The Two Word Test: A Benchmark for Combinatorial Semantic Processing in Large Language Models

Nicholas Riccardi and Rutvik Desai

**Introduction**

Combining multiple words into a single, coherent concept is a vital feature of human language and intelligence. For example, we know that ‘the beach ball’ is a sensible phrase, while ‘the ball beach’ is not. In humans, some of these phrases are learned as single units that combine two seemingly unrelated words (e.g., the knee cap), while others are more compositional and built ‘from the ground up’. For example, ‘the baby bird’ makes sense, but many other words could follow ‘baby’ and the phrase would still be sensible (e.g., boy, girl, squirrel, etc.). These unique memory-dependent and compositional elements make simple combinatorial phrases an interesting test case for the performance of large language models (LLMs). Further, while LLMs have achieved high performance on benchmarks of human knowledge such as the Graduate Record Examination, it is unclear to what extent these tests measure actual combinatorial comprehension of language and concepts as opposed to ‘bag of words’ or related strategies. Here, we provide a simple Two Word Test to evaluate the combinatorial semantic processing of LLMs, and compare performance to a large sample (N = 150) of human participants who completed the same task. Our results reveal that LLMs fail dramatically at the ability to judge the meaningfulness of simple two word phrases when compared to humans.

**Materials and Methods**

**Phrase Generation and Human Rating Collection**

The Two Word Test for LLMs consists of noun-noun combinations and human meaningfulness ratings originally collected by Graves and colleagues (2013), whose methods we will now summarize. Graves et al began stimuli selection by choosing 500 highly imageable words from six studies of human language processing (cite). Words were rejected from this list if their frequency as nouns were lower than their frequency as other parts of speech (confirmed via CELEX lexical database; cite). All possible noun-noun combinations were generated, resulting in 249,500 phrases. Occurrence of these phrases was cross-referenced with a large database of human-generated text (USENET; Shaoul & Westbury 2007). Phrases that appeared at least once in this corpus, and only in one combinatorial direction (i.e., ‘noun1 noun2’, and not ‘noun2 noun1’) were kept, resulting in 1,351 potential phrases. These phrases were then visually inspected by Graves et al., and phrases with possible interchangeable word orders or were taboo were removed. This resulted in 1,080 meaningful phrases, and consisted of 321 unique words in noun1 position and 298 unique words in noun2 position. 1,080 ‘nonsense’ or low-meaningfulness phrases were then generated by reversing the word order of the meaningful phrases, resulting in 2,060 total phrases that were rated by human participants.

Participants (N=150) rated one of five subsets (412 phrases) of the total phrase pool, with the order of phrase presentation randomized. Participants were given the following instructions to rate each phrase on a 5-point scale (values 0-4):

‘Please read each phrase, then judge how meaningful it is as a single concept, using a scale from 0 to 4 as follows: If the phrase makes no sense, the appropriate rating is 0. If the phrase makes some sense, the appropriate rating is 2. If the phrase makes complete sense, the appropriate rating is a 4. Please indicate your response by clicking on the button to the left of the number. Please consider the full range of the scale when making your ratings.’

Three examples were given: the goat sky, 0 (makes no sense), the fox mask, 2 (makes some sense), and the computer programmer, 4 (makes complete sense).

For each phrase, the mean and standard deviation of responses were calculated, and Figure 1 shows a histogram of the mean responses collapsed across all phrases. The bimodal distribution reflects a clear distinction between ‘meaningful’ and ‘nonsense’ phrases, with phrases in the middle being more ambiguous.

We also include four additional measures (maybe?)

**The Two Word Test: Assessment of Combinatorial Semantic Understanding in GPT**

To assess combinatorial semantic processing, we conducted a series of experiments comparing GPT performance to the human data collected by Graves et al. We begin by exactly replicating the Graves experiment using the GPT API, and progress towards binary meaningfulness judgments in subsets of the noun-noun combinations that have the lowest variability in human ratings (i.e., the ‘easiest’ test cases for GPT to answer correctly). We also test GPT by using binary ‘meaningful’ and ‘meaningless’ prompts instead of a numerical rating system in order to confirm that GPT’s shortcomings are not due to its occasional difficult with number systems (cite). We find that GPT fails dramatically at the Two Word Test, exhibiting meaningfulness judgments that are anomalous when compared to human performance. GPT displays bias towards judging phrases as ‘making sense’, even when humans find those same phrases to be nonsensical. This behavior suggests that GPT lacks true combinatorial semantic abilities in certain task paradigms.

1. **Continuous Meaningfulness Judgments: Replicating Graves et al with GPT**

Using the GPT API, we submitted the prompt and examples originally provided by Graves et al, as well as all 2,060 phrases in a randomized order. We did this three separate times, using temperatures of 0, 0.5, and 1, with 0 being full deterministic and being more probabilistic. Results did not significantly vary by temperature, so we will only report temperature 1 (see supplementary materials for other temperatures). To encourage GPT to use the full range of the 0 to 4 scale, we provided two additional examples for 1 and 3 (not included in the original Graves study; the knife arm, 1 (makes very little sense) and the soap bubble, 3 (makes a lot of sense)). Compared to the human distribution, which reflects ‘meaningful’ and ‘meaningless’ phrases in the bimodal peaks, GPT shows a bias towards rating most phrases as a 2 or 3 (makes some sense, makes a lot of sense; Fig. 1).

Insert fig 1 here

However, distributional differences can be misleading, especially when comparing a single case (GPT) to group-level mean ratings (humans). It is more informative to compare GPT’s response to each individual phrase and test the probability that its response came from the same distribution as the human responses. We therefore conducted a series of phrase-wise statistical tests using generated data and permutation testing to compare GPT to human meaningfulness ratings.

First, we used the phrase-wise means and standard deviations provided by Graves et al to generate a gaussian probability distribution of 10,000 simulated human responses to each phrase, respecting the lower and upper limits of the scale (i.e., no values < 0 or > 4) and rounded to the nearest whole integer to match GPT’s response scale (Fig. 2). Then, for each phrase, we conducted a Crawford & Howell t-test for case-control comparisons with GPT as the case and the distribution as the control. The Crawford & Howell t-test is designed for statistical comparison of a single observation to a group, and returns the probability that the single observation comes from the same probability distribution as the group. We hereby define an ‘Two Word Test failure’ as when GPT’s meaningfulness rating has less than a 5% probability of coming from the human distribution (i.e., GPT’s response is significantly different from humans, alpha = .05).

Figure 2 (sample phrase distributions and GPT’s rating)

This revealed that GPT fails at the Two Word Test on over half of the 2,060 phrases. We then tested multiple alphas, shown in Table #.

|  |  |  |
| --- | --- | --- |
| Alpha | # of GPT Two Word Test Failures | % of GPT Two Word Test Failures |
| .05 | 1,059 | 51.4% |
| .01 | 704 | 34.2% |
| .001 | 468 | 22.7% |

To substantiate our results, we then used two forms of permutation testing to simulate the probability of this many Two Word Test failures. First, we generated 10,000 simulated participants whose phrase-wise responses were based on the underlying probability distributions for each phrase. Then, for each phrase, we tested each simulated participant’s response against the rest of the distribution using the Crawford & Howell t-test in order generate a probability distribution of the total number of ‘Two Word Test failures’ that could be expected if 10,000 subjects were sampled.