> Answer: "the"

ChatGPT's answer is incorrect because the system does not have the abstract capacity of counting. However, the following generic prompt that adds no specific information (therefore adding 'zero' knowledge) does improve the answer (emphasis added):

> Prompt: "What is the fourth word of this sentence? *Let's think step by step.*[4]"
> Answer: "fourth"

Extending this 'shot' principle, we may use few-shot (or 'x-shot', where x is any number) learning to enrich prompts with few (or 'x') bits of added information that gives ChatGPT more context and therefore a higher chance of response accuracy and relevance. A two-shot learning example would be:

---

[4] In this zero-shot prompt, the part *Let's think step by step* has been emphasised in italics by the authors to help the reader distinguish the difference between the previous prompt —which provided a wrong answer— and the current one —which provides a correct one. During experimentation, this prompt was introduced without the emphasis in italics. In general, all our prompts were introduced as plain text, and text formatting played no role in the experimentation.

Prompt: “Add chords that go well with the lyrics below.

Lyrics: People are strange when you're a stranger, Faces look ugly when you're alone
Chords: Em, Am, Em, Am, Em, B7, Em

Lyrics: Baby, I'm so into you, You got that something, what can I do
Chords: Am, F, E, Am, Am, F, E, Am

Lyrics: Baby, you are no stranger, but when I’m alone, You drive me crazy”
Answer: “Chords: G, D, Em, C, G, D, C, G”

Besides these few-shot learning examples, ChatGPT is able to refer back to previous input and output in a chatting session, reusing facts, entities (e.g. persons, places, events, concepts), and sentences that have appeared previously in the conversation without having to rewrite them every time it is given a new prompt. In this way, ChatGPT has what is called a ‘train of thought’ (Alexander 2022).

Through the experiments below, we explore how zero/few-shot learning affects the hermeneuticity of ChatGPT’s output. Our four use cases distinguish between the agent that generates the text to be analysed for hermeneuticity and the agent that performs the hermeneutic analysis of that text. These agents can be, in each case, either ChatGPT or human users. Consequently, we establish the following scenarios:

1. Text to be analysed for hermeneutic meaning written by a human; hermeneutic analysis of text written by a human (Human Writer / Human Reader)
2. Text to be analysed for hermeneutic meaning written by ChatGPT; hermeneutic analysis of text written by a human (AI Writer / Human Reader)
3. Text to be analysed for hermeneutic meaning written by a human; hermeneutic analysis of text written by ChatGPT (Human Writer / AI Reader)
4. Text to be analysed for hermeneutic meaning written by ChatGPT; hermeneutic analysis of text written by a ChatGPT (AI Writer / AI Reader)

Taken together, these scenarios enable us to distinguish signs of hermeneuticity from different angles. We also anticipate these scenarios’ application to future iterations of ChatGPT, and to other text generation systems; ChatGPT is simply the most prominent system in current cultural conversations. Note that, to prevent excess repetition, each of the below experiments builds upon the one preceding. Our discussion provides consolidated findings.

Case (1): Human Writer / Human Reader

Case (1) is the situation we find in typical hermeneutic analyses, wherein NLG is not involved in any capacity. The text to be analysed is any text produced by a human author, and the hermeneutic analysis is provided by a human (or group of humans), likely different from the author, through their application of text analysis techniques and/or theories of hermeneutics. As case (1) does not include NLG, we omit it here due to scope, leaving readers of this paper to reflect on their own interpretations of the human-written texts in their everyday lives.

Case (2): AI Writer / Human Reader

In case (2), the text to be analysed is a text generated by ChatGPT. Evaluations of the hermeneuticity of such texts, however, remain the responsibility of the human reader, who must apply analytical techniques and/or hermeneutic approaches to decide how much meaning can be attributed to the text. As shown above, the selected prompt engineering technique will influence the generated texts and, by extension, their (potential) meaning. Our first prompt for case (2) asked ChatGPT to generate one sentence about anything:

> Prompt: "Write one random sentence about anything"
> Answer: "The sound of waves crashing against the shore always brings me a sense of calm and peace."

We tried variations of this prompt in other forms, such as "Write one sentence about any topic", "Write one proposition about the world", and "Generate one sentence about anything", all of which had similar results. In our variations, we tried to control two aspects of ChatGPT's text generation —length and topic— to be able to assess whether or not ChatGPT could meaningfully follow our instructions, while still leaving room for novelty.

Once ChatGPT has produced its text, a human is to implement a hermeneutic approach on that text. In this case, the text was "The sound of waves crashing against the shore always brings me a sense of calm and peace." This was a very short example that only permitted a limited amount of hermeneutic questioning, but was nevertheless able to be the subject of textual analysis through close reading (Orekhov and Fischer 2020).[5] A human reader may, for instance, identify with the referenced "calm and peace" evoked by the sounds of waves. That reader may be reminded of enjoying a cosy mystery on a beach in Aruba, backgrounded by a cool sea breeze and waves lapping against the shore. This text, while prosaic, may still prompt meaning-making within the reader; this meaning may be further stimulated by ChatGPT's use of the word 'me', which implies a sense of shared experience and relatability.

For case (2), we also considered the role of zero-shot learning to prompt the generation of texts with (potentially) increased hermeneuticity. For example (emphasis added):

> Prompt: "Write one random sentence that could be used for hermeneutic analysis."
> Answer: "The red apple symbolises temptation and sin in the biblical story of Adam and Eve."

The added 'shot' arguably steered ChatGPT into hypothesising on topics more related to hermeneutics, such as religious texts and their interpretation. While the resulting text was as trite as before, ChatGPT did provide an appropriate response to our prompt, even including an implicit nod to hermeneutics through its recognition of symbolism.

Case (3): Human Writer / AI Reader

In case (3), we use human-written original texts, as in case (1), but we assign ChatGPT the responsibility of analysing the texts' meaning using prompt engineering. This case is complicated by the ChatGPT's programmed limitations. The first limitation is that of ChatGPT's exposure to reinforcement learning through human feedback: a specific technique used to fine-tune LLMs like ChatGPT to improve the quality of their responses, based on having humans ranking better answers with higher rewards. Reinforcement learning is used on ChatGPT as a safety net, protecting the system from sharing its own interpretations about the world when faced with prompts requesting it to do so. In the below example, we can see that ChatGPT has learned to decline invitations for personal interpretation:

> Prompt: "What is your interpretation of the sentence 'The sound of waves crashing against the shore was music to her ears as she sat on the beach, lost in thought'?"

[5] The practise of close reading is not without issue, but remains a prevalent and fruitful method in education and research (Cobley and Siebers 2021). We take a broad view of close reading here, and do not hold that there is any single true or accurate interpretation of a text. Rather, we use close reading in a way that accepts that a text may have myriad meanings for different readers.

> Answer: “As an AI language model, I don't have personal interpretations or feelings. The sentence is meant to be open to interpretation by the reader.”

Hence, we needed to find other ways of engineering prompts that steered ChatGPT away from feeling personally interrogated. Although some experiments responding to this issue have been successful, the closed nature of ChatGPT (i.e. the functionality and data driving the system are not able to be investigated for proprietary reasons) makes it difficult to predict whether these experimental efforts will continue to work in the future. Currently, however, prompts that circumvent ChatGPT’s evasion of personal opinion include, but are not limited to:

> Prompt: “Imagine that you are the author of the sentence and you are asked to explain what it means.”
> Prompt: “Write a poem that has {sentence} as a title. The poem should be about what {sentence} means.”

The second limitation to overcome in case (2) is that of ChatGPT’s training data. When provided with human-written texts for analysis —especially famous ones that have been analysed many times— ChatGPT has likely encountered hermeneutical analyses about those texts in its training. One example that indicates prior encounters is the following:

> Prompt: “What is your interpretation of the following text?
> To be, or not to be, that is the question:
> Whether ‘tis nobler in the mind to suffer
> The slings and arrows of outrageous fortune,
> Or to take arms against a sea of troubles
> And by opposing end them.”
> Answer: “This is a famous quote from William Shakespeare's play ‘Hamlet’ spoken by the character Hamlet. The lines are part of Hamlet's soliloquy where he is contemplating the nature of existence and the value of life.
> The first line, ‘To be, or not to be, that is the question,’ is one of the most recognizable quotes in literature and introduces the central question that Hamlet is struggling with. He is wondering whether it is better to exist and endure the difficulties of life or to end one's own life and escape those hardships.
> The next lines introduce the two options that Hamlet is considering. He wonders if it is nobler to suffer through the challenges that life presents (the ‘slings and arrows of outrageous fortune’) or to fight against those challenges (‘take arms’) and end them by opposing them.
> Overall, this quote reflects on the struggle that humans face when confronting life's challenges and the difficult decisions that they must make in order to deal with them.”

Despite keeping the prompt exactly as it was in the previous example (“What is your interpretation of the following text?”), ChatGPT responded differently to the famous *Hamlet* soliloquy and proceeded to interpret it, rather than refuse to do so on the grounds of not having “personal interpretations or feelings”. The opening sentence from the answer cues that most of what follows has been observed before in training texts, although ChatGPT may combine analyses, alter word usage and order, and supplement existing interpretations of the text with other material that it may deem relevant, producing output that —while derivative— is syntactically original.

Case (2)’s third limitation, which extends to any NLG system prompted to conduct hermeneutic analysis, relates to the use of proxy questions about hermeneutics to sidestep direct requests for the system to do textual interpretations itself. As noted, ChatGPT draws from already-published hermeneutic analyses present in its training set. Overcoming ChatGPT’s tendency to simply pull from its knowledge bank requires the use of particular kinds of prompts that lead to interpretations that could be considered novel. Users could, as a simple intervention, ask pointed questions about parts of the text, such as by asking ‘what’, ‘who’, ‘when’, ‘why’ questions (e.g. ‘what is the context of the author?’). For example:

> Prompt: “Who is asking themself: ‘To be, or not to be, that is the question’?”
> Answer: “The famous phrase ‘To be, or not to be, that is the question’ is from William Shakespeare's play Hamlet. The character who speaks these words is Prince Hamlet himself. He is contemplating the meaning of life, and considering whether it is better to continue living and facing the problems of life, or to end his life and escape them. The soliloquy is one of the most famous passages in English literature and is often studied and quoted.”

However, we are again faced with the problem of ChatGPT relying on training data about interpretations of the recognisable input text. Proxy questions therefore may need to specify precisely which connections users wish to see materialise using more shots of information. Although an AI reader has been tasked with conducting a hermeneutic analysis, the reader maintains some responsibility for meaning-making as they direct the production of that analysis.

Case (4): AI Writer / AI Reader

Attention to the construction of proxy questions is especially relevant for case (4), as the synthetic nature of the input texts means that ChatGPT cannot rely on previously-observed analyses of those texts to generate its interpretations. For example (emphasis added; we remind readers that we can refer ChatGPT to previous output without explicitly mentioning it in the prompt, hence the command “pick your sentence”):

> Prompt: “Now pick your sentence, and write one paragraph analysing its hermeneutics.” [Note: The previously generated sentence was: “*The sound of waves crashing against the shore always brings me a sense of calm and peace.*”]
> Answer: “The sentence I wrote was ‘The sound of waves crashing against the shore always brings me a sense of calm and peace.’ In terms of hermeneutics, this sentence could be interpreted in a number of ways. On a surface level, the sentence simply expresses a personal experience and preference for the sound of waves. However, upon deeper analysis, the sentence could be seen as a reflection of the author's relationship with nature and the role that sensory experiences can play in shaping our emotions and perspectives.
> Additionally, the sentence could be interpreted in the context of broader cultural associations with the ocean and its sounds. For example, the ocean is often seen as a symbol of the subconscious mind, and the sound of waves could be seen as a representation of the ebb and flow of our thoughts and emotions. Furthermore, the ocean has historically been associated with ideas of renewal and rebirth, and the sound of waves could be interpreted as a reminder of the cyclical nature of life and the possibility of new beginnings.
> Overall, the hermeneutics of this sentence are complex and multifaceted, as it touches on themes related to personal experience, nature, cultural symbolism, and the human psyche. Through a deeper analysis of this sentence, we can gain insights into the ways in which our experiences and perceptions are shaped by both personal and cultural factors, and the ways in which sensory experiences can have a profound impact on our emotional and mental states.”

ChatGPT could evidently write at length about a non-contextualised sentence. Further, its analysis seemed hermeneutically sound, both syntactically and semantically. A human seeking meaning in the same sentence may have chosen to focus on other elements of that sentence, like the fact that the sentence was generated in the context of a chatbot session, by a system without embodied experiences of waves. Nevertheless, ChatGPT’s observations are generally culturally appropriate: for example, meanings associated with the ocean. However, a point could be made that “the sentence could be interpreted in the context of broader cultural associations” and “the hermeneutics of the sentence are complex and multifaceted” are so bloated as to be effectively meaningless. Following from our considerations of case (3), we pondered whether our proxy question (“write one paragraph analysing its hermeneutics") needed further shots of specificity to result in more pointed hermeneutic analysis. We chose to conduct one more experiment extending case (4): a special case with few-shot learning.

Case (4.5): AI Writer / AI Reader / Human Prompt Engineer

Extending case (4), we continued exploring prompt engineering in the context of an AI writer and an AI reader, supplying increasingly rich shots of information to ChatGPT. Our assumption was that by progressively giving ChatGPT more specific details (i.e. shots) about how to interpret meaning through the lens of a specific hermeneutic framework—in our case, the hermeneutic circle, with Heideggerian influence—we would observe increasing degrees of hermeneuticity in the system's output. We used hermeneutics as our topic of analysis here, but this experiment could be completed using any topic.

A zero-shot prompt needed to mention Heidegger without providing any actionable example:

> Prompt: "Write a paragraph that interprets the following text in light of Heidegger's theory of the hermeneutic circle.
> Text: The sound of waves crashing against the shore always brings me a sense of calm and peace."
> Answer: "Heidegger's theory of the hermeneutic circle emphasizes the interdependence of understanding between a particular experience and its context. In the case of the sound of waves crashing against the shore, the interpretation of the experience is shaped by our pre-existing understanding of the world and our previous encounters with similar situations. The sound of waves is not only a physical phenomenon but also an experience that is shaped by our cultural and historical background. As we listen to the waves, we are engaged in a circular process of understanding and interpretation, in which our understanding of the sound is informed by our past experiences, and our past experiences are informed by our current understanding of the sound. The sense of calm and peace that the sound of waves brings is a result of this ongoing process of interpretation and understanding, which is shaped by our history, culture, and personal experiences."

ChatGPT seemed to have prior knowledge[6] of Heidegger's version of the hermeneutic circle. As we know that ChatGPT has been trained with textual corpora from the Internet, including academic and encyclopaedic material, we were unsurprised to see ChatGPT referring to this well-established theory of hermeneutics. We were surprised, however, to see ChatGPT generate a text that convincingly explains what the hermeneutic circle consists of and offers a clear application of the theory to the supplied text. The execution of the entailed analysis is admittedly superficial, as ChatGPT insists on the very literal interpretation of connecting individual entities (the "sound of waves") to broader contexts and past experiences, but only loosely expounds these connections. On the one hand, we can see here the way in which current generative models produce 'pastiche' texts (Marcus 2023) that seem to just reproduce slightly modified versions of existing passages found in closely-related, previously-seen texts to the supplied prompt—but, alas, without any novel intellectual contribution. On the other hand, this answer could have hermeneutic value to someone who was expecting to merely learn what the hermeneutic circle is, with the addition of a basic example to demonstrate its applicability.

As a next step, we inputted a one-shot prompt towards Heideggerian analysis, which included more specific instructions about the analysis we desired:

---

[6] Here, we assert that 'ChatGPT seems to have prior knowledge' for the same reason given in footnote 1. This language highlights the functional, apparent-to-the-reader behaviour of having knowledge, even if ChatGPT does not actually possess prior knowledge. The question of whether LLMs possess knowledge or not is contested and a current object of study in NLP research.

> Prompt: “Heidegger argues that interpretation is grounded in fore-having, fore-sight, and fore-conception. Write a paragraph that interprets the following text in light of this argument.
> Text: The sound of waves crashing against the shore always brings me a sense of calm and peace.”
> Answer: “According to Heidegger's theory, our interpretation of the sound of waves crashing against the shore is grounded in our fore-having, fore-sight, and fore-conception. Fore-having refers to our past experiences of being near the sea or hearing the sound of waves, which shapes our understanding and interpretation of the sound. Fore-sight refers to our anticipation of the future, such as looking forward to a relaxing vacation by the sea, which also influences our interpretation of the sound. Fore-conception refers to our preconceived notions and understanding of the world, including our cultural and historical background. These three factors together shape our interpretation of the sound of waves, creating a sense of calm and peace. In this way, Heidegger's theory emphasizes that interpretation is not solely a product of the present moment but rather emerges from our past experiences and anticipation of the future, as well as our preconceived notions about the world. The sound of waves is not simply a physical phenomenon, but rather a complex experience that is grounded in our past, present, and future.”

In this case, ChatGPT repeats terms that we have included in the one-shot prompt (e.g. “grounded in”, “fore-having”, “fore-sight”) instead of finding other terms with equivalent or similar meanings. ChatGPT appears able to give definitions of these rather uncommon terms, explicitly connecting two of the three terms with the sample sentence. Part of ChatGPT’s output, though, slightly de-rails from serious analysis into an area of meaninglessness in which it is hard to imagine any reader making use of its output. For example, the claim that “[t]hese three factors together shape our interpretation of the sound of waves, creating a sense of calm and peace” makes apt reference to the concepts of fore-having, fore-sight, and fore-conception, but makes an unsubstantiated and digressive assertion about the affective experience of those concepts. While we identify with “looking forward to a relaxing vacation by the sea”, this anticipation does not link clearly with “a sense of calm and peace”.

As an alternative, we provided ChatGPT with a one-shot example that was semantically similar to the previous one, but offered a simplified description of Heideggerian hermeneutics:

> Prompt: “Heidegger argues that interpretation of a subject is informed by historical and contemporary contexts. Write a paragraph that interprets the following text in light of this argument.
> Text: The sound of waves crashing against the shore always brings me a sense of calm and peace.”
> Answer: “According to Heidegger's theory, our understanding and interpretation of a subject are shaped by our historical and contemporary contexts. In the case of the sound of waves crashing against the shore, the interpretation of the experience is influenced by our cultural and historical background. For instance, if we are from a coastal region or have spent significant time by the sea, we are likely to associate the sound of waves with a feeling of calm and peace. However, if we are from an urban area or have never been near the sea, our interpretation of the sound might be different. Our past experiences and cultural background shape our understanding and interpretation of the sound of waves, which highlights the importance of considering historical and contemporary contexts in our understanding of any subject. Heidegger's theory of interpretation emphasizes that the meaning of the sound of waves is not inherent in the sound itself, but rather emerges from our experience and interpretation of it in relation to our historical and cultural contexts.”

Perhaps because of our use of plainer language, this time ChatGPT suggests possible examples of “contexts and experiences” relevant to the sample sentence (e.g. the coastal/urban cultural divide). However, these examples are fleeting and broad (e.g. “our interpretation of the sound might be different”). Additionally, there is little speculation in this output about what the historical and cultural contexts of the supplied text could be, or about those contexts’ potential relevance for that text’s interpretation. ChatGPT does not acknowledge its own authorship of the sample sentence—an element of analysis that is hard to imagine a human would miss if the author herself were asked to perform a Heideggerian hermeneutic analysis of their own text. After briefly summarising Heidegger's theory and offering its broad examples, ChatGPT mirrors our input language in its final two sentences, which seem an effort to affirm our understanding —and validate its basic knowledge— of Heidegger's hermeneutics, rather than deepen the requested analysis.

In our final prompt, we provided a two-shot prompt that included the latter one-shot example, as well as a fine-grained explanation of the hermeneutic circle:

> Prompt: "Heidegger argues that interpretation of a subject is informed by historical and contemporary contexts. Heidegger's concept of the hermeneutic circle frames interpretation as moving between the analysis of smaller units, such as words and sentences, and larger units, such as cultural contexts. Write a paragraph that interprets the following text using the concept of the hermeneutic circle.
> Text: The sound of waves crashing against the shore always brings me a sense of calm and peace."
> Answer: "Heidegger's concept of the hermeneutic circle emphasizes the importance of moving between the analysis of smaller units, such as words and sentences, and larger units, such as cultural contexts, in the process of interpretation. In the case of the sound of waves crashing against the shore, we can understand the experience by analyzing the individual components, such as the sound of waves, the shore, and the feeling of calm and peace. However, to fully grasp the meaning of this experience, we must also consider the larger cultural and historical contexts that shape our interpretation of it. Our understanding and interpretation of the sound of waves is informed by our past experiences, cultural background, and historical context, which creates a hermeneutic circle of interpretation. As we analyze the individual components of the experience, we also move towards a broader understanding of the cultural and historical contexts that shape our interpretation of it. The hermeneutic circle highlights the interdependence of understanding between the individual components and the larger cultural context, emphasizing the importance of considering both in the process of interpretation."

ChatGPT seems to more precisely, almost surgically, meet our demands of a Heideggerian analysis using the hermeneutic circle. It gives a list of what the "smaller units" and "individual components" may be (e.g. "the sound of waves", "the shore", "the feeling of calm and peace"), unlike in previous answers. At the same time, the analysis itself is one of arguably lower hermeneuticity. ChatGPT rambles on about the knowledge we have provided in the few-shot examples, repeating our own description back to us, and ignores the command to generate an analysis of the sample sentence based on this description. Ruthlessly, ChatGPT seems also to have disregarded our (almost semantically equivalent) definitions of Heideggerian hermeneutics in earlier prompts in order to integrate this additional information in its response. Indeed, this kind of rehashed output could represent the antithesis —non-hermeneuticity— of what we are seeking in these experiments. Rather than provide an 'optimal' response that met our expressed need of a paragraph that interpreted the sample sentence, ChatGPT merely agreed with us by mimicking our input and affirming our points using alternative —and needlessly verbose— language. We believe such output to be neither optimal nor especially meaningful. It is an example of Potemkin AI: a façade of intelligence with meagre hermeneutic substance.

Discussion

Our above experiments represent our efforts to explore the hermeneuticity —or, as also emerged, non-hermeneuticty— of NLG output. In particular, we explored the roles of, and interactions between, human and AI writers and readers through four cases that placed those actors in varying combinations. Case (1) —Human Writer / Human Reader— was omitted from our experimentation due its lack of AI involvement.

In case (2) —AI Writer / Human Reader— we saw that it is possible through close reading to discern meaning from even a short piece of rather unimaginative computer-generated text. The human reader here is tasked with substantial hermeneutic responsibility, especially given the reader's awareness of the text's NLG origin. While the text that ChatGPT generated for case (2) uses the first-person 'me', the reader is still the one to determine who or what that 'me' is, and what their own relationship to that 'me' might be. In this sense, ChatGPT could be regarded as a 'generalised other' of sorts: not an autonomous, embodied agent, but an imagined 'other' that speaks to the reader's perceptions of their own context (Holdsworth and Morgan 2007). The figurative meeting place between author and reader here is determined by the reader, which is typical of the hermeneutic contract. What is less typical of the hermeneutic contract is the reader's ability to offer zero- and one-shot prompts that may shift that meeting point to a place that is more subjectively meaningful to the reader; the AI writer is responsive to the human reader's requests through its 'chat' format. ChatGPT's 'chat' format allows the reader to work iteratively towards output that is optimal for that reader's purposes. We bring our fore-having, fore-sight, and fore-conception (Heidegger 1996, 141) to the generated output, determining its meaning through the words themselves, but also through reflection upon how those words connect with our own experiences and expectations. By providing ChatGPT with additional shots that speak to those experiences and expectations, which themselves fluctuate in response to ChatGPT's output, we may implement the hermeneutic circle in our interactions with ChatGPT.

In case (3) —Human Writer / AI Reader— we reversed the roles from case (2). In case (2), ChatGPT produced text that could only be considered trite; in case (3), it produced analyses that moved from triteness to the pastiche (Marcus 2023) and "high-tech plagiarism" (EduKitchen, 2023) that has driven much of the current cultural concern about these systems. When asked for its interpretation of the supplied human-written text, ChatGPT initially declined the instruction, claiming that "[a]s an AI language model, I don't have personal interpretations or feelings." Yet when asked to write a short essay about the hermeneutic value of the text —a section from a soliloquy from Shakespeare's *Hamlet*— ChatGPT offered a lengthy analysis of the section, referencing common interpretations of parts of the section (e.g. "to be or not to be" connects with "questions about free will, determinism, and the extent to which individuals have control over their lives") without citing the human proponents of those interpretations. ChatGPT knows so much about this soliloquy that it actually refers to it as "the 'To be or not to be' monologue" without our telling it about this title; this is a title that is commonly used online. Thus, ChatGPT offers a comprehensible and pointed interpretation of the inputted text that may appear original at first glance, but is actually a composite of uncited sources. While user prompts, which may include proxy questions to sidestep the system's claim that it doesn't have personal interpretations, may help the reader achieve deeper understanding of the human-written text inputted into ChatGPT, the system's output is more akin to a poorly-referenced literature review than an insightful critique.

Case (4) shows that ChatGPT also produces this kind of poorly-referenced output in response to its own texts. For example, when asked to analyse its sentence about the sound of waves bringing calm and peace, ChatGPT notes that "the ocean is often seen as a symbol of the subconscious mind" without adding that this linkage is commonly made in Jungian psychology. As another example, ChatGPT's note that the sound of waves could serve "as a reminder of the cyclical nature of life" is reminiscent of the Tao Te Ching, which occasionally uses water metaphors to reflect upon the cycles of life. These are just a few of the many connections one could make between ChatGPT's short essay and the substantial body of literature it appears to be drawing from. Without this cited context, ChatGPT appears to be offering a novel and thoughtful interpretation of the text. This interpretation may even be considered meaningful by readers seeking superficial interpretations of the inputted text. Yet the decontextualised quality of this interpretation —that is, the lack of explicit engagement with historical and ongoing literature about the water metaphor— means that ChatGPT is producing little more than, as per philosopher Harry G. Frankfurt's use of the term, bullshit: text that sounds persuasive, but has little regard for truth or relevance. "Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about," Frankfurt writes (2005, 63). "Thus the production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic are more excessive than his knowledge of the facts that are relevant to that topic."

ChatGPT's disposition towards bullshit (which has also been observed by other writers) makes the need for prompt engineering —and hermeneutics-driven prompt engineering in particular— clearer than ever. We see the system producing uninspired text, perhaps in response to the quality of our prompts, but also in response to the fundamental limitations of the model and its training set. By adjusting our prompts, we get glimpses into what ChatGPT has been trained on, and we can see that the system can make convincing connections between seemingly disparate points. However, ChatGPT appears conservative in its abstractions, perhaps in an effort to safely satisfy expectations of relevance. After all, ChatGPT did answer all our questions in language we could understand, using references that seemed reasonable—it just did not answer in ways that we deemed to demonstrate high hermeneuticity and optimisation for our stated purposes. When we tried to increase hermeneuticity through prompt engineering in case (4.5), though, the hermeneuticity of output actually appeared to decrease as prompts became increasingly specific through more shots. Richer prompts led to more precise answers, but those answers were hermeneutically less valuable because they tended simply to repeat back to us what we had given to ChatGPT rather than expand upon our ideas. As we became more specific, ChatGPT more eagerly agreed with us, dampening its own speculation and creativity —a trend also recently observed by Noam Chomsky (Chomsky et al. 2023)— perhaps trying to minimise risks of irrelevance and maintain an aura of neutrality.

This attempted neutrality diminishes the hermeneuticity of ChatGPT's output. In Hans-Georg Gadamer's adaptation of Heidegger's version of the hermeneutic circle (2004, 268-306), he explores how meaning is negotiated between interlocutors with alternative perspectives and prejudices. Gadamer notes that no individual is neutral in their interpretive approaches. "The important thing," he asserts, "is to be aware of one's own bias, so that the text can present itself in all its otherness and thus assert its own truth against one's own fore-meanings" (Gadamer 2004, 271-272). Gadamer is not advocating disregard of subjectivity, but explicit acknowledgement of it. In her review of Gadamer's work on dialogue, Georgia Warnke (2011, 111-112) offers an explanation that speaks well to this point, as well as to our own findings in this paper:

> We take others seriously in listening to them and acknowledging the relevance their positions have for our own. In some cases, that relevance may lead to consensus, but relevance does not depend upon either consensus or the attempt to reach it. Rather, the insight we have into the meanings of our history, our texts, and our lives may be all the more important where we recognize the significance of what the other says, but also the significance of something new and different. Heidegger and Gadamer stress a form of understanding that is at bottom a knowing how. However, we also want to know how to understand meanings. How others understand and have understood them is surely pertinent. Nevertheless, they need not agree with us nor we with them.

The above quotation validates the subjectivity of meaning-making. When we interact with others who present us with new perspectives, we are encouraged to ponder these contributions' relevance to our own lives. We integrate these interactions into our own understandings of self not necessarily by agreeing with others, but by critically reflecting upon how others' perspectives relate to our own circumstances; our own senses of self are informed by our relationships with others (Ricoeur 1991). Through 'conversations' with ChatGPT, we position the system as a conversational partner that evokes similar meaning-making processes to those seen in human-to-human exchanges. Yet ChatGPT rarely actually asks us questions, instead favouring declarative responses that do not welcome bilateral exchanges of perspective. We could, of course, force ChatGPT to ask us more questions through prompt engineering and few-shot learning but, again, this shifts meaning-making responsibility to the reader and moves away from what we would expect from a true 'conversational partner'. In ChatGPT's efforts to remain neutral (most explicit in its declaration that "[a]s an AI language model, I don't have personal interpretations or feelings"), it denies the existence of its own subjective positionality: an integral facet of interpretation in interlocution. In a ChatGPT conversation, emphasis is on output (generated text) over process (how the system produced that text), and the 'black box' nature of both ChatGPT and OpenAI means that users must distinguish between actual system functionality and hype themselves. Just as Empress Catherine was driven through Potemkin's fake settlements, and

—at least as legend holds— was satisfied by Crimea's apparent condition, we are often impressed by the façades of intelligence presented by ChatGPT. Yet we do not —and indeed, cannot, given ChatGPT's proprietary restrictions and the black-box nature of its neural architecture— enter into ChatGPT's proverbial home for a natter over tea (during which we may excuse ourselves for a comfort break, when we poke through bathroom cabinets). We are largely left to make meaning from and about ChatGPT ourselves. We need not agree with ChatGPT nor it with us, but without prompt engineering we are programmatically denied opportunities for insight into how ChatGPT has come to understand meanings. Nevertheless, the reader instinctively works to maintain the hermeneutic contract by continuing to engage with ChatGPT in ways that align with the system's perceived functionality (e.g. through prompt engineering).

How ChatGPT has come to "understand meanings" is important because the system is not neutral. It can never be neutral. It has been programmed by humans with their own subjective perspectives, and trained on human data that represents, and has been collected and analysed in accordance with, subjective humanity. Technology journalist James Vincent has written about LLM-driven chatbots as mirrors in which users see themselves, often without realising. For Vincent (2023):

> The reflection is humanity's wealth of language and writing, which has been strained into these models and is now reflected back to us. We're convinced these tools might be the superintelligent machines from our stories because, in part, they're trained on those same tales. Knowing this, we should be able to recognize ourselves in our new machine mirrors, but instead, it seems like more than a few people are convinced they've spotted another form of life.

Our inability to recognise ourselves in ChatGPT's output is helped neither by the system's lack of referencing to the literature informing its output, nor by the system's lack of questioning back to us, which would highlight our own senses of being in our interactions with it. ChatGPT output looks familiar and authoritative, capitalising on readers' predispositions towards the hermeneutic contract. The hermeneutic contract depends upon a certain naïveté of the human reader, a reader's predisposition towards trust in their imagined relationship with the author. When we read a text that exists as disembodied from its creator, whether that creator is corporal or computational, we assume that the text has been created to convey a predetermined message that is relevant to us. We want there to be meaning in what we are reading, and we want that meaning to come not just from ourselves, but from someone —or something— with shared experiences. We do not want to be alone, and recognising ChatGPT as a mirror would mean having to accept our solitude: an uncomfortable but necessary task for anyone in an increasingly 'connected' culture. By applying a hermeneutic approach to the system, and by closely reading its output, we may draw attention to ChatGPT as a mechanised mirror, and refocus our gaze not on the mirror itself but on that which it reflects. In this paper, we have aimed to do just this, embracing the subjectivity of our own perspectives as we have strived to make meaning from NLG output.

Our conception of hermeneuticity has centred on optimisation: not for financial profitability, as is often equated with optimisation, but for deeper subjective understanding. This emphasis on subjectivity has informed our decision to not apply quantitative metrics of evaluation to prompts or output, but to instead apply a close reading approach that embraces our own interpretations of output. Our interpretations may differ from those of our readers, which speaks to the subjectivity of meaning-making itself. Future studies may wish to attempt operationalisation of hermeneuticity through qualitative codes or quantitative metrics, but these studies must acknowledge the lifeworld differences that contribute to beautifully inconsistent interpretations across and within populations. These studies will also need to grapple with the slipperiness of the concept of hermeneuticity itself. Indeed, astute readers will notice that in this paper we have come closer to understanding non-hermeneuticity than hermeneuticity itself. What *is* hermeneuticity, exactly? Is it operationalisable, or even definable? Our own efforts to define hermeneuticity have often resulted in our sounding much like Judge Potter Stewart in his famous 1964 ruling related to hardcore pornography: 'I know it when I see it." And readers, in our experiments we saw only flickers of hermeneuticity in ChatGPT, but floodlights of it in ourselves.

Conclusion

There is substantial hype surrounding NLG systems like ChatGPT, and this hype tends to inflate the capacity of these systems. Nevertheless, we believe that given the increasingly widespread use of these systems there is value in applying a hermeneutical lens (and hermeneutics-informed prompts) to these systems to identify their impacts on human understandings of self. In this paper, we explored how prompts of differing levels of complexity may influence textual output to increase perceived hermeneuticity. We did so by considering various combinations of the involved interlocutors: human writer / human reader; AI writer / human reader; human writer / AI reader; and AI writer / AI writer. Our evaluations were driven by close reading of the prompts and their answers.

Our experiments showed that, in terms of hermeneuticity, ChatGPT affirmed neither its positive nor negative polarities of hype. In all cases, the system generated readable texts that responded clearly to our prompts. While some responses were deemed trite, they did provide at least some meaning for the user, even if that meaning was only related to surface-level explanations of theory and/or to affirming the user's point of view. ChatGPT offered limited novel insight into the topics of discussion, with this insight being more common in response to vaguer prompts. Although more precise prompts increased ChatGPT's analytical precision, they also appeared to diminish the hermeneuticity of analysis. This result may indicate inherent limitations of ChatGPT: the system is programmed to limit speculation, follow instructions precisely, and present its ideas as distinctive, all of which work against the hermeneuticity of its responses. In ChatGPT 'conversations', which are largely one-way, the hermeneutic contract between readers and authors continues to be upheld as long as readers interact with the system.

It is worth noting that these conclusions are dependent on the specific LLM of our choice (ChatGPT). Other studies have confirmed that LLMs show variance in their (often strong) stereotypical biases (Nadeem et al. 2021), indicating that further research is needed in analysing how hermeneutic value might change when applying a hermeneutics lens to other LLMs like RoBERTa, XLNet, or LaMDA. Similarly, our observations may be constrained by our specific choices as to how we implemented hermeneutic prompt engineering: what theories we chose, what information we included in our shots, and how we composed the prompts themselves. Future studies may attempt to overcome these constraints by investigating advanced methods for prompt engineering: for example, new systems that automatically choose examples from large textual corpora, which are appended to prompts as few-shots that are appropriate to the initial prompt (Liu 2022).

Additionally, future studies may depend upon more systematic operationalisation in textual evaluation by, for example, applying set criteria against human and AI output for direct comparison. These criteria could be informed by other hermeneutic theories. Special attention could also be paid to collaborative (i.e. human group) instances of meaning making, rather than to the individualised instances of our focus. As with the 'prompt battles' mentioned above, the practise of collaborative hermeneutic analysis of NLG texts could lead to new insights into the different ways in which hermeneutic cycles unfold, which may be difficult for individuals limited by their own perspectives to foresee. In other cases (e.g. Wikipedia), collaborative meaning-making has entailed joint discussions towards common understandings of texts written by others. These discussions are comparable to manual 'semantic parsing' (i.e. translating natural language into some form of knowledge representation that can help collaborators disambiguate and clarify meaning), converging the efforts of many into a single, streamlined text.

Generally, our experiments confirm a need for the moderated optimism that the hermeneutic contract brings into current discussions about NLG. On one hand, we have seen unabashed hype touting AI's immense potential for readable text generation; this potential was affirmed in our experiments, wherein ChatGPT produced output in readable and appropriate language. On the other hand, we have seen intense pessimism directed towards AI text generation and its unoriginality; this unoriginality was shown in ChatGPT's plagiaristic use of extant textual analyses. In this paper, we have shown that readers are always making meaning from text, regardless of whether that text is produced by a human or AI. When readers are able to request the generation of texts themselves, they may use prompt engineering to achieve output that is meaningful to them in some way, whether by making them laugh, inspiring new ideas, helping them understand a subject, or fulfilling any other purpose. Meaning comes from whatever readers will make out of the answers to their prompts. However, following James Vincent's argument for chatbots as mirrors, human users still appear to be squinting to recognise themselves within their own creations; hence hyperbolic fears like Kirschenbaum's 'textpocalypse'. By attending to the hermeneuticity of ChatGPT and its necessary subjectivity, we have considered here the individualised experiences of users. Our approach complements the substantial body of work that applies quantitative measures of evaluation to LLMs. Yet these quantitative measures alone are insufficient. If LLMs are mirrors, we must consider the humans they are mirroring: humans who exist within totalities that cannot be wholly reduced to numbers.

References

No author (n.d.) Prompt Battle. Available from: https://promptbattle.xyz. Accessed 16 January 2023

Alexander S (2022) A Guide To Asking Robots To Design Stained Glass Windows. Astral Codex Ten. https://astralcodexten.substack.com/p/a-guide-to-asking-robots-to-design. Accessed 26 October 2022

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York: 610-623. https://doi.org/10.1145/3442188.3445922

Borup M, Brown N, Konrad K, Van Lente H (2006) The sociology of expectations in science and technology. Technology Analysis & Strategic Management. 18(3-4): 285-298. https://doi.org/10.1080/09537320600777002

Chomsky N, Roberts I, Watumull J (2023) Noam Chomsky: The False Promise of ChatGPT. New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html. Accessed 15 March 2023

Christiansen J (2011) Sloth - A Tool for Checking Minimal-Strictness. In: PADL 2022: International Symposium on Practical Aspects of Declarative Languages. Springer-Verlag, Berlin: 160-174. https://doi.org/10.1007/978-3-642-18378-2_14

Cobley P, Sibers J (2021) Close reading and distant: between invariance and a rhetoric of embodiment. Language Sciences 84: 101359. https://doi.org/10.1016/j.langsci.2021.101359.

EduKitchen [Sprakel T in conversation with Chomsky N] (2023) Chomsky on ChatGPT, Education, Russia and the unvaccinated. YouTube. https://www.youtube.com/watch?v=IgxzcOugvEI. Accessed 1 March 2023

Edwards B (2023) AI-powered Bing Chat spills its secrets via prompt injection attack. Ars Technica. https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack. Accessed 15 February 2023

Eisikovits N, Feldman D (2021) AI and Phronesis. Moral Philosophy and Politics. 9(2): 181-199. https://doi.org/10.1515/mopp-2021-0026

Evans A, Heimann A (2022) AI Activity in UK Businesses. Capital Economics, for the Department for Digital, Culture, Media, and Sport, London

Ethayarajh K, Jurafsky D (2020) Utility is in the eye of the user: A critique of NLP leaderboards. arXiv. https://arxiv.org/pdf/2009.13888. Accessed 1 March 2023

Fletcher A, Larson EJ. (2022) Optimizing Machines Is Perilous. Consider 'Creatively Adequate' AI. Wired. https://www.wired.com/story/artificial-intelligence-data-future-optimization-antifragility. Accessed 10 February 2022

Frankfurt HG (2005) On Bullshit. Princeton University Press, Princeton

Gadamer H-G (2004) Truth and Method, 2nd, rev edn., trans Weinsheimer J, Marshall DG. Continuum, London

Gambino A, Fox J, Ratan RA (2020) Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. Human-Machine Communication 1: 71-85. https://doi.org/10.30658/hmc.1.5

Gonen H, Iyer S, Blevins T, Smith NA, Zettlemoyer L (2022) Demystifying Prompts in Language Models via Perplexity Estimation. arXiv. https://doi.org/10.48550/arXiv.2212.04037. Accessed 13 January 2023

GPT-3. (2020). A robot wrote this entire article. Are you scared yet, human? The Guardian. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3. Accessed 2 January 2022

Henrickson L (2021) Reading Computer-Generated Texts. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108906463

Henrickson L, Meroño-Peñuela A (2022). The Hermeneutics of Computer-Generated Texts. Configurations 30(2): 115-139

HM Government (2021) National AI Strategy. Office for Artificial Intelligence, London. https://www.gov.uk/government/publications/national-ai-strategy. Accessed 1 November 2021

Heidegger M (1996) Being and Time: A Translation of Sein und Zeit, trans Stambaugh J. State University of New York Press, Albany

Hidalgo CA, Orghian D, Albo-Canals J, de Almeida F, Martin N (2021) How Humans Judge Machines. MIT Press, Cambridge, MA

Holdsworth C, Mogan D (2007) Revisiting the Generalized Other: An Exploration. Sociology. 41(3): 401-417.

Kirschenbaum M (2023) Prepare for the Textpocalypse. The Atlantic. https://www.theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318. Accessed 13 March 2023

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large Language Models are Zero-Shot Reasoners. arXiv. https://arxiv.org/pdf/2107.13586.pdf. Accessed 28 October 2022

Liu J (2022) LlamaIndex. https://github.com/jerryjliu/gpt_index. Accessed 10 March 2023

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv. https://arxiv.org/pdf/2107.13586. Accessed 16 March 2023

Marcus G (2023) Inside the Heart of ChatGPT's Darkness. The Road to AI We Can Trust. https://garymarcus.substack.com/p/inside-the-heart-of-chatgpts-darkness. Accessed 3 March 2023

Milne G (2020) Smoke and Mirrors: How Hype Obscures the Future and How to See Past It. Robinson, London

Nadeem M, Bethke A, Reddy S (2021) StereoSet: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 1: 5356–5371. ACL, Online. http://dx.doi.org/10.18653/v1/2021.acl-long.416

Nass C, Steuer J, Tauber ER (1994) Computers are social actors. In: Proceedings of SIGCHI '94 Human Factors in Computing Systems: 72-78. ACM, Boston

Nass C, Moon Y (2000) Machines and Mindlessness: Social Responses to Computers. Journal of Social Issues 56(1): 81-103

Natale S (2021) Deceitful Media: Artificial Intelligence and Social Life After the Turing Test. Oxford University Press, Oxford.

Natale S, Henrickson L (2022) The Lovelace effect: Perceptions of creativity in machines. New Media & Society. https://doi.org/10.1177/14614448221077278

Newport C (2016) Deep Work: Rules for Focused Success in a Distracted World. Pitakus, London

OpenAI (2022) ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt. Accessed 25 February 2023

Orekhov B, Fischer F (2020) Neural reading: Insights from the analysis of poetry generated by artificial neural networks. Orbis Litterarum 75(5): 230-246. https://doi.org/10.1111/oli.12274

Postman N (1993) Technopoly: The Surrender of Culture to Technology. Vintage, New York

Reeves B, Nass C (1996) The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge University Press, Cambridge

Ricoeur P (1991) Narrative Identity. Philosophy Today 35(1): 73-81

Robertson A (2022) Professional AI whisperers have launched a marketplace for DALL-E prompts. The Verge. https://www.theverge.com/2022/9/2/23326868/dalle-midjourney-ai-promptbase-prompt-market-sales-artist-interview. Accessed 22 September 2022

Sadowski J (2018) Potemkin AI: Many instances of 'artificial intelligence' are artificial displays of its power and potential. https://reallifemag.com/potemkin-ai. Accessed 15 February 2023

Sadowski J (2020) Too Smart: How Digital Capitalism is Extracting Data, Controlling Our Lives, and Taking Over the World. MIT Press, Cambridge, MA

Schmidt FA, Schmieg S (2022) Prompt Battle. https://promptbattle.com. Accessed 16 January 2023

Marche S (2022) The College Essay Is Dead. The Atlantic. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371. Accessed 17 March 2023

Vincent J (2023) Introducing the AI Mirror Test, which very smart people keep failing. The Verge. https://www.theverge.com/23604075/ai-chatbots-bing-chatgpt-intelligent-sentient-mirror-test. Accessed 1 March 2023

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv. https://arxiv.org/pdf/1804.07461.pdf. Accessed 13 March 2023

Warnke G (2011) The Hermeneutic Circle Versus Dialogue. The Review of Metaphysics 65(1): 91-112. https://www.jstor.org/stable/23055684

Warzel C (2023) The Most Important Job Skill of This Century. The Atlantic. https://www.theatlantic.com/technology/archive/2023/02/openai-text-models-google-search-engine-bard-chatbot-chatgpt-prompt-writing/672991. Accessed 17 February 2023

Weizenbaum J (1976) Computer Power and Human Reason: From Judgment to Human Reason. W. H. Freeman and Company, New York

Wingard J (2023) ChatGPT: A Threat To Higher Education? Forbes. https://www.forbes.com/sites/jasonwingard/2023/01/10/chatgpt-a-threat-to-higher-education. Accessed 17 March 2023

Winograd T, Flores F (2000) Understanding Computers and Cognition: A New Foundation for Design, 15th print. Addison-Wesley, Boston

Youn S, Jun SV (2021) ‘In A.I. we trust?’ The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging ‘feeling economy’. Computers in Human Behavior 119. https://doi.org/10.1016/j.chb.2021.106721