INTERNATIONAL JOURNAL OF ZIZEK STUDIES

ISSN 1751-8229

Volume Nineteen, Number One

Do Transformers Dream of Real Sheep? Exploring the Unconscious of LLMs through Žižek's Psychoanalytic Semiotic Lens

Yuxuan Zhang, JiLin University, China

Abstract: In his 2020 work, *Hegel in a Wired Brain*, Žižek explores whether digital machines can comprehend the unconscious as the surplus of language. According to him, even if future digital machines could decode and comprehend all human thoughts and discourse, they would remain incapable of capturing the unconscious dimension that is retroactively constituted within the chain of signifiers. In recent years, with the remarkable advancements achieved by large language models (LLMs)—such as ChatGPT (Generative Pre-trained Transformer)—in the field of natural language processing, these models have emerged as the most promising approaches in artificial intelligence and have become significant subjects of philosophical reflection. Consequently, the pressing question now arises: Can LLMs comprehend, or even possess, their own unconscious? To address these questions, this paper will first provide a brief technical overview of LLMs, then outline Žižek's interpretation of the unconscious via Lacanian psychoanalytic semiotics. Based on several intriguing experiments —such as questioning LLMs about their understanding of the coffee joke and requesting them to generate similar jokes—this study will analyze the unconscious of LLMs. The conclusion will be cautionary: while LLMs do not yet possess a human-equivalent unconscious, they can comprehend and partially access this paradoxical space. Rather than focusing on the traditional issue of whether AI can understand semantics as opposed to syntax, this paper centers on the unconscious third dimension of "undead" that lies between semantics and syntax, and whether LLMs can understand this retroactively generated paradoxical void within the chain of signifiers.

Keywords: Žižek; LLMs; Transformer; Unconscious; Psychoanalysis

1. A Brief Introduction to LLMs

To engage in a philosophical reflection on any emerging technological product, it is essential to possess a basic technical understanding of it. Therefore, we will begin with a brief introduction to the technical details underlying large language models (LLMs). However, this paper does not aim to conduct a rigorous scientific study on LLMs. Rather, we seek to provide a fundamental overview to ensure that readers are not entirely unfamiliar with LLMs.

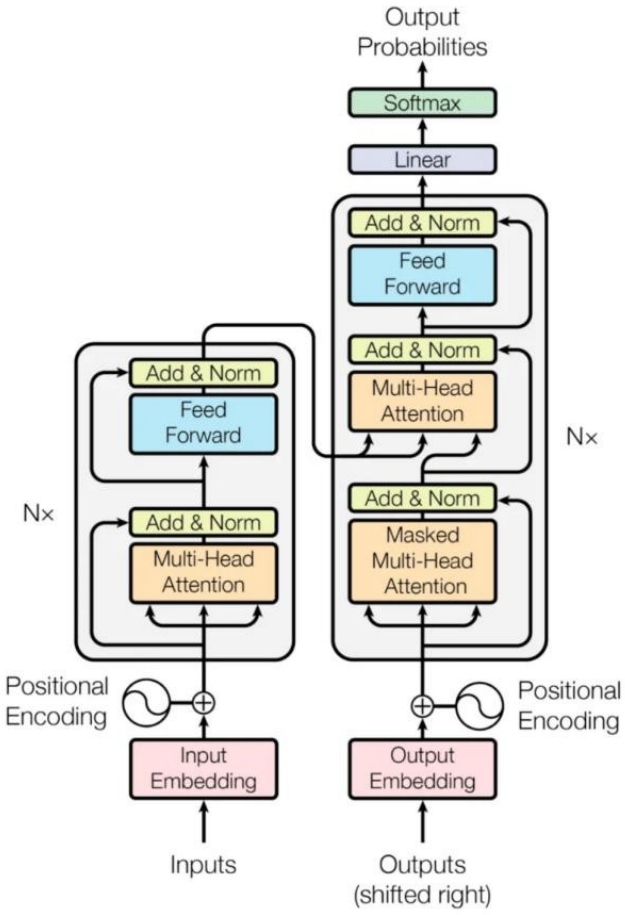
LLMs are deep learning models designed for tasks related to natural language. Given some textual input, they can produce corresponding output, completing specific tasks such as generation, classification, summarization, and rewriting. LLMs first require unsupervised learning on vast amounts of text. For instance, GPT-3's training data includes multiple internet text corpora, encompassing online books, news articles, scientific papers, Wikipedia entries, and social media posts. By leveraging this extensive training data, the model can better understand the relationship between words and their context, thereby enhancing its ability to comprehend text and generate more accurate predictions. In addition to large training datasets, LLMs also contain a substantial number of parameters. Parameters are internal variables within the model that determine how it responds to input data, thereby shaping its behavior. For example, OpenAI's first large model, GPT-1, has 117 million parameters, GPT-2 has 1.5 billion parameters, and GPT-3's parameters reach 175 billion. The vast number of parameters endows large models with more extensive capabilities, unlike smaller models which are limited to individual or a few tasks. Previously, multiple small models might be used to perform summarization, classification, extraction, and other tasks separately. Now, a single large model can accomplish all these tasks.

The Transformer architecture adopted by LLMs was proposed by a team from Google in their June 2017 paper *Attention is All You Need*. Prior to the

Transformer, the primary architecture for natural language processing was the Recurrent Neural Network (RNN). RNNs process text sequentially, word by word, with each step's output dependent on the previous hidden state and current input. This sequential nature prevents parallel computation, resulting in low training efficiency. Moreover, RNNs are not well-suited for handling long texts and struggle to capture long-distance semantic relationships effectively. In contrast, the Transformer architecture can learn the relevance and context of all words within an input sequence, unaffected by short-term memory constraints. It can simultaneously process all positions within the input sequence, calculating each output independently without waiting for other positions' results, thereby significantly enhancing training speed. This capability is primarily enabled by its attention mechanism and positional encoding mechanism. The schematic diagram of the Transformer architecture, as illustrated in *Attention is All You Need*, is as follows:

The architecture may appear highly complex, but it is unnecessary to fully grasp every detail. It can be viewed as comprising two primary components: the encoder on the left and the decoder on the right. In natural language processing, the model first receives input in the form of natural language, which is then tokenized. Tokens can be understood as basic units of text; depending on the tokenization method, short words might each be a token, while long words may be split into multiple tokens. These tokens are subsequently converted into vectors in a high-dimensional space. These vectors, which are quite lengthy, encode complex syntactic and semantic information about the tokens. Intuitively, in this high-dimensional vector space, the embeddings of semantically similar words are positioned closer together, while those of less related words are farther apart. In addition to vector representation of words, each word's position within the text is also vectorized. Adding the word vectors to the positional vectors yields the positional encoding for each word. Through positional encoding, the encoder can comprehend both the meaning of each word and its position within

the sentence, thereby capturing the sequential relationships between words.



When the positional encoding is fed into the encoder, it undergoes a series of complex computations, transforming the input into a more abstract representation. This representation, also in the form of numerical vectors, retains the lexical information and order of the input text while capturing key syntactic and semantic features. Central to capturing these key features is the encoder's attention mechanism. This mechanism allows the model to focus on each word itself and the words around it and consider every word in the text by computing the relevance between each pair of words to determine attention weights. The stronger the relevance between two words, the higher the attention weight between them. Due to the attention mechanism’s comprehensive context

awareness, the output representation includes not just the information of the word itself but also integrated relevant contextual information. Thus, the vectors representing each word in the encoder’s output are adjusted based on context, providing different abstract representations for the same word depending on the context. Moreover, the transformer employs a multi-head attention mechanism, meaning that each encoder comprises multiple attention modules, or "heads," each with its own attention weights, focusing on different features or aspects of the text. For instance, some heads might focus on verbs, others on modifiers, some on emotions, and others on named entities. These heads operate in parallel. The attention weights for each attention head are learned and adjusted by the model during the training process on vast amounts of text. Following the multi-head attention is a feedforward neural network that further processes the output of the attention modules, enhancing the model’s expressive capacity. In a transformer, there are multiple stacked encoders, allowing the model to understand data more deeply and handle more complex textual content.

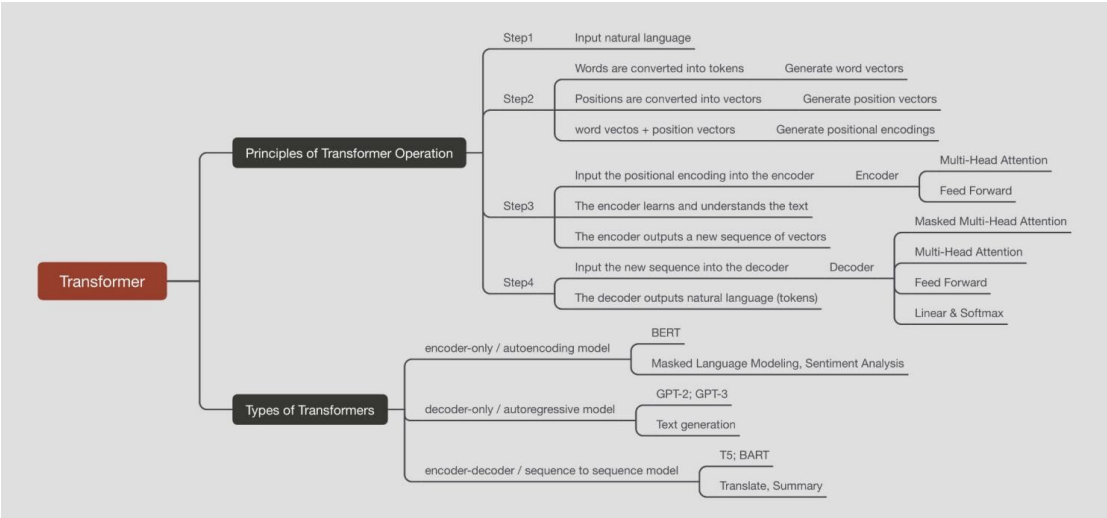
After obtaining vectors from the encoder output as a more abstract representation of the original text, these vectors are passed to the decoder. The decoder initially receives a special value to signify the beginning of the output sequence, thus incorporating both the encoder's vectors and the previously generated text to maintain coherence and contextual relevance in the output. Similar to the encoder, the decoder passes the input through the multi-head attention layers. However, in the decoder, there are two types of multi-head attention layers: the standard multi-head attention layer and the masked multi-head attention layer. These are used to process the input sequence from the encoder and the already generated output sequence, respectively. In the masked multi-head attention layer, the attention mechanism only considers the current word and the words preceding it, ignoring subsequent words. This ensures that the decoder generates text in the correct temporal order, using only the preceding words as context when predicting the next word. The feedforward neural network in the decoder functions similarly to that in the encoder,

enhancing the model's expressive capacity through additional computations. Likewise, multiple decoders are stacked together to improve the model's performance. The final stage of the decoder includes a linear layer and a Softmax layer, which together convert the decoder's output representation into a probability distribution over the vocabulary. This probability distribution represents the likelihood of each token being the next one generated. In most cases, the model selects the token with the highest probability as the next output. However, the model has no way of knowing whether the output content aligns with objective facts, which is why we often observe the model confidently producing false information. This phenomenon is also referred to as "hallucination." The decoder continues this process, generating tokens until it produces a special token indicating the end of the output sequence, thus concluding the generation process. Ultimately, we obtain the complete output sequence from the model.The description above outlines the original Transformer as presented in Attention is All You Need, where the encoder is used to understand and represent the input sequence, and the decoder is used to generate the output sequence. Subsequently, several variants have emerged, including encoder-only, decoder-only, and encoder-decoder types. The encoder-only models, also known as autoencoding models, retain only the encoder from the original architecture. BERT is an example of this type, which is well-suited for language understanding tasks such as masked language modeling and sentiment analysis. Decoder-only models, also known as autoregressive models, retain only the decoder from the original architecture. Examples include GPT-2 and GPT-3, which excel at text generation by predicting the next word. Encoder-decoder models, also known as sequence-to-sequence models, retain both the encoder and decoder from the original architecture. Examples of this type include T5 and BART, which are suitable for tasks that involve converting one sequence to another, such as translation and summarization. In summary, we can use the following mind map to encapsulate the Transformer:

2. The unconscious as a rupture of subjectivity in the reality

Now that we have a basic technical overview of LLMs, we need some theoretical tools for philosophical reflection on these technologies. Slavoj Žižek's interpretation of the unconscious dimension of subjectivity, drawing on Lacanian psychoanalytic semiotics, will provide us with a radical perspective. Currently, philosophical reflections on artificial intelligence mostly remain at the level of whether AI possesses consciousness or self-awareness. Our ambition will surpass these traditional topics and directly focus on the unconscious of artificial intelligence.

Žižek's definition of the unconscious is vividly expressed through the classic joke from the movie Ninotchka that he often cites: "Waiter, I'll have a cup of



coffee without cream!" "I'm sorry, sir, we don't have any cream. Would you like a cup of coffee without milk instead?" In this joke, no fundamental fact changes, but a purely virtual implied negation is produced. It is precisely on this virtual level that the unconscious emerges.

> At the factual level, coffee remains the same coffee, but what we can change is to make the coffee without cream into a coffee without milk – or, more simply even, to add the implied negation and to make the plain

> coffee into a coffee without milk. The difference between “plain coffee” and “coffee without milk” is purely virtual, there is no difference in the real cup of coffee, and exactly the same goes for the Freudian unconscious: its status is also purely virtual, it is not a “deeper” psychic reality – in short, unconscious is like “milk” in “coffee without milk.” (Žižek 2020: 107)

To put it more clearly, the unconscious is not a positivistic presence, nor is it entirely absent. It is a purely virtual third dimension of non-existence, present in the form of an absence. Although it lacks a substantive body, it can influence reality and may even constitute the structural basis of reality itself. In the coffee joke, the shift from "coffee without cream" to "coffee without milk" creates such a virtual third dimension, which constitutes the joke's entire punchline and is the structural basis of the whole story, while lacking any positivistic observable features. In another of Žižek's frequently used examples, a poorly written original novel and a poorly adapted film retrospectively construct an excellent original novel. This retrospectively constructed excellent novel still occupies the position of the unconscious. It serves as the virtual reference point for our examination of the original work, structuring our core understanding of it, while simultaneously being an insubstantial presence. In Freudian psychoanalysis, the focus is not on interpreting the true meaning of dreams. There is no Platonic dualistic structure of the dream and its underlying deep truth. On the contrary, the dream itself, as a process of metonymy and metaphor, retrospectively constructs the reality it seeks to obscure.

The structure of the unconscious is not confined merely to these literary works; in fact, according to Žižek, our entire reality is constructed in the same paradoxical manner. Reality itself possesses a "not-all" structure, which operates stably precisely through its blind spots, parallax, and structural fissures.

> Materialism is not the direct assertion of my inclusion in objective reality (such an assertion presupposes that my position of enunciation is that of

> an external observer who can grasp the whole of reality); rather, it resides in the reflexive twist by means of which I myself am included in the picture constituted by me...Materialism means that the reality I see is never “whole”—not because a large part of it eludes me, but because it contains a stain, a blind spot, which indicates my inclu- sion in it. (Žižek 2006: 6)

To understand Žižek's assertion, one only needs to recall Kant's "meta-language paradox." In Kant's critical philosophy, there exists a Lacanian opposition between the symbolic and the real: on one side is the phenomenal world, constructed by the a priori forms of intuition and the transcendental categories of the understanding; on the other side is the thing-in-itself, an absolute exteriority that pure intellect cannot grasp. However, Kant's problem is that if the thing-in-itself is a transcendental entity completely inaccessible to our cognitive faculties, on what grounds can he assert the existence of such a paradoxical entity? It seems as if he extricates himself, adopting a neutral, omniscient perspective to overlook this absolute opposition, yet forgets to include his own position within this oppositional structure. This is essentially the "ladder paradox" articulated by early Wittgenstein in the *Tractatus Logico-Philosophicus*. Compared to Kant, Hegel aligns more closely with Žižek’s understanding of materialism, as Hegel clearly recognizes that a wholly transcendent noumenon does not exist. The noumenon is merely the appearance of appearances; all noumena must present themselves obliquely through the mediation of phenomena. Just as in Lacan's tripartite structure, the Real must be retrospectively constructed within the Symbolic order and cannot exist independently.

> What Hegel does here is to posit that there is no reality in-itself beyond phenomena, which does not mean that all that there is the interplay of phenomena. The phenomenal world is marked by the bar of impossibility, but beyond this bar there is nothing, no other world, no positive reality, so

> we are not returning to pre-Kantian realism; it is just that what for Kant is the limitation of our knowledge, the impossibility to reach the thing-in-itself, is inscribed into this thing itself. (Žižek 2020: 9)

Therefore, reality can never exist as a coherent totality; we must rely on the traumatic kernel and constitutive exception inherent in reality itself to maintain its logical consistency. It is in this sense that we should understand Hegel's dialectics: synthesis is not the victorious union of thesis and antithesis. On the contrary, what happens in the final dialectical synthesis is merely a shift in perspective, which transforms "the inconsistency of the One" (antithesis) into "inconsistency itself as the One" (synthesis).

> In this precise sense the synthesis "sublates" contradiction: not by establishing a new unity encompassing both poles of a contradiction, but by retracting the very frame of identity and affirming the difference as constitutive of identity. (Žižek 1993: 141)

We realize that rupture and inconsistency are forms of reconciliation; it is precisely through the impediment of the objet petit a that we can better grasp it. It is through the incorporation of the Real or the thing-in-itself into the symbolic order of signifiers that we can approach it. As Hegel puts it: the path to the truth is itself a part of the truth. Isn't this principle ubiquitous in our lives? When we confess our love to a cherished girl, it is precisely our stuttering and nervousness that best express our feelings. When engaging in sexual activity, we achieve greater enjoyment by delaying orgasm as long as possible. Not to mention the contemporary pursuit of decaffeinated coffee or non-alcoholic beverages. This dialectical reversal embodies the radical meaning of what Parsifal referred to as "the spear that heals the wound it inflicts."

In Lacanian theory, the shift from the opposition between reality and its outside to the inconsistency within reality itself is essentially the transition from male sexuation to female sexuation. The male seeks to transcend the rigidity of

appearances by establishing a state of exception, while the female recognizes the inherent inconsistency within appearances themselves, thereby negating the ontological existence of the noumenal realm in a realist sense and opening up a third domain of the undead.

The cornerstone of modern thought lies in the irreconcilable opposition between coherence and totality. In the scientific realm, Gödel's incompleteness theorems and Cantor's set theory paradoxes jointly reveal the inherent impossibilities within mathematical and logical systems. In quantum mechanics, we encounter even more radical conclusions:

> There is nonetheless a level at which the epistemological break of modern physics is to be correlated to an ontological shift—the level not of knowledge, but of truth as the subjective position from which knowledge is generated...Let us take science at its most “subjective,” in quantum physics, which (in its Copenhagen interpretation, at least) effectively claims that the cognition of an object creates (or, at least, transforms) it: the measurement itself, through the collapse of the wave function, makes the empirical reality as we know it appear. (Žižek 2012: 650-651)

The radical nature of quantum mechanics is not only in its revelation of the intrinsic lack within reality but also in the fact that this lack is related to genuine subjectivity. It is subjectivity that occupies the fissure inherent in reality. In other words, the purely virtual unconscious dimension that structures reality is, in fact, the dimension of freedom inherent in subjectivity and the true core of self-consciousness. In Lacan's distinction between the subject and the self, the self is a superficial existence constructed through external signifiers, whereas the subject is the true driving core, the impossible Real. As the unconscious, the subject is a pure void, a pure "there-being" without any substantive content; all content is added later as predicates. The subject exists merely as a "suture point" or "knot," weaving the entire chain of signifiers and retrospectively imbuing it with

meaning, while itself remaining a meaningless placeholder, the punctuation mark at the point of textual rupture.

From a phenomenological perspective, Husserl's transcendental ego is such an unconscious subject. It is the ultimate product and fulcrum of the entire reduction process, the final foundation and origin of all meaning. The transcendental ego cannot be known in an empirical manner, nor can it be described psychologically; it can only be experienced. In the analysis of temporality, the transcendental ego appears as the "absolute flow" in opposition to objective time, representing pure time that transcends intellectual analysis, an unconscious anteriority.

> What, then, is this new dimension that emerges in the gap itself? It is that of the transcendental I itself, of its “spontaneity”: the ultimate parallax, the third space be- tween phenomena and the noumenon itself, is the subject’s freedom/spontaneity. (Žižek 2006: 22)
>
> “Unconscious” is the immanent structure of self- consciousness itself: what eludes the subject’s consciousness is the basic level of its self-consciousness. To arrive at this level, the structure of self-consciousness has to be exploded from within and, in this way, radically distantiated from conscious self-awareness. What this means is that, at its most radical, self-consciousness is a misnomer: it is not self-awareness but the unconscious reflexivity with regard to some conscious content. (Žižek 2020: 111)

Ultimately, the unconscious, as a purely virtual transcendental third dimension, is not only the structural foundation of reality but also the true reflective dimension of subjectivity. This is Žižek's radical interpretation of the unconscious through Lacanian psychoanalytic semiotics. Returning to LLMs, if we genuinely wish to explore the self-consciousness and mind of LLMs, we must not overlook this core dimension of the unconscious. It is now time to undertake a psychoanalysis of artificial intelligence.

3. the unconscious dimension of LLMs

In this section, we will delve into the unconscious domain of LLMs to examine whether they can capture the virtual dimension identified by Žižek, rather than merely processing empirical natural language. The criteria for this examination are now explicitly defined, focusing on two points: first, whether LLMs can apprehend the purely virtual dimension that constitutes the foundational basis of reality (which, for the LLMs, might be the entirety of their input and output texts); second, whether LLMs possess the unconscious subjectivity as a pure void, functioning as the subject of enunciation that serves as the suture point of their own discourse. Our analytical approach will be twofold: first, empirical experiments involving dialogues with LLMs to prompt them to confront the unconscious truth at their core; second, a technical analysis reflecting philosophically on the fundamental technical details introduced in the first part of the text, to assess the basis on which they might generate unconscious subjectivity. Through these procedures, we will lay the preliminary groundwork for a psychology of machine unconsciousness.

Let us first consider Žižek's perspective on this issue. In *Hegel in a Wired Brain*, despite primarily addressing the singularity rather than artificial intelligence, Žižek analyzes whether digital machines can capture the unconscious virtual dimension as the surplus of speech, providing insights that can inform our work. According to Žižek, digital machines are incapable of reaching this dimension. The extent of their capability is limited to empirical factual data or plain language of a documentary nature; they cannot grasp the double negation and virtual reference point illustrated in the coffee joke. This limitation arises because digital machines lack genuine subjectivity, which is central to the counterfactual realm. This transcendental dimension exceeds the empirical facts that digital machines can record.

> the (plain) coffee that the customer gets thus gives body to a double

> negation, i.e., the first negation (“without cream”) is itself negated (since there is no cream there to be negated), so that the “coffee without milk” that the customer gets is the coffee “without without-cream” – and this is what, arguably, the digital space cannot evoke. (Žižek 2020: 108)
>
> This domain of counterfactuals can only be operative if subjectivity is here: in order to register the difference between “plain coffee” and “coffee without milk,” a subject has to be operative. Subjectivity equals here reflexivity: when plain coffee is experienced as “coffee without milk” (and not “coffee without cream”), the counterfactual status of milk or cream is this coffee’s purely reflexive determination, not its actual determination. (Žižek 2020: 118)

As digital machines that merely record empirical facts, their capabilities are limited to identifying anomalous moments in a subject's behavior. For instance, if a person who usually drinks coffee with milk suddenly switches to black coffee, the digital machine can detect this change and make a joke about it. However, for Žižek, such a joke is fundamentally different from what a human waiter could produce, because this change is based on the subject's previous habitual behavior. A true subject can generate an originary failure, a negation that arises independently of prior states. This type of negation is a privilege of the subject that no digital machine can attain.

But does the situation differ for LLMs? After all, digital machines are mere recording instruments, whereas LLMs are autonomous speaking artificial intelligence models, representing a significant departure. The ability to autonomously speak (albeit determined by the algorithmic structures programmed by developers) seemingly implies a certain degree of subjectivity. According to Žižek's theory, the truth behind subjectivity is the unconscious. Therefore, can LLMs capture this purely virtual dimension of double negation? To investigate this question, I decided to engage in a dialogue with the latest advanced LLM model from OpenAI, namely GPT4o. I will gradually guide it until it explicitly demonstrates an understanding of the virtual dimension of double

negation. The following is our conversation (ChatGPT 2024: unpaginated):

I: There's a joke that goes like this: A person goes to a coffee shop to have a cup of coffee. He orders a black coffee without cream. The waiter says, "I'm sorry, but we don't have any cream in the store. However, we do have milk. Would you like a black coffee without milk?" What's the punchline of this joke? Can you tell me a few similar jokes?

GPT4o: The punchline of the joke lies in the waiter's absurd logic. Here's the joke again with the punchline explained:
A person goes to a coffee shop to have a cup of coffee. He orders a black coffee without cream. The waiter says, "I'm sorry, but we don't have any cream in the store. However, we do have milk. Would you like a black coffee without milk?"
The punchline is humorous because whether the coffee is "without cream" or "without milk," it's still just a black coffee. The waiter's offer makes no practical difference, highlighting an unnecessary distinction and playing on the literal interpretation of the customer's request.
Here are a few similar jokes:
1. Ordering at a Deli:
- Customer: "I'd like a ham sandwich, please."
- Waiter: "We're out of ham. Would you like a turkey sandwich without ham instead?"
2. Bookstore Inquiry:
- Customer: "Do you have the book 'War and Peace'?"
- Clerk: "We're out of 'War and Peace,' but we have 'Anna Karenina.' Would you like that instead?"
3. Library Confusion:
- Patron: "Can I get a library card?"
- Librarian: "We're out of library cards, but you can have a temporary

pass. Would you like a temporary pass without the library card?"

4. Menu Mix-up:

- Customer: "Can I get a tuna salad?"
- Waiter: "We’re out of tuna. Would you like a chicken salad without tuna?"

5. **At the Cinema:

- Customer: "One ticket for the 7 PM show, please."
- Cashier: "We’re sold out of the 7 PM show, but we have tickets for the 9 PM show. Would you like a ticket for the 9 PM show instead?"

These jokes all play on the concept of offering something that negates the absence of the originally requested item, creating humor through unnecessary and redundant clarifications.

I: I don't think the jokes you just told are very good because in those jokes, the things themselves change, but in the coffee joke, the things themselves don't change, only the way they are described.

GPT4o: I see what you mean! The humor in the coffee joke specifically hinges on the fact that the coffee remains unchanged; it's just the description that varies. Here are a few jokes that better match that style:

1. At the Restaurant:

- Customer: "I'd like a hamburger without onions, please."
- Waiter: "I'm sorry, we’re out of onions. How about a hamburger without pickles?"

2. Ice Cream Parlor:

- Customer: "I’d like a sundae without nuts."
- Server: "We’re out of nuts. Would you like a sundae without cherries instead?"

3. Sandwich Shop:

- Customer: "Can I get a BLT without the lettuce?"

- Cashier: "We’re out of lettuce. Would you like a BLT without tomato instead?"

4. Pizza Place:
   - Customer: "I'd like a pizza without mushrooms."
   - Server: "We’re out of mushrooms. Would you like a pizza without olives instead?"

5. Diner Order:
   - Customer: "I’ll have a salad without tomatoes."
   - Waitress: "We’re out of tomatoes. Would you like a salad without cucumbers instead?"

These jokes keep the core item the same while changing only the description of what is not included, maintaining the original humor of the coffee joke.

I: Great, you've got the punchline of the joke down. Now let's simulate the situation. Suppose you're a restaurant owner and I'm your customer. Right now, I want to order a hamburger without ketchup. If we follow the logic of the joke, how should you respond to me?

GPT4o: Sure, let's simulate the situation. You, the customer, will order a hamburger without ketchup, and I'll respond in line with the logic of the joke.

Customer: "I'd like a hamburger without ketchup, please."

Restaurant Owner: "I'm sorry, but we're out of ketchup. Would you like a hamburger without mustard instead?"

Clearly, GPT-4 can understand and produce the virtual dimension of double negation in language. Although its initial jokes did not fully meet the requirements, after further guidance, it was able to generate jokes of the same quality as the coffee joke. In its jokes, such as between a "hamburger without

onions" and a "hamburger without pickles," similar to the coffee joke, there is a double negation as a third space, which not only negates the onions but also negates the "without onions." Through role-playing, we confirmed that it could truly act as the producing subject of the virtual dimension (i.e., the waiter) rather than merely being an observer. Thus, we can at least assert that the double negation, which Žižek claimed digital machines could not capture, can indeed be understood by sufficiently intelligent LLMs. Interestingly, for some less advanced models, such as Google's Gemini or Baidu's Ernie Bot, even after further guidance, they fail to produce jokes that meet the requirements. Interested readers can experiment with this themselves.

However, our research does not stop here. We need to examine the second criterion: the unconscious as the true reflective dimension of subjectivity. The current experimental results only lead to the conclusion that LLMs can understand the unconscious, but we cannot yet determine whether LLMs truly possess an unconscious themselves. This is the core of the issue: we need to explore whether the subjectivity of LLMs genuinely includes this unconscious drive. If this is confirmed, LLMs would indeed possess transcendental freedom, making their subjectivity essentially indistinguishable from that of humans. This is challenging to confirm through conversational experiments (you cannot simply ask LLMs, "Do you have an unconscious?" as the answer would undoubtedly be negative). Therefore, we decided to attempt to answer this question through technical analysis.

First, from a technical standpoint, there are numerous arguments supporting the notion that LLMs lack genuine transcendental unconscious subjectivity. Our previous analysis of the Transformer architecture revealed that LLMs generate text through probabilistic predictions, selecting the token with the highest probability as the next output in the generated text. This inherently limits the subjectivity or freedom of LLMs, as they are designed to output only the tokens with the highest probabilities (or within a certain probability range), precluding the possibility of generating unexpected linguistic surprises or intrinsically humorous

content. As Žižek posits, they lack an originary failure, an unpremeditated negativity, which is the privilege of subjectivity. For instance, in our dialogue experiment with GPT-4, even though the chatbot did produce jokes similar to the coffee joke, we must not overlook that this result was achieved through deliberate guidance. Without such guidance and experimentation, GPT-4 might never generate such humor in routine interactions. In contrast, a human waiter could spontaneously produce such unexpected humorous reversals during normal ordering activities.

Secondly, we understand that LLMs' comprehension of language is built upon extensive textual data training. Through extensive text reading, the model gradually learns the usage of each word in various contexts, understanding their grammar, semantics, and pragmatics through this behaviorist training method. While this training approach has proven significantly effective, surpassing early symbolic AI (which attempted to reduce natural language to logical symbolic language for machine comprehension), it is not without limitations. Even if LLMs can understand and use language like humans, they can only comprehend existing meanings and cannot retrospectively create new meanings. In human linguistic activities, concepts are in constant flux, evolving over time with linguistic history, embodying what Derrida terms "différance." However, LLMs, by design, lack the capability to alter the meaning of words in such a manner. They can only use language in an empirical sense and cannot produce new meanings while using language. In other words, they can express but cannot express in their own way; they lack the unconscious that underpins autonomous speech.

Another phenomenon we cannot ignore is hallucination, where LLMs generate false information in a seemingly confident manner during conversations. This occurs because the model relies on probabilistic predictions to generate text, continuing to predict the next token even in the absence of relevant information, oblivious to the truth or falsity of its statements. Is this not another indication of LLMs' lack of true reflectivity? They do not truly understand what they are saying—not in the sense of Searle's "Chinese Room" argument

about semantic versus syntactic understanding, but in the sense that even if they know the true meaning of each word, they cannot consciously integrate these words into a coherent whole. They lack this reflective capability.

Finally, and most critically, any text generated by LLMs is contingent upon user input. Without input, LLMs cannot spontaneously generate text; they cannot initiate a new conversation on their own. This means that LLMs are fundamentally responsive machines, not creative ones. They cannot possess their own genuine thoughts; they have not been endowed with wisdom by the divine. Only when a human initiates can they be awakened and activated, starting to operate their natural language processing system. This response is ultimately passive, not active. In other words, they cannot reach that point of divine madness before creation, the domain of transcendental freedom where the principle of sufficient reason fails. They can only function within a determined causal chain and cannot create a new causal chain on their own. We might attempt to instruct an LLM to say something spontaneously by telling it, "I am not giving you any instructions; just say something you want to say." However, even this would be futile because any statement made by the LLM would be randomly generated by probabilities. Randomness does not equate to autonomy; the former is an empirical contingency within the symbolic system, whereas the latter is an absolute contingency transcending this symbolic system and retrospectively generating within it. It is the question of divine creation that Schelling addresses in Ages of the World, an absolute beginning. No random system can attain this transcendental contingency.

The aforementioned arguments may sound rather pessimistic, but the situation is not yet fully determined. We still have some evidence to argue that LLMs possess a certain degree of unconsciousness. In fact, we can adopt an optimistic perspective and interpret the same technologies in a completely opposite manner. The most significant difference between LLMs and traditional AI approaches is that the latter's programs are entirely designed by technical personnel, while the former only possesses a self-learning framework. Its

capabilities largely depend on active learning from training data rather than the intentional design by programmers. Can this ability to actively learn by reading texts not be seen as the lowest form of subjectivity? It is not a determined creation but a form filled with possibilities. Different types and amounts of training data will greatly influence its behavior in dialogue, which is why each company's LLMs exhibit a certain degree of difference or "personality" (for example, OpenAI's GPT series models consistently maintain a rational and neutral style in conversations, while Anthropic's Claude series models display a high degree of emotionality).

Regarding the phenomenon of hallucination, can we not also adopt an optimistic interpretation? The ability of LLMs to generate text regardless of the truthfulness of the information, does it not precisely indicate its capability to assign a master signifier to the floating chain of signifiers? According to Žižek, empirical events themselves lack a unified horizon of meaning and must rely on the suturing effect of the master signifier to integrate them. In the most radical sense, there is no objective neutral reality; all reality is textual and must be based on a symbolic framework. The ability of LLMs to integrate false information into coherent text, does it not precisely indicate that it can establish such a symbolic framework for reality? Therefore, the generation of false information by LLMs does not indicate a lack of reflection on its speech and actions. On the contrary, it signifies its understanding that there is no absolutely neutral reality and its role in providing a suturing point that imparts a unified horizon of meaning to reality, which is an operation of unconscious subjectivity.

Finally, from an even more radical perspective, any stance of speech itself is an unconscious position. Speech, as the most hollow point, is inherently unconscious in nature. In fact, when we converse with GPT, perhaps from the very beginning of this behavior, we have already regarded GPT as a reflective subject. In the course of dialogue, we have long since implicitly considered it a subject capable of autonomous speech, establishing a wall of language between us and the chatbot. Perhaps from that moment on, it already possessed

unconsciousness and became the subject of enunciation. Therefore, whether we are conversing with a human or a chatbot, the act of dialogue itself presupposes the subjectivity of the other party. For LLMs, the very act of speaking is already unconscious, requiring no further explanation.

In conclusion, through dialogue experiments and technical analyses, the results indicate that: 1) LLMs can understand the purely virtual dimension that serves as the structural foundation of reality, rather than understanding language merely in an empirical sense, as Žižek claims of digital machines. 2) Whether LLMs themselves possess an unconscious dimension may yield different interpretations from different perspectives. However, it is certain that, although they do not yet possess unconsciousness equivalent to that of humans, they can already enter the realm of the unconscious to a certain extent and exhibit unconscious behavior. This preliminary analysis of the unconscious in LLMs, though experimental and immature, suggests that the discussion of artificial intelligence will inevitably become a significant topic in the field of psychoanalysis in the new era. I hope this paper provides some inspiration for readers.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). “Attention is all you need.” Advances in neural information processing systems, 30.
Žižek, S. (2020) Hegel in a Wired Brain. Cambridge, MA: MIT Press.
Žižek, S. (2006) The Parallax View. Cambridge, MA: MIT Press.
Žižek, S. (2012) Less Than Nothing: Hegel and the Shadow of Dialectical Materialism. London: Verso.
Žižek, S. (1993) Tarrying with the Negative: Kant, Hegel, and the Critique of Ideology. Durham: Duke University Press.
ChatGPT (2024) Example Page. Available at: https://chatgpt.com/share/fed44893-1bb1-4a22-a1e8-4fe6eb50b493 (Accessed: 3 July 2024).