Is the ‘Calculator for Words’ analogy useful for communicating about LLMs?

Dr Brian Ballsun-Stanton Dr Inês Hipólito

This paper is available as a preprint

Feedback would be *delightful*.

- This presentation is CC-BY.
- Code available at https://github.com/Denubis/calculator-for-words-presentation
- Presentation available at: https://denubis.github.io/calculator-for-words-presentation/

Figure 1: doi.org/10.5281/zenodo.12602858

Understanding the capabilities of technology is not a new problem

Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out? (Babbage 1864)

Understanding the capabilities of technology is not a new problem

> *Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out? (Babbage 1864)*

For an LLM: the answer is **YES**.

> *Large language models (LLMs) are fundamentally different from search engines, functioning more as 'vibe-machines' than information retrieval systems. (Ballsun-Stanton and Hipólito 2024)*

Motivation and Aim

- Why are we here: The need for effective LLM comprehension.
- Is the ‘calculator for words’ analogy the most effective one?
- How do we teach, communicate, and write policy about these tools without effective analogies to communicate their major capabilities and limitations?

Our goal

Find an effective analogy to communicate LLM affordances to avoid error-prone usage patterns.

The Problem

- Too many conflicting stories about “AI”
- Anthropomorphization: They respond using “I”
- Skeuomorphic lenses vs. understanding true limitations and capabilities

For ‘Calculators for words’

Willison on LLMs

One of the most pervasive mistakes I see people using with large language model tools like ChatGPT is trying to use them as a search engine. ... I like to think of language models like ChatGPT as a calculator for words. ... Want them to work with specific facts? Paste those into the language model as part of your original prompt! (Willison 2023)

Unreliability and Locus of Control

The analogy ‘calculator for words’:

- Moves locus of control to the user’s perspective
- User input is primary, not the ‘creative’ output of the machine
- Grounding inputs to reduce confabulation
- Maps to prior affordances users expect

Difficulty of Effective Use

- Using LLMs effectively is deceptively difficult
- Requires building an accurate mental model of capabilities and limitations
- Users need to spend time with LLMs to understand their potential and pitfalls

LLMs vs. Search Engines

Argument 2.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If an AI system lacks inherent intentional agency in its core operation (token inference) and its outputs are primarily bounded by human-defined service layers, then it is fundamentally different from search engines that link to **intentionally created content**.

LLMs vs. Search Engines

Argument 2.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If an AI system lacks inherent intentional agency in its core operation (token inference) and its outputs are primarily bounded by human-defined service layers, then it is fundamentally different from search engines that link to **intentionally created content**.

P2. LLMs **lack inherent intentional agency** in their core operation (token inference), and their outputs are primarily determined by human-defined service layers (system prompts and post-hoc interactions applied to token inference).

LLMs vs. Search Engines

Argument 2.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If an AI system lacks inherent intentional agency in its core operation (token inference) and its outputs are primarily bounded by human-defined service layers, then it is fundamentally different from search engines that link to **intentionally created content**.

P2. LLMs **lack inherent intentional agency** in their core operation (token inference), and their outputs are primarily determined by human-defined service layers (system prompts and post-hoc interactions applied to token inference).

C. Therefore, LLMs are fundamentally different from search engines that link to intentionally created content.

Against ‘Calculators for words’

Bucci's Response

To put it differently, a calculator has a well-defined, well-scoped set of use cases, a well-defined, well-scoped user interface, and a set of well-understood and expected behaviors that occur in response to manipulations of that interface. ... Large language models, when used to drive chatbots or similar interactive text-generation systems, have none of those qualities. They have an open-ended set of unspecified use cases. (Bucci 2023)

Determinism and Affordances

Theorem 3.1 (Ballsun-Stanton and Hipólito 2024):

A calculator, upon receiving invalid input, will give the user an error. A LLM, upon receiving invalid input, will give the user a valid response.

- LLMs lack the deterministic nature of calculators
- The interface of LLMs is deceptively simple, hiding complex and often unpredictable behavior
- Affordances of LLMs are vastly different from those of calculators

“ChatGPT is Bullshit” and Appropriateness of Use

- LLMs generate text without adherence to an underlying reality (Hicks, Humphries, and Slater 2024)
- Output is a stream of tokens with the highest statistical likelihood, appearing as coherent thought
- Unlike calculators, which have clear appropriate use cases, LLMs’ appropriate applications are still being defined
- **Therefore:** The analogy fails to capture the open-ended nature of LLM interactions and outputs

Discussion

The Pragmatics of Experience

> *When I speak in front of groups and ask them to raise their hands if they used the free version of ChatGPT, almost every hand goes up. When I ask the same group how many use GPT-4, almost no one raises their hand. I increasingly think the decision of OpenAI to make the "bad" AI free is causing people to miss why AI seems like such a huge deal to a minority of people that use advanced systems and elicits a shrug from everyone else. (Mollick 2023)*

- The gap in user experience between different LLM versions affects perception and understanding
- This disparity influences how we communicate about and conceptualize LLMs

Evaluating the ‘Calculator for Words’ Analogy

Intuitions:

- The analogy falls apart on deeper inspection
- Lacks useful ontological or epistemological similarities with LLMs

Evaluating the ‘Calculator for Words’ Analogy

Pedagogical utility:

- Useful for teaching attention to context window and token inference
- Helps adjust audiences’ epistemic orientation to LLM interfaces

Evaluating the ‘Calculator for Words’ Analogy

Comprehensibility:

- Requires specific interpretation to be useful
- May lead to misunderstandings if taken too literally

Maps of No Territory

Maps of No Territory: A New Analogy

- Inspired by Borges' "On Exactitude in Science" (Borges and Hurley 1998)
- LLMs as maps without a corresponding territory
- Aims to provide more nuanced intuitions for general audiences
  LLMs, as our proposed map-analogues, generate language based on statistical relationships learned from vast amounts of text, creating abstract representations of language patterns and structures. (Ballsun-Stanton and Hipólito 2024)

Traces on maps

- "Training" an LLM:
  - it stores the representations of the statistical relationships between tokens.
  - In our map-metaphor, we can call them **traces**.
- The relationships stored in the model's weights are a map of no territory.

A trace is a correspondence of a sign/token output by an LLM which has a referent useful to the user.

LLMs as Maps of No Territory

Argument 5.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If LLMs generate language based on abstract representations without direct access to real-world referents, then LLMs are like maps of no territory.

LLMs as Maps of No Territory

Argument 5.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If LLMs generate language based on abstract representations without direct access to real-world referents, then LLMs are like maps of no territory.

P2. LLMs generate language based on abstract representations without direct access to real-world referents.

LLMs as Maps of No Territory

Argument 5.1 (Modus ponens) (Ballsun-Stanton and Hipólito 2024):

P1. If LLMs generate language based on abstract representations without direct access to real-world referents, then LLMs are like maps of no territory.

P2. LLMs generate language based on abstract representations without direct access to real-world referents.

C. Therefore, LLMs are like maps of no territory.

Effective Interaction and Navigation

- Interacting with LLMs is like being a librarian in Borges' Infinite Library
- Requires skillful navigation to find meaningful content
- Three levels of (mental) mapping:
  1. Expertise Map: User's foundational understanding
  2. Incomplete but Sufficient Map: Framework for effective engagement
  3. Map of a Map of no Territory: Abstract representations within the LLM

Conclusion

- The ‘calculator for words’ analogy serves as an effective negative heuristic
  - Discourages treating LLMs as search engines or fact repositories
  - Falls short in providing positive intuition for effective LLM utilization
- ‘Maps of no territory’ offers a more comprehensive understanding
  - Captures the nature, capabilities, and limitations of LLMs
  - Encourages more informed and responsible engagement
- Effective use of LLMs requires:
  1. Skillful interpretation of traces provided by LLM interaction
  2. Developing a framework for effective engagement
  3. Understanding that outputs reflect training data, not direct representations of reality

References

Babbage, Charles. 1864. "Passages from the Life of a Philosopher." https://www.gutenberg.org/files/57532/57532-h/57532-h.htm.

Ballsun-Stanton, Brian, and Inês Hipólito. 2024. "Is the 'Calculator for Words' analogy useful for communicating about LLMs?" Zenodo. https://doi.org/10.5281/zenodo.12602858.

Borges, Jorge Luis, and Andrew Hurley. 1998. *Collected Fictions*. Penguin Classics Deluxe Edition. New York, NY: Penguin Books.

Bucci, Anthony. 2023. "Word Calculators Don't Add Up." https://bucci.onl/notes/Word-calculators-dont-add-up.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT Is Bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Mollick, Ethan. 2023. "An Opinionated Guide to Which AI to Use: ChatGPT Anniversary Edition."