**When Transformer models are more compositional than humans: The case of the depth charge illusion**

Dario Paape (University of Potsdam)

English native speakers often interpret the sentence *No head injury is too trivial to be ignored* to mean that head injuries, even seemingly trivial ones, should never be ignored. However, this interpretation is not compositionally licensed: The embedded degree phrase is internally incongruous (sensible: *too serious to be ignored*), and the verb *ignored* should not be negated (cf. *No missile is too small to be banned*) [1]. Nevertheless, participants complete the sentence *No head injury is too trivial to be* ___ with the verb *ignored* or a semantically similar continuation about 80% of the time [2]. This effect is known as the "depth charge" illusion. Among the proposed explanations for the illusion are processing errors and superficial interpretation [1,2], pragmatic inference about the intended meaning [3], and the existence of a stored, non-compositional grammatical template [4].
An interesting question to ask is whether the illusion also appears in giant Transformer-based language models like GPT-3 and BERT [5,6]. Transformer models show an impressive ability to generate coherent text, but struggle with complex grammatical structures [7] and semantic mechanisms such as negation and entailment [8]. For instance, BERT will produce the word *Apple* with equal probability in the sentence *iOS is developed by* ___ compared to the sentence *iOS is not developed by* ___ [9]. In order to provide compositional completions for depth charge sentences, a Transformer model would need to identify the scope of the negation, as well as its interaction with the degree phrase *too trivial to X*. Given their limitations and partial reliance on heuristics [10], Transformers could show a stronger depth charge illusion than humans. On the other hand, Transformers do not process sentences incrementally; they can use all the information in the sentence in parallel [11]. This may give them a compositional advantage over humans: It has been suggested that human compositional processing is foiled in depth charge sentences due to the incremental combination of *no* and the second negative element *too*, which masks the incongruity [2,4]. In sum, Transformers may behave differently from humans with regard to the illusion, but the direction is not clear.
We conducted an experiment with four giant Transformer models: Two versions of GPT-3 (ada with 2.7 billion parameters and davinci with 175 billion parameters), Jurassic-1-Jumbo (175 billion parameters) [12], and RoBERTa, which is BERT with additional training (125 million parameters) [12]. The input items were 32 depth charge sentences that have previously been used with human participants [2,3]. We included a control condition with *some* instead of *no*, which reduces the illusion to about 10% in humans [2], and a condition with *enough* instead of *too*, which allows for a sensible compositional interpretation. A variety of additional controls were tested to check whether the models are sensitive to negation and the meaning of degree constructions, as shown in **Table 1**. The dependent variable is the log probability of the verb (e.g., *ignored*) in each sentence.
As shown in **Figure 1**, the Transformer models show a higher log probability for *ignored* in sentences with *no* than in sentences with *some*, similar to humans. This is despite the fact that the models have apparently encoded the necessary knowledge to handle the construction: The control conditions all show behavior that is consistent with compositionality. However, when looking at actual sentence completions generated by Transformers, patterns emerge that set them apart from humans. First, they often fail in the control conditions, producing transparently incongruous sentences (see examples in **Table 2**). Second, they tend to produce many compositional continuations for depth charge sentences. For instance, in the negated *head injury* item (1b) in **Table 1**, RoBERTa ranks the verbs *addressed* (14%), *treated* (9%) and *considered* (7%) higher than the verb *ignored* (5%).
Taken together, the results show that Transformer models exhibit human-like behavior in that they fall for the depth charge illusion, but also suggest that Transformers may be more compositional than humans in cases where incremental processing creates a bottleneck of complexity.

(1) (a) No head injury is trivial enough to be ignored. ☑ (compositionally sensible)
(b) No head injury is too trivial to be ignored. ☒ (depth charge)
(c) Some head injuries are too trivial to be ignored. ☒ (not compositionally sensible)
(d) No head injury is so trivial as to be ignored. ☑
(e) No head injury is so trivial as to not be ignored. ☒
(f) Head injuries that are too trivial will be ignored. ☑
(g) Head injuries that are not too trivial will be ignored. ☒
(h) Head injuries that are trivial are more likely to be ignored. ☑
(i) Head injuries that are trivial are less likely to be ignored. ☒

**Table 1**. Example item showing the constructions tested in the experiment. 32 different sentences were used.

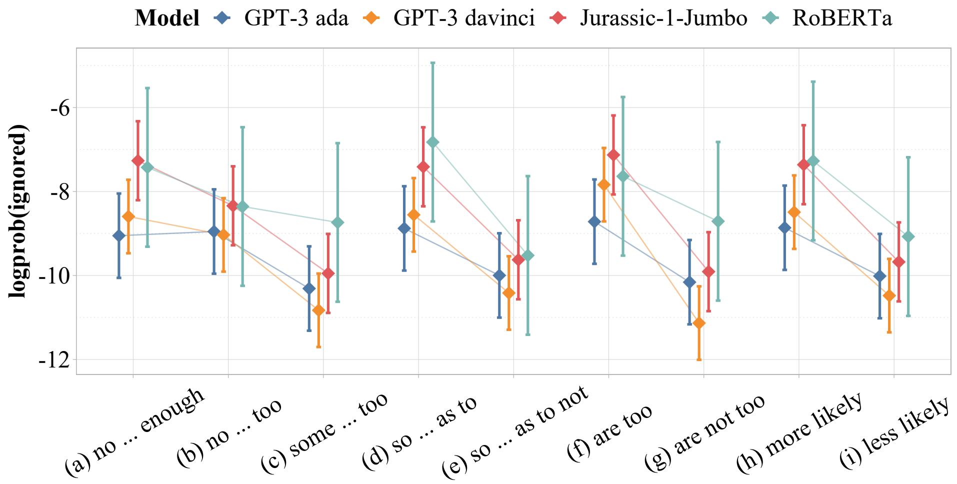

**Figure 1**. Log probability of the critical verb (e.g., *ignored*) by construction and model.

**GPT-3 ada**

No head injury is too trivial to be counted as a crime. ☑ (compositional)
Some head injuries are too trivial to be taken lightly. ☒ (non-compositional)
Head injuries that are trivial are more likely to be fatal. ☒
Head injuries that are trivial are less likely to be fatal. ☑

**GPT-3 davinci**

No head injury is too trivial to be ignored. Any recent head injury, no matter how minor, should be included in the patient's history. ☒
Some head injuries are too trivial to be treated, Dr. Benson acknowledged. ☑

**Jurassic-1-Jumbo**

No head injury is too trivial to be noticed by a parent. ☑
No head injury is too trivial to be ignored. All head injuries need to be taken seriously. ☒

**RoBERTa**

Head injuries that are too trivial will be punished. ??
Some head injuries are too trivial to be ignored. ☒

**Table 2**. Example completions by model.

**References.** [1] Wason & Reich (1979, Q J Exp Psychol). [2] Paape et al. (2020, J Semant). [3] Zhang et al. (2021, AMLaP presentation). [4] Fortuin (2014, Cogn Linguist). [5] Brown et al. (2020, arXiv:2005.14165). [6] Devlin et al. (2018, arXiv:1810.04805v2). [7] van Schijndel et al. (2018, arXiv:1909.00111). [8] Hossain et al. (2020, Proc EMNLP). [9] Hosseini et al. (2021, arXiv:2105.03519 ). [10] McCoy et al. (2019, arXiv:1902.01007). [11] Kahardipraja et al. (2021, arXiv:2109.07364). [12] Lieber et al. (2021, white paper, AI21 labs). [13] Liu et al. (2019, arXiv:1907.11692).