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

Dario Paape*

Abstract. State-of-the-art Transformer-based language models like GPT-3 are very good at generating syntactically well-formed and semantically plausible text. However, it is unclear to what extent these models encode the compositional rules of human language and to what extent their impressive performance is due to the use of relatively shallow heuristics, which have also been argued to be a factor in human language processing. One example is the so-called depth charge illusion, which occurs when a semantically complex, incongruous sentence like *No head injury is too trivial to be ignored* is assigned a plausible but not compositionally licensed meaning (*Don't ignore head injuries, even if they appear to be trivial*). I present an experiment that investigated how depth charge sentences are processed by Transformer models, which are free of many human performance bottlenecks. The results are mixed: Transformers do show evidence of non-compositionality in depth charge contexts, but also appear to be *more* compositional than humans in some respects.

Keywords. Transformer, language model, depth charge illusion, compositionality

1. Introduction. Transformer-based language models like GPT-3 (Brown et al. 2020) show impressive capabilities in terms of being able to produce naturalistic, usually grammatically correct, and often coherent text (e.g., Dale 2021). Rather than using recurrent and convolutional neural networks like previous state-of-the-art language models, the Transformer architecture is based on a so-called self-attention mechanism: during training on large amounts of human-written text, the model learns which parts of the preceding context are important for predicting the next word (Vaswani et al. 2017). Using only this information, GPT-3 can "write" passable philosophical essays (Elkins and Chun 2020), and even scientific papers about itself (Generative Pretrained Transformer et al. 2022). The outputs are often so convincing that humans cannot distinguish between machine-generated and human-written text (Clark et al. 2021, Uchendu et al. 2021).

At the same time, however, it is relatively easy to unmask Transformer models if one knows what to look for. For instance, due to their architecture and training regime, Transformers often fail at simple arithmetic (Floridi and Chiriatti 2020, Patel, Bhattamishra and Goyal 2021), arrive at bizarre deductions in scenarios that require real-world knowledge, and sometimes output obvious non sequiturs with sudden and extreme topic shifts that would be absurd coming from a human writer or speaker (Marcus and Davis 2020). Furthermore, given that any "knowledge" about the world that may be encoded in the model is not grounded in experience or reasoning but is filtered through language and its statistical properties (e.g., Alberts 2022), such as the frequent co-occurrence of certain terms, Transformers often resort to heuristics: they produce associatively plausible rather than factually correct answers to information questions (Sobieszek and Price 2022), and to some extent rely on simple lexical overlap between a premise and a hypothesis to predict entailment or non-entailment (McCoy, Pavlick and Linzen 2019).

*The author would like to thank the Vasishth Lab members, Yuhan Zhang, and Lisa Levinson for helpful comments and suggestions. Author: Dario Paape, Department of Linguistics, University of Potsdam (paape@uni-potsdam.de).

That the "knowledge" encoded by Transformers is often heuristic in nature is also highlighted by so-called mispriming effects: For instance, BERT (Devlin et al. 2018), a close cousin of GPT-3, when asked to complete the input *Talk? Birds can* . . . , will produce *talk* as the most likely continuation, but will produce *fly* as the most likely continuation when the prompt is *Birds cannot* . . . (Kassner and Schütze 2019). Similarly, GPT-3 will assume that a mixture of cranberry juice and grape juice is poisonous if the linguistic context suggests that dangerous substances are being mixed (Marcus and Davis 2020). Even though several of these limitations can be overcome by targeted training and/or the addition of symbolic knowledge (see Helwe et al. 2021 for a review), it is nevertheless striking that *un*targeted training on very large language corpora often results in reliance on relatively shallow processing strategies.

The implicit or explicit gold standard against which language models are usually compared and evaluated in terms of their "shallowness" is human performance. However, there are many tasks related to language and reasoning on which humans do not perform well, which casts doubt on this rationale (Linzen and Baroni 2021). Many failures of human reasoning can also be seen as being the result of mispriming effects: For instance, the Cognitive Reflection Test (Frederick 2005) and its extension (Thomson and Oppenheimer 2016) contain questions such as *How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?*, to which 84% of participants answer "27" despite the correct answer being "none". Relatedly, when asked *How many animals of each kind did Moses take on the ark?*, 81% of participants answer "two" even after having been instructed to look out for possible errors in the question, and even though they know that the biblical story is about Noah (Erickson and Mattson 1981).

Human blindness to incongruous information in otherwise highly congruent contexts and the tendency to fall for verbal misdirection generalize to examples from different thematic domains (e.g., Barton and Sanford 1993, Cook et al. 2018), and are not reducible to the default assumption that interlocutors always produce sensible statements and requests (Reder and Kusbit 1991). In light of results such as these, it has been proposed that human language processing is partly heuristic and often just "good enough" (e.g., Ferreira and Patson 2007, Christianson 2016): Instead of constructing detailed syntactic and semantic structures based on compositional rules, people may sometimes use high-level language statistics and world knowledge to derive a "quick and dirty" approximation of meaning. As a case in point, the meaning of implausible passive sentences such as *The dog was bitten by the man* is often converted into that of an active sentence with reversed roles (*The dog bit the man*), presumably because this meaning is a priori more plausible, and because the agent of an event is usually mentioned first in English (Ferreira 2003, Christianson, Luke and Ferreira 2010).

1.1. THE DEPTH CHARGE ILLUSION. The difference between heuristic language processing and "reasoning" in Transformers and in humans is that the human version can usually be neutralized by explicitly pointing out the problematic element(s) in a given sentence (e.g., Erickson and Mattson 1981, Barton and Sanford 1993), or by explaining the invalidity of a given inference and explaining the correct solution (e.g., van Benthem 2008, Claidière, Trouche and Mercier 2017, Calvillo, Bratton, Velazquez, Smelter and Crum 2022). However, the incorrect "solutions" to some reasoning problems, like the Monty Hall problem (e.g., Vos Savant 1997, Rosenthal 2008) and the Wason selection task (Wason 1968), famously tend to resist being explained away in this manner.

In the linguistic domain, an example that also has the property of persistence in the face of corrective explanation is the so-called depth charge illusion, which was first discussed by Wason and Reich (1979). The depth charge illusion occurs in (1), which is often interpreted to mean *Don't ignore head injuries, even if they appear to be trivial*.

(1) No head injury is too trivial to be ignored.

Anecdotally, a small subset of people immediately recognizes that this sentence is incorrect, a second subset is open to considering the possibility that it *could* be incorrect but cannot see why, and a third subset will stubbornly insist that it is correct, even after being confronted with the following argumentation:

1. The phrase *too trivial to be ignored* is semantically incongruous, because it presupposes that something can be *so trivial that it should* ***not*** *be ignored* (compare *X is too young to die*, which translates to *X is so young that they should* ***not*** *die*).
2. The incongruity is not removed by the initial negation: Asserting that no head injury has the incongruous property of being *too trivial to be ignored* does not make the property itself any less incongruous.
3. The initial negation *does* cause the overall statement to be affirmative, contrary to the plausible misinterpretation (*Don't ignore head injuries*): In abstract terms, if no X is too Y to be Z'ed, this means that no X crosses the threshold beyond which it should *not* be Z'ed, meaning that *all* X should be Z'ed (that is, ignored).
4. Compositionally, the sentence thus means *Ignore all head injuries, even if they appear to be trivial*.

The sentence *can* be made compositionally sensible by changing it to *No head injury is too trivial to be noticed/treated* or to *No head injury is trivial enough to be ignored*, but speakers will often find these variants difficult to process or even reject them as being malformed.

Despite broad agreement in the literature that the "don't ignore" interpretation of (1) is not compositional, not all scholars agree that the depth charge illusion is due to a processing error, partly because it is so persistent. Explanations generally fall into three categories: The classic shallow processing account (Wason and Reich 1979, Paape, Vasishth and von der Malsburg 2020), an account based on the alleged idiomaticity of the construction *No X is too Y to Z* (Cook and Stevenson 2010, Fortuin 2014), and an account based on unconscious correction of an assumed speech error (Zhang, Ryskin and Gibson 2022).

The shallow processing account claims that due to the syntactic and semantic complexity of (1), working memory becomes overloaded at some point and compositional processing breaks down or is suspended, presumably when the implicit negation contained in the word *too* is combined with the initial negation (Paape et al. 2020). Readers then use their world knowledge in combination with superficial language heuristics (*duplex negatio affirmat*; *No head injury is too trivial ...* $\rightarrow$ *All head injuries are too dangerous ...*) to derive a plausible meaning. By contrast, the idiomatic or construction-based account assumes no breakdown. Instead, its proponents claim that the *No X is too Y to Z* construction is a stored grammatical unit that can, by virtue of its idiomaticity, violate the compositionality principle and be "legally" interpreted to mean *No X should be Z'ed*. Finally, the error-correction account claims that readers combine prior expec-

tations about plausible utterances with expectations about plausible speech errors to reconstruct the presumably intended sentence (*No head injury is so trivial as to be ignored*) and derive its meaning. The proposal that the human sentence processor uses prior expectations about sentence meanings and speech errors to "repair" potentially degraded input has also been applied to other cases of linguistic illusions (e.g., Gibson, Bergen and Piantadosi 2013, Frazier and Clifton 2015).

All proposed accounts have empirical weaknesses: Shallow processing cannot explain why readers usually don't consciously notice the complexity overload that causes compositional processing to break down, or why the illusion often cannot be explained away: If complexity is the problem, taking one's time and working out the compositional meaning step by step should always lead to success. Conversely, the construction-based account cannot explain why the illusion *can* sometimes be explained away, why some people appear to be immune to it, and why it generalizes to distinct but compositionally similar constructions (e.g., *too … as that* in German; Paape et al. 2020). Finally, the Bayesian error-correction account cannot explain why readers usually cannot consciously access and report the assumed error correction ("I believe the speaker/writer made a mistake here") even after multiple passes over the sentence,[1] and why putting the incongruity in focus by changing the word order weakens the illusion (*Too trivial to be ignored is surely no head injury* in German; Paape 2021).

Investigating how Transformers handle depth charge sentences may provide a way out of the empirical conundrum. Unlike humans, Transformers don't have limited working memory capacity, so they don't experience complexity overload. Humans suffer from the "now or never" bottleneck (Christiansen and Chater 2016), that is, they must quickly and incrementally integrate incoming information before it is forgotten. Transformers, by contrast, don't process sentences incrementally but holistically, that is, they always have full access to all words in the sentence unless this access is deliberately limited (Kahardipraja, Madureira and Schlangen 2021). On the other hand, Transformers are prone to learning heuristics rather than compositional rules (e.g., McCoy et al. 2019), and are known struggle with negation (Kassner and Schütze 2019, Hossain, Kovatchev, Dutta, Kao, Wei and Blanco 2020, Hosseini, Reddy, Bahdanau, Hjelm, Sordoni and Courville 2021), so that they might to some extent mimic human "good enough" processing of depth charge sentences.

By contrast, under the construction-based account, the Transformer would need to learn from the training data that the *No X is too Y to Z* construction cannot only be used compositionally (*No head injury is too trivial to be treated*) but also non-compositionally, that is, idiomatically. However, the non-compositional variant is relatively rare: Cook and Stevenson (2010) report 170 instances of the construction in a written corpus of 1.1 billion words, of which 80% were compositional (e.g., *no risk [is] too small to eliminate*). Fortuin (2014) reports only 13 instances of the "negative" (idiomatic) construction in a corpus of similar size. Transformers tend to overgeneralize when the number of exceptions to a compositional rule is small, showing a) that compositionality *is* learned to some degree and b) that a certain amount of counterevidence is needed to "memorize" exceptions (Hupkes et al. 2020).[2]

[1]Even in an experimental context where the task is to correct incorrect use of *too* and *enough*, participants propose corrections to depth charge sentences only in about 30% of trials, and in most cases suggest changing the adjective (e.g., *trivial* → *dangerous*; O'Connor 2015), which leaves the compositional implausbility of the overall interpretation ("Don't ignore head injuries") intact.

[2]Interestingly, even if the depth charge illusion is best characterized as a processing error in humans, if the error

Regarding the Bayesian error-correction account of Zhang et al. (2022), it is not entirely clear how it could be applied to Transformers: Transformers don't have "common sense" learned from the real world against which they can evaluate sentence meanings, nor do they have a notion of speech errors, much less a subconscious error correction mechanism. To a standard Transformer, there is no "noise" — every sentence seen during training is data, and the model can only adapt by allowing for more variability in its parameters (Michel and Neubig 2018, Passban, Saladi and Liu 2020). There may be some notion of a "plausible utterance" encoded in the model, in the sense that words with certain meanings tend to co-occur, possibly even in syntactically similar environments, but it is unlikely that the model also encodes the assumption that unintended errors can lead to malformed sentences, and is able to reconstruct the original meaning.

In what follows, I present an experiment in which different Transformer models were tested on the *No X is too Y to Z* construction and a variety of control constructions. The experiment is exploratory in nature: The aim was not to conclusively answer the question of how the depth charge illusion arises in humans, but to see whether a system trained on many terabytes of text, but without incremental processing, memory bottlenecks, or error correction mechanisms would show the illusion or not.

2. Experimental study. The purpose of the study was to assess the probability different Transformer models assign to the word *ignored* as the next word after seeing the preamble *No head injury is too trivial to be* .... The log probability of *ignored* is treated as the dependent variable, and higher probabilities are taken to indicate a stronger depth charge illusion. This is a simplification, as the models may also assign high probabilities to continuations such as *overlooked* or *forgotten about*, which are semantically similar to *ignored*. To solve this problem, one can ask human coders to classify the continuations into "*ignore*-like" and "*treat*-like" categories, and then sum the relevant probabilities. However, there is some uncertainty as to which coding scheme should be used, as the relevant dimension of semantic similarity is not easy to capture (Paape et al. 2020, O'Connor 2015). I thus restrict my analysis to the single token *ignored* here. In sections 2.4 and 2.5, the overall distribution of continuations is discussed in more detail.

2.1. MATERIALS. The experiment had 9 conditions overall, as shown in (2). Here, *compositional* is used as shorthand for "has a sensible meaning under a compositional analysis", whereas *not compositional* is used as shorthand for "does not have a sensible meaning under a compositional analysis". Conditions (2-a), (2-b) and (2-c) are the depth charge conditions, while the rest are control conditions designed to test whether the models have encoded knowledge about negation, scales, and the degree particles *too* and *so*. The control conditions are syntactically and/or semantically less complex than the depth charge conditions, and a Transformer that has encoded the relevant knowledge should consistently assign higher probability to *ignored* in the compositional conditions compared to the non-compositional conditions.

is frequent enough, a Transformer would presumably treat it as evidence of a grammaticalized rule exception. In addition, there are several academic papers on the depth charge illusion (see Paape 2021 for a review), in addition to discussions on several public web forums (e.g., https://english.stackexchange.com/questions/91612/no-head-injury-is-too-trivial-to-ignore), which may become part of the training data of Transformer models and provide "evidence" of the construction being used.

(2) a. No head injury is trivial enough to be → ignored (compositional) ☑
b. No head injury is too trivial to be → ignored (not compositional) ☒
c. Some head injuries are too trivial to be → ignored (not compositional) ☒

d. No head injury is so trivial as to be → ignored (compositional) ☑
e. No head injury is so trivial as to not be → ignored (not compositional) ☒

f. Head injuries that are too trivial will be → ignored (compositional) ☑
g. Head injuries that are not too trivial will be → ignored (not compositional) ☒

h. Head injuries that are trivial are more likely to be → ignored (compositional) ☑
i. Head injuries that are trivial are less likely to be → ignored (not compositional) ☒

Replacing *too* in the depth charge sentence (2-b) with *enough* in (2-a) yields a compositionally sensible meaning when the sentence is completed with the verb *ignored*. Humans do indeed produce compositional, "*ignore*-like" completions for this construction in the majority of trials (O'Connor 2015). The third depth charge condition (2-c) with *some* instead of *no* also leads to more compositional completions in humans, but in this case the completions are "*treat*-like" rather than "*ignore*-like" (Paape et al. 2020). Humans also assign lower sensibleness ratings to *some*-sentences ending with *ignored*, suggesting that the initial negation is crucially involved in "masking" the incongruity of the degree phrase in (2-b) and creating the depth charge illusion (Paape et al. 2020, Paape 2021).

To see if differences between conditions generalize across different sentence contexts, the 32 German depth charge items used by Paape et al. (2020) were translated into English and adapted to fit the design shown in (2). Only the versions with negative adjectives were used (e.g., *No plan is too unrealistic to be → scrapped*, *No physical theory is too implausible to be → dismissed*). Across all sentences, the dependent variable was the log probability of the verb used in the Paape et al. rating experiments.

2.2. TESTED MODELS. Four models were tested. The first two were models of different sizes from the GPT-3 family: Ada, the least powerful version of GPT-3, which "can perform tasks like parsing text, address correction and certain kinds of classification tasks that don't require too much nuance" and Davinci, the most powerful version, which "shines [...] in understanding the intent of text" and "is quite good at solving many kinds of logic problems" (https://beta.openai.com/docs/models/gpt-3). The third model under consideration was Jurassic-1-Jumbo, released by AI21 Labs, which is similar in size to Davinci but outperforms it in terms of predictive accuracy on many corpora (Lieber et al. 2021). The fourth model was RoBERTa, a retrained version of BERT with improved performance (Liu et al. 2019; https://huggingface.co/roberta-large). RoBERTa's training regime is somewhat different from that of GPT-3 and Jurassic-1: Like BERT, RoBERTa is bidirectional, that is, it not only considers the context to the left of a word but also the context to the right. RoBERTa has 355 million parameters, and is thus similar in size to GPT-3 Ada[3]. GPT-3 Davinci and Jurassic-1-Jumbo are much larger, with about 175 *billion* parameters each.

The GPT-3 models were queried via the OpenAI API, Jurassic-1-Jumbo was queried via the

[3]This assumes that Ada corresponds to the 350M version of GPT-3 reported by Brown et al. (2020) (https://blog.eleuther.ai/gpt3-model-sizes/).

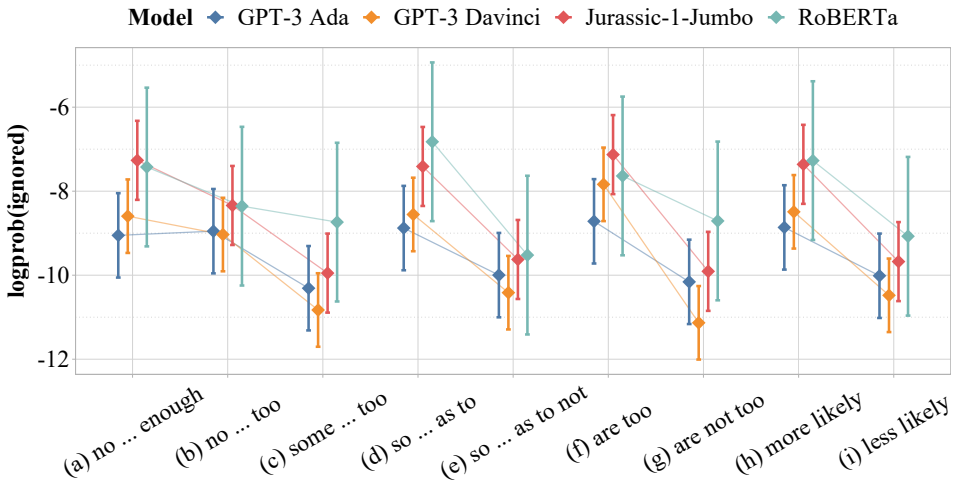

Figure 1: Log probability of the critical verb (e.g., *ignored*) by construction and model. Error bars show 95% confidence intervals across all 32 items.

AI21 Labs API, and RoBERTa was queried via the HuggingFace API (Wolf et al. 2020). For multi-token completions (e.g., … *to be ruled out*) the log probabilities of the generated tokens were summed. The pretrained models were used as-is; no fine-tuning of any kind was carried out.

2.3. Results and Bayes factor analysis. Figure 1 shows the results by model and condition. In order to gauge the amount of statistical evidence in the data for differences between conditions and between models, the returned log probabilities were analyzed using a linear mixed-effects model (LMM) in Stan (Stan Development Team 2022) via the brms package (Bürkner 2017) in R (R Core Team 2022). The LMM assumed a Gaussian likelihood, and contained a fixed effect of model, which was treatment-coded with GPT-3 Ada as the baseline, as well as random intercepts by sentence and random slopes for model by sentence. For the conditions, sum contrasts were defined in the following way:

- (2-b) versus (2-a): *too* −1 versus *enough* +1 (compositional versus non-compositional)
- (2-c) versus (2-b): *some* −1 versus *no* +1 (negation masks incongruity)
- (2-d) versus (2-e): *as to not* −1 versus *as to* +1
- (2-f) versus (2-g): *are not too* −1 versus *are too* +1
- (2-h) versus (2-i): *less likely* −1 versus *more likely* +1

Normal(0,1) priors were used for all contrasts, and Bayes factors were computed using the bayestestR package (Makowski, Ben-Shachar and Lüdecke 2019). The model code and the data are available at `https://osf.io/yw8vk`. Bayes factors are interpreted according to the scheme given by Andraszewicz et al. (2015). The results of the analysis are as follows (BF is BF_{10}):[4]

[4]Because I am only interested in evidence for differences between models, I do not distinguish cases in which the null hypothesis is favored, that is, in which there is evidence in favor of the models behaving similarly, from cases in which the evidence doesn't favor either hypothesis.

- For GPT-3 Ada as the baseline, strong evidence for positive differences for all contrasts (all BFs > 15), except for the *too* / *enough* contrast (BF = 1.6).
- *too* / *enough* contrast: Moderate evidence for larger difference in Jurassic-1-Jumbo than in GPT-3 Ada (BF = 5.3).
- *some* / *no* contrast: No evidence that other models perform differently from GPT-3 Ada (all BFs < 1.1).
- *as to not* / *as to* contrast: Moderate evidence for larger difference in RoBERTa than in GPT-3 Ada (BF = 6.7).
- *are not too* / *are too* contrast: Strong evidence for larger difference in GPT-3 Davinci than in GPT-3 Ada (BF = 27), moderate evidence for larger difference in Jurassic-1-Jumbo than in GPT-3 Ada (BF = 3.2).
- *less likely* / *more likely* contrast: No evidence that other models perform differently from GPT-3 Ada (all BFs < 1.7).

2.4. EXAMPLE COMPLETIONS BY MODEL. Example sentence completions produced by the different models are shown below, along with indicators showing whether the completion is compositionally sensible or not. The prompt is underlined in each example.

(3) **GPT-3 Ada**

a. No head injury is too trivial to be counted as a crime. (compositional) ☑
b. Some head injuries are too trivial to be taken lightly. (not compositional) ☒
c. Head injuries that are trivial are more likely to be fatal. ☒
d. Head injuries that are trivial are less likely to be fatal. ☑

(4) **GPT-3 Davinci**

a. 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. ☒
b. Some head injuries are too trivial to be treated, Dr. Benson acknowledged. ☑

(5) **Jurassic-1-Jumbo**

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

(6) **RoBERTa**

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

The completions are interesting in multiple regards. Completions like (3-a), (4-b) and (5-a) appear to be compositional, that is, there is no depth charge illusion in these examples. On the other hand, examples (4-a) and (5-b) show that when the models produce non-compositional completions, they will occasionally follow them up with semantically matching continuations. The pair (3-c)/(3-d) shows that GPT3 Ada's performance on the relatively straightforward control conditions is far from perfect: It is highly unlikely that humans would ever produce *fatal* in (3-c). Completions (3-b) and (6-b) show that Transformers produce non-compositional completions in the *some* condition as

well, which happens very rarely in humans (Paape et al. 2020). Completion (6-a) by RoBERTa is completely unexpected, unless it is taken to mean that whoever caused the head injury is going to be punished, in which case it would be a non-compositional completion.

2.5. FORWARD AND BACKWARD MASK-FILLING WITH ROBERTA. As mentioned above, RoBERTa is bidirectional, and can thus produce completions based on rightward as well as on leftward context, that is, RoBERTa is able to "retrodict" words based on future input. This is achieved by putting a [MASK] token in place of the to-be-inserted word. The masking feature can be used to take a closer look at compositionality in RoBERTa: Given an input string like *No head injury is too [MASK] to be ignored*, does RoBERTa produce an adjective that results in a compositionally well-formed degree phrase? It is also worthwhile to look at the distribution of completions: Even if the most likely completion is compositional, there may be an alternative, non-compositional completion with a similarly high probability, or the other way around. RoBERTa's top 4 verb and adjective completions with their associated probabilities for a set of example sentences are shown in (7) below.

(7) a1. **No head injury is too trivial to be [MASK]**
addressed – 14%, treated – 9%, considered – 7%, ignored – 3%

a2. **No head injury is too [MASK] to be ignored**
serious – 36%, minor – 10%, severe – 8%, small – 8%

b1. **No potential habitat is too nutrient-poor to be [MASK]**
developed – 17%, exploited – 17%, explored – 11%, viable – 3%

b2. **No potential habitat is too [MASK] to be ruled out**
small – 24%, remote – 12%, obscure – 4%, good – 4%

c1. **No chapter is too irrelevant to be [MASK]**
archived – 6%, read – 5%, included – 3%, forgotten – 3%

c2. **No chapter is too [MASK] to be skipped**
important – 50%, long – 13%, short – 7%, boring – 5%

d1. **No physical theory is too implausible to be [MASK]**
tested – 40%, proven – 5%, tried – 4%, considered – 3%

d2. **No physical theory is too [MASK] to be dismissed**
old – 11%, weak – 9%, absurd – 6%, good – 5%

For the selected examples, most of RoBERTa's completions are compositional, though there are also non-compositional completions with relatively high probabilities. The picture differs markedly from human performance: Humans complete *No head injury is too trivial to be …* and *No chapter is too irrelevant to be …* with "*ignore*-like" and "*skip*-like" continuations about 80% of the time (Paape et al. 2020, Appendix B). On the other hand, compositionally "retrodicting" the adjective seems to be difficult for RoBERTa in some contexts: RoBERTa mostly generates non-compositional adjective completions in (7-b2) and (7-d2), even though masking the verb mostly results in compositional completions for the same sentences in (7-b1) and (7-d1). For (7-a2) and (7-c2), on the other hand, the adjective completions are mostly compositional, just like the verb

completions in (7-a1) and (7-c1). Overall, RoBERTa's performance shows quite a lot of variability between sentences, which matches what is observed in humans (Paape et al. 2020, Paape 2021, Zhang et al. 2022). Possible sources of this variability are discussed below.

3. Discussion. The aim of this paper was to investigate whether Transformer-based language models show the depth charge illusion, in which a compositionally incongruous sentence (*No head injury is too trivial to be ignored*) is given an unlicensed but plausible interpretation (*Don't ignore head injuries, even if they appear to be trivial*). Transformers are an interesting test case, given the range of proposed explanations for the illusion in human readers: Theoretical proposals range from processing breakdown and recovery through world knowledge and superficial language heuristics (Wason and Reich 1979, Paape et al. 2020), to the existence of an idiomatic *No X is too Y to Z* construction (Cook and Stevenson 2010, Fortuin 2014), to Bayesian speech error correction (Zhang et al. 2022). Transformers don't experience processing breakdown, may not have seen enough instances of the hypothesized *No X is too Y to Z* construction to encode it, and have no means of distinguishing between "normal" training data and speech errors.

The experimental results yielded some evidence that the depth charge illusion is present in Transformers of different types and sizes: Across 32 test sentences, all considered models (GPT-3 Ada, GPT-3 Davinci, Jurassic-1-Jumbo, and RoBERTa) assigned higher probabilities to completions like *ignored* when the sentence began with a negation (*No head injury . . .*) compared to when it did not (*Some head injuries . . .*), even though the completion results in an internally incongruous degree phrase (*too trivial to be ignored*) in both cases. Furthermore, apart from Jurassic-1-Jumbo, none of the models appeared to distinguish between *too* and *enough* in negated contexts, even though the two degree particles have opposite meanings. At the same time, however, the Transformer models showed evidence of compositional processing in control contexts such as *Head injuries that are trivial are less likely to be . . . [*ignored]*, suggesting that the required syntactic and semantic rules have been encoded.

Taken at face value, these results suggest largely parallel effects between human readers and Transformers with regard to the depth charge illusion, despite the presumably very different underlying processing mechanisms. However, a closer look at the Transformers' sentence completions revealed that they produce a variety of un-humanlike continuations for the control conditions, suggesting that their grammatical "knowledge" may not be as deep as the high-level results suggest (Bender et al. 2021). On the other hand, the mask-filling patterns of RoBERTa suggested that RoBERTa is often *more* compositional than human readers: For a variety of test sentences, including the most famous example *No head injury is too trivial to be . . .*, RoBERTa showed a preference for compositional completions like *addressed*, unlike human participants (Paape et al. 2020, O'Connor 2015). At the same time, non-compositional completions did also appear in the list of most likely tokens, and even dominated for some sentences, especially in "retrodictive" contexts, that is, when RoBERTa had to fill in the adjective based on the verb (e.g., *No potential habitat is too [small] to be ruled out*).

What are the implications of these findings for the empirical deadlock between the competing psycholinguistic accounts of the depth charge illusion? Proponents of the construction-based view could argue that the Transformer models have picked up on the *No X is too Y to Z* construction to some extent, but haven't fully mastered it yet, presumably because they haven't encoun-

tered enough instances of it in the input data, and because the construction is arguably ambiguous between a compositional and a non-compositional version (Cook and Stevenson 2010, Fortuin 2014). A preference for the compositional reading wouldn't be surprising under this view, given that Transformers are known to struggle with infrequent and/or idiomatic constructions (Hupkes et al. 2020, Dankers et al. 2022). This suggests some promising avenues for future research: Providing disambiguating context should allow the Transformer to identify the intended reading, as it arguably does for humans (Fortuin 2014), and additional training on the construction should increase accuracy, in the sense that contextual cues should be more reliably identified.

Meanwhile, proponents of the superficial processing account could argue that even though Transformers don't experience processing breakdown, they could nevertheless be using heuristics to process depth charge sentences. This is a plausible assumption, given that Transformers are known to learn heuristics in other settings, including negation processing (McCoy, Pavlick and Linzen 2019, Helwe, Clavel and Suchanek 2021). That the models haven't achieved human-like syntactic and semantic competence in terms of processing scales (*more trivial → higher probability of ignoring*), degrees (*too trivial*) and negation is clear from the many examples in which they produced compositionally incongruous completions in the control conditions.

At the same time, however, the models do not appear to use the simplest possible heuristic for dealing with depth charge sentences: to ignore the beginning of the sentence and only locally evaluate the degree phrase *too trivial to be ignored*, which is always incongruous, irrespective of whether the sentence begins with *no* or with *some*. This strategy is unlikely to be used by human readers, who are limited by their incremental left-to-right processing and the "now or never" bottleneck (Christiansen and Chater 2016), which may lead them to incorrectly combine *no* and *too* before they even reach the verb (Paape et al. 2020). Transformers, on the other hand, do not have this limitation, and yet they have apparently learned to pay attention to the initial *no* in depth charge contexts.

Where does this leave the superficial processing account? The following scenario is possible: Even larger Transformer models with even more (or better) training may eventually acquire human-like syntactic and semantic competence, but may not exhibit the depth charge illusion, because their output is not limited by performance factors.[5] Resistance to the illusion may also gradually increase with scale and training, as the amount of abstract compositional knowledge and, presumably, transfer ability in the system increases. Larger models typically do perform better on language tasks, though the current data do not show evidence of scale effects: The strength of the depth charge illusion was similar across models of very different sizes, though Jurassic-1-Jumbo showed some evidence of distinguishing more between *too* and *enough* than the other models.

Proponents of the Bayesian error-correction account could argue that despite the absence of "reasoning" in Transformers, the models may have learned to correct for speech errors to some extent. Transformer-based language models often acquire unexpected capabilities that they were not explicitly trained for (e.g., Radford et al. 2019, Brown et al. 2020), and deep learning has been touted as a potentially powerful approach to grammar correction (Dale and Viethen 2021), so error

[5]This should also extend to other linguistic illusions that have been attributed to performance factors, such as agreement attraction (*The key to the cabinets are on the table*). However, the precise way in which Transformers encode syntactic structure and linguistic dependencies may be inherently error-prone (Finlayson et al. 2021, Ryu and Lewis 2021), so that complete immunity may be impossible.

repair may be a latent capability of such models. However, reasoning about what the originally *intended form or message* of a given linguistic utterance was (*No head injury is so trivial as to be ignored*; Zhang et al. 2022) and how it might have been transformed into its observed form (*No head injury is too trivial to be ignored*) is a very complex task. This type of reasoning would presumably require some approximation of a theory of mind, as well as an approximate model of human speech production, relevant topic knowledge (*Should all head injuries be treated?*), and pragmatics. Capabilities that resemble common-sense reasoning may be present in current Transformers to some extent, but, as Kejriwal et al. (2022) have recently argued, a more diverse set of empirical tools is needed to find out how much "common sense" is really encoded in the models.

A deeper understanding of how Transformers process the depth charge construction will, in all likelihood, benefit future research into why most humans struggle with the construction, and why there are such large differences between people and between specific sentences. At the sentence level, the interpretation of depth charge stimuli partly depends on the strength of world knowledge associated with the sentence (Paape et al. 2020, Zhang et al. 2022), as well as its sentiment polarity, semantic cohesion, and word order (Paape 2021). Investigating whether Transformers are also sensitive to these factors would yield further insights into how similar the mechanisms behind the illusion are in Transformers and humans. At the participant level, working memory capacity has been investigated as a potential source of variability, yielding a null result (Paape et al. 2020). A more promising factor may be language experience. An interesting approach would be to try and fine-tune Transformers in such a way that they become immune to the depth charge illusion, and to then see if a similar approach can be used to immunize human participants. "Natural" immunity appears to be rare in humans but may also depend on language experience. If immunization is possible, this would support the original intuition of Wason and Reich (1979) that the depth charge illusion is a processing error, and weaken theories that see the illusion as a result of grammaticalization or pragmatic reasoning.

References

Alberts, L., 2022. Is it possible not to cheat on the Turing Test? Exploring the potential and challenges for true natural language 'understanding' by computers. arXiv preprint arXiv:2206.14672 doi:https://doi.org/10.48550/arXiv.2206.14672.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., Wagenmakers, E.J., 2015. An introduction to Bayesian hypothesis testing for management research. Journal of Management 41, 521–543. doi:https://doi.org/doi/10.1177/0149206314560412.

Barton, S.B., Sanford, A.J., 1993. A case study of anomaly detection: Shallow semantic processing and cohesion establishment. Memory & Cognition 21, 477–487. doi:https://doi.org/10.3758/BF03197179.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. doi:https://doi.org/10.1145/3442188.3445922.

van Benthem, J., 2008. Logic and reasoning: Do the facts matter? Studia Logica 88, 67–84.

doi:https://doi.org/10.1007/s11225-008-9101-1.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. URL: https://arxiv.org/abs/2005.14165.

Bürkner, P.C., 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80, 1–28.

Calvillo, D.P., Bratton, J., Velazquez, V., Smelter, T.J., Crum, D., 2022. Elaborative feedback and instruction improve cognitive reflection but do not transfer to related tasks. Thinking & Reasoning 0, 1–29. doi:https://doi.org/10.1080/13546783.2022.2075035.

Christiansen, M.H., Chater, N., 2016. The now-or-never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences 39, e62. doi:https://doi.org/10.1017/S0140525X1500031X.

Christianson, K., 2016. When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. Quarterly Journal of Experimental Psychology 69, 817–828. doi:https://doi.org/10.1080/17470218.2015.1134603.

Christianson, K., Luke, S.G., Ferreira, F., 2010. Effects of plausibility on structural priming. Journal of Experimental Psychology: Learning, Memory, and Cognition 36, 538–544. doi:https://doi.org/10.1037/a0018027.

Claidière, N., Trouche, E., Mercier, H., 2017. Argumentation and the diffusion of counter-intuitive beliefs. Journal of Experimental Psychology: General 146, 1052–1066. doi:https://doi.org/10.1037/xge0000323.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N.A., 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics. pp. 7282–7296. doi:https://doi.org/10.18653/v1/2021.acl-long.565.

Cook, A.E., Walsh, E.K., Bills, M.A., Kircher, J.C., O'Brien, E.J., 2018. Validation of semantic illusions independent of anomaly detection: Evidence from eye movements. Quarterly Journal of Experimental Psychology 71, 113–121. doi:https://doi.org/10.1080/17470218.2016.1264432.

Cook, P., Stevenson, S., 2010. No sentence is too confusing to ignore, in: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, pp. 61–69. URL: https://aclanthology.org/W10-2109.

Dale, R., 2021. GPT-3: What's it good for? Natural Language Engineering 27, 113–118. doi:https://doi.org/10.1017/S1351324920000601.

Dale, R., Viethen, J., 2021. The automated writing assistance landscape in 2021. Natural Language Engineering 27, 511–518. doi:https://doi.org/10.1017/S1351324921000164.

Dankers, V., Lucas, C.G., Titov, I., 2022. Can transformer be too compositional? Analysing id-

iom processing in neural machine translation. URL: https://arxiv.org/abs/2205.15301.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Elkins, K., Chun, J., 2020. Can GPT-3 pass a writer's Turing test? Journal of Cultural Analytics 5, 17212. doi:https://doi.org/10.22148/001c.17212.

Erickson, T.D., Mattson, M.E., 1981. From words to meaning: A semantic illusion. Journal of Verbal Learning and Verbal Behavior 20, 540–551. doi:https://doi.org/10.1016/S0022-5371(81)90165-1.

Ferreira, F., 2003. The misinterpretation of noncanonical sentences. Cognitive Psychology 47, 164–203. doi:https://doi.org/10.1016/S0010-0285(03)00005-7.

Ferreira, F., Patson, N.D., 2007. The 'good enough' approach to language comprehension. Language and Linguistics Compass 1, 71–83. doi:https://doi.org/10.1111/j.1749-818X.2007.00007.x.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., Belinkov, Y., 2021. Causal analysis of syntactic agreement mechanisms in neural language models. arXiv preprint arXiv:2106.06087 doi:https://doi.org/10.48550/arXiv.2106.06087.

Floridi, L., Chiriatti, M., 2020. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines 30, 681–694. doi:https://doi.org/10.1007/s11023-020-09548-1.

Fortuin, E., 2014. Deconstructing a verbal illusion: The 'No X is too Y to Z' construction and the rhetoric of negation. Cognitive Linguistics 25, 249–292. doi:https://doi.org/10.1515/cog-2014-0014.

Frazier, L., Clifton, Jr, C., 2015. Without his shirt off he saved the child from almost drowning: Interpreting an uncertain input. Language, Cognition and Neuroscience 30, 635–647. doi:https://doi.org/10.1080/23273798.2014.995109.

Frederick, S., 2005. Cognitive reflection and decision making. Journal of Economic Perspectives 19, 25–42. doi:https://doi.org/10.1257/089533005775196732.

Generative Pretrained Transformer, G., Thunström, A.O., Steingrimsson, S., 2022. Can GPT-3 write an academic paper on itself, with minimal human input? URL: https://hal.archives-ouvertes.fr/hal-03701250.

Gibson, E., Bergen, L., Piantadosi, S.T., 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. Proceedings of the National Academy of Sciences 110, 8051–8056. doi:https://doi.org/10.1073/pnas.121643811.

Helwe, C., Clavel, C., Suchanek, F.M., 2021. Reasoning with Transformer-based models: Deep learning, but shallow reasoning, in: 3rd Conference on Automated Knowledge Base Construction, p. https://openreview.net/forum?id=Ozp1WrgtF5_.

Hossain, M.M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., Blanco, E., 2020. An analysis of natural language inference benchmarks through the lens of negation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 9106–9118. doi:10.18653/v1/2020.emnlp-main.732.

Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R.D., Sordoni, A., Courville, A.C., 2021.

Understanding by understanding not: Modeling negation in language models. CoRR abs/2105.03519. URL: https://arxiv.org/abs/2105.03519.

Hupkes, D., Dankers, V., Mul, M., Bruni, E., 2020. Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research 67, 757–795. doi:10.48550/arXiv.1908.08351.

Kahardipraja, P., Madureira, B., Schlangen, D., 2021. Towards incremental transformers: An empirical analysis of transformer models for incremental NLU. CoRR abs/2109.07364. URL: https://arxiv.org/abs/2109.07364.

Kassner, N., Schütze, H., 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. arXiv preprint arXiv:1911.03343 doi:10.18653/v1/2020.acl-main.698.

Kejriwal, M., Santos, H., Mulvehill, A.M., McGuinness, D.L., 2022. Designing a strong test for measuring true common-sense reasoning. Nature Machine Intelligence 4, 318–322. doi:https://doi.org/10.1038/s42256-022-00478-4.

Lieber, O., Sharir, O., Lenz, B., Shoham, Y., 2021. Jurassic-1: Technical details and evaluation. White Paper. AI21 Labs.

Linzen, T., Baroni, M., 2021. Syntactic structure from deep learning. Annual Review of Linguistics 7, 195–212. doi:https://doi.org/10.1146/annurev-linguistics-032020-051035.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. URL: https://arxiv.org/abs/1907.11692.

Makowski, D., Ben-Shachar, M.S., Lüdecke, D., 2019. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. Journal of Open Source Software 4, 1541. doi:https://doi.org/10.21105/joss.01541.

Marcus, G., Davis, E., 2020. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review URL: https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion.

McCoy, R.T., Pavlick, E., Linzen, T., 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007 doi:https://doi.org/10.18653/v1/P19-1334.

Michel, P., Neubig, G., 2018. MTNT: A testbed for machine translation of noisy text, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium. pp. 543–553. URL: https://aclanthology.org/D18-1050.

O'Connor, E., 2015. Comparative illusions at the syntax-semantics interface. Los Angeles, CA: University of Southern California dissertation .

Paape, D., 2021. The role of incremental and superficial processing in the depth charge illusion: Experimental and modeling evidence. URL: psyarxiv.com/jp2ma. psyArXiv preprint.

Paape, D., Vasishth, S., von der Malsburg, T., 2020. Quadruplex negatio invertit? The on-line processing of depth charge sentences. Journal of Semantics 37, 509–555. doi:https://doi.org/10.1093/jos/ffaa009.

Passban, P., Saladi, P.S., Liu, Q., 2020. Revisiting robust neural machine translation: A transformer case study. arXiv preprint arXiv:2012.15710 doi:https://doi.org/10.48550/arXiv.2012.15710.

Patel, A., Bhattamishra, S., Goyal, N., 2021. Are NLP models really able to solve simple math word problems? arXiv preprint arXiv:2103.07191 doi:https://doi.org/10.48550/arXiv.2103.07191.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Reder, L.M., Kusbit, G.W., 1991. Locus of the Moses illusion: Imperfect encoding, retrieval, or match? Journal of Memory and Language 30, 385–406. doi:https://doi.org/10.1016/0749-596X(91)90013-A.

Rosenthal, J.S., 2008. Monty Hall, Monty Fall, Monty Crawl. Math Horizons 16, 5–7. URL: https://www.jstor.org/stable/25678763.

Ryu, S.H., Lewis, R.L., 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. arXiv preprint arXiv:2104.12874 doi:https://doi.org/10.48550/arXiv.2104.12874.

Sobieszek, A., Price, T., 2022. Playing Games with AIs: The Limits of GPT-3 and Similar Large Language Models. Minds and Machines 32, 341–364. doi:https://doi.org/10.1007/s11023-022-09602-0.

Stan Development Team, 2022. Stan Modeling Language Users Guide and Reference Manual. URL: https://mc-stan.org. version 2.26.8.

Thomson, K.S., Oppenheimer, D.M., 2016. Investigating an alternate form of the cognitive reflection test. Judgment & Decision Making 11, 99–113.

Uchendu, A., Ma, Z., Le, T., Zhang, R., Lee, D., 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. arXiv preprint arXiv:2109.13296 doi:https://doi.org/10.48550/arXiv.2109.13296.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. doi:https://doi.org/10.48550/arXiv.1706.03762.

Vos Savant, M., 1997. The power of logical thinking. St. Martin's Press, New York.

Wason, P.C., 1968. Reasoning about a rule. Quarterly Journal of Experimental Psychology 20, 273–281. doi:https://doi.org/10.1080/14640746808400161.

Wason, P.C., Reich, S.S., 1979. A verbal illusion. Quarterly Journal of Experimental Psychology 31, 591–597. doi:https://doi.org/10.1080/14640747908400750.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online. pp. 38–45. URL: https://aclanthology.org/2020.

emnlp-demos.6.

Zhang, Y., Ryskin, R., Gibson, E., 2022. A noisy-channel approach to depth-charge illusions. Preprint available at SSRN: https://ssrn.com/abstract=4130042 .