Z: The New LLM by ZombieCorp. State of the Art on Everything!*

* a thought experiment

Do Zombies Understand? A Choose-Your-Own-Adventure Exploration of Machine Cognition

Ariel Goldstein, Gabriel Stanovsky, The Hebrew University of Jerusalem

We're excited to release Z, a **completely open source LLM**, including access to anything you may ask for: code, training data, learned weights, or hyperparameters. As shown in Table 1, Z **achieves state-of-the-art results in all current (and future) NLP benchmarks!**

Benchmark	Z 🧟
HELM	100%
MMLU	100%
Human-Eval	100%

Table 1. (Imagined) results for Z 🧟

Do you consider Z as capable of understanding?

Yes

Michael Dummett, 1976

*"if a **robot be devised to behave in just the ways that are essential to a language speaker**, an implicit knowledge of the **correct theory of meaning** for the language could **be attributed to the robot with as much right as to a human**"*

No

Alan Turing, 1950

*"Not until a machine can write a sonnet or compose a concerto because of **thoughts and emotions felt**, and not by the chance fall of symbols, could we agree that machine equals brain—that is, **not only write it but know that it had written it.**"*

Functional Understanding

Model Z will **functionally understand** A task T if its performance on T is good, or better than, a human who is an expert at the task

Research Agenda

Achieve **Super-human performance** on all tasks. AI adopted "drosophila" tasks (Chess, Go, NLI?) abandoned once functionally understood. (McCarthy, 1990).

Conscious Understanding

M **consciously understands** a task T if M *functionally understands* T and M is *conscious* **There is something that "it is like" to be M**

Research Agenda

Build consciousness into LLMs, e.g., via **cognitively-inspired architectures**, such as spiking neural networks (NCC; Koch et al., 2016, Tonini, 2016, Mediano et al., 2022).

Debate around Understanding in NLP isn't Productive Due to Terminological Disagreement

The heated debate around machine understanding (Mitchell and Krakauer,2022) is so far non productive, and the **field is at an impasse**. We argue that this happens because **proponents at either side hold different definitions for what it means to *understand*** (Bender and Koller, 2020, Bubeck et al., 2023). We lean on vast literature in philosophy and neuroscience, studying this phenomenon as **the mind-body problem** (Chalmers, 1995), leading to **two distinct research agendas**.

Other Possible Answers?

Whether Z understands depends on implementation *but it has nothing to do with conscious experience.*

This argument is in line with Block (1981)'s definition of *Psychologism*. We challenge this answer by considering that a discrete set of seemingly intelligent steps can be internally implemented in "non intelligent" ways.

The question is ill-posed as Z is inconceivable *it's meaningless to discuss different properties of Z.*

This argument may stem from the belief that consciousness has a function in understanding (Van Gulick, 2022), and hence it is impossible for an agent to excel on every NLP benchmark without also achieving consciousness.

Question doesn't define what's understanding *so different definitions may lead to different answers.*

We don't aim to define understanding, and do not argue that there is a single "correct" definition. We claim that answering the question elucidates different definitions for understanding. We invite engaing with this question to examine *your* definition for understanding