GOLI: Goal-Optimized Linguistic Stimuli for Psycholinguistics and Cognitive Neuroscience

Shashank Srikant[1,2] **Greta Tuckute**[3] **Sijia Liu**[2,4] **Una-May O'Reilly**[1,2]
[1]CSAIL, MIT [2]MIT-IBM Watson AI Lab BCS, MIT [4]Michigan State University
{shash, gretatu}@mit.edu

Abstract

Experiments in psycholinguistics and the cognitive neuroscience of language rely on linguistic stimuli (sentences) with either specific linguistic properties or which target specific cognitive processes. Such stimuli are generally assembled using manual or semi-manual methods, limiting their quality, quantity, and diversity. We propose GOLI - a gradient-based optimization method that transforms a random sentence into a novel linguistic stimulus which fulfills an experimenter's goal. We apply GOLI to two deliberately different tasks—creating minimal pairs of counterfactual sentences, and deriving constrained stimuli that predict specific responses in human brain regions. We demonstrate how GOLI supports diverse experiment goals and efficiently generates stimuli that are not subject to experimenter biases which may arise from manual methods.

1 Introduction

Experiments in psycholinguistics and cognitive neuroscience of language aim to understand the representations and computations that support human comprehension and production abilities. In comprehension studies in particular, experiments record behavioral (eye-tracking and self-paced reading times) or neural outcomes (electroencephalogram, EEG, and functional magnetic resonance imaging, fMRI) while humans process carefully designed linguistic input (Lai et al., 2015; Shain et al., 2019; Wehbe et al., 2021; Heilbron et al., 2022). Similar methods have been recently extended to probe how computational and language models process such linguistic input as well (Warstadt et al., 2019; Jeretic et al., 2020).

Linguistic stimuli (sentences) used in such experiments are typically hand-constructed (Martín-Loeches et al., 2012; Lai et al., 2015). While handcrafting provides the experimenter with significant control over the goals of the constructed stimuli (*e.g.* the stimuli should adhere to grammatical rules such as subject-verb agreement or convey information about a particular topic), assembling a sizeable set of stimuli is time- and resource-intensive. Another important concern is the diversity of the resulting stimuli—they are generally limited by the experimenter's vocabulary and assumptions. Experimenters could easily be misguided by their top-down assumptions or an inaccurate formulation of the hypothesis being tested and use sets of words, sentence structures, or concepts that are biased in some way. Studies have shown how such biased stimuli have led to incorrect scientific conclusions (Chaves and Dery, 2018; Siegelman et al., 2019).

	Specify goals?	Data-driven?	Automate?
Handcrafted	✓	✗	✗
Template-based	✓	✗	✓
Naturalistic corpora	✗	✓	✓
GOLI (this work)	✓✓	✓	✓

Table 1: GOLI automates generating stimuli which satisfy experimenter-supplied goals. It handles a broader set of goals than handcrafted and template-based methods while being data-driven.

Another approach to constructing stimuli is to use templates (Warstadt et al., 2020). An experimenter defines templates which structure and constrain stimuli. Stimuli are generated with a template by filling them with words sampled from different naturalistic vocabularies (*e.g.* of parts of speech). While this automates the process of creating stimuli and allows goal specification similar to handcrafting (Warstadt et al., 2019; Jeretic et al., 2020), the generated stimuli are still constrained to the experimenter's notions of a 'correct' template. Templates also often generate unnatural and incorrect sentences, which then need to be manually filtered out by the experimenter.

Yet another popular approach, which we refer to as the search-based method (SBM), involves randomly sampling from naturalistic text corpora (Kennedy et al., 2013; Nastase et al., 2021; Heilbron et al., 2022). While this approach circumvents biases potentially introduced in handcrafted and

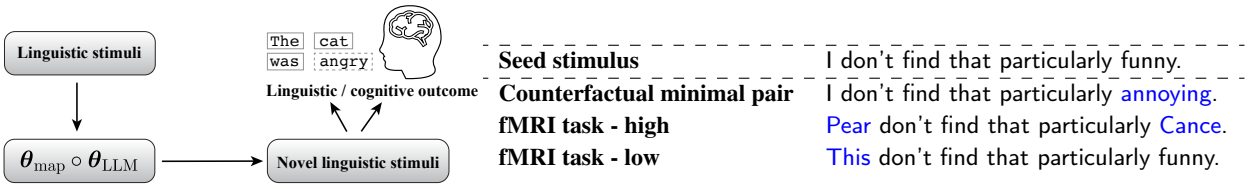

Figure 1: Overview. GOLI transforms a seed linguistic stimulus into a novel stimulus which either contains a desired linguistic property or elicits a desired cognitive outcome. It uses a language model ($\boldsymbol{\theta}_{\text{LLM}}$) to represent the seed sentence, a mapping model ($\boldsymbol{\theta}_{\text{map}}$) to map it to the desired property, and uses a gradient-based method to modify the seed sentence (propagates gradients through the composed model $\boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}$) into a novel one. The table (right) shows an example of stimuli generated from a seed stimulus for the three objectives we demonstrate in this work.

template-guided stimuli, it has no efficient way to identify goals like targeted phenomena (*e.g.* sentiment polarity, surprisal, agreement, garden-path effects, licensing, gross syntactic expectation, center embedding, long-distance dependencies, and others mentioned in Marvin and Linzen (2018); Hu et al. (2020)) or infrequent phenomena (Bresnan and Kanerva, 1989; Losiewicz, 1992; Hoffmann, 2004; Ross, 2018; Turner, 2020) within the large corpora that need to be sampled to find suitable sentences. Prior work has manually edited such sampled texts and inserted linguistic properties of interest to create naturalistic stimuli (Futrell et al., 2020). Further, sampling from corpora does not enable creating minimal pair stimuli–pairs of sentences that differ only in a very specific linguistic property. Minimal pairs are extensively used in psycholinguistic and language research to isolate causal attributes of behavior (Bemis and Pylkkänen, 2011; Kochari et al., 2018; Parrish and Pylkkänen, 2021). Thus, minimal pairs, targeted and infrequent phenomena are mostly studied through stimuli that experimenters handcraft or create from templates.

It then seems that handcrafted and template-based stimuli offer significant control over the created stimuli, but may introduce undesirable experimenter-biases and are also time- and resource-expensive to create. On the other hand, automated sampling from naturalistic corpora avoids experimenter-biases, but is not suited to test targeted or infrequent phenomena (Table 1). We propose a method that is data-driven, automated, efficient, and can fulfill a large set of experimenter goals which includes targeted and infrequent phenomena.

GOLI (Figure 1) is an automated approach to generate **g**oal-**o**ptimized **l**inguistic stimul**i**. GOLI starts from a seed sentence and modifies it until it satisfies experimenter-specified outcomes (linguistic or cognitive) by solving a gradient-based optimization formulation. Constraints on the generated stimuli can be easily enforced via the optimization formulation, providing the necessary control over stimuli that is typically offered by handcrafting and template-based approaches. In fact, we show that GOLI-generated stimuli can satisfy a broader set of goals than what handcrafting or template-use satisfies. GOLI is data-driven and is not bound to inductive biases of experimenters since it relies on data-driven computational models to transform the seed sentence and optimize it to achieve the desired goal.

We demonstrate GOLI on two deliberately different tasks: generation of minimal-pair counterfactuals and the generation of stimuli which predict specific responses in the human brain. These tasks differ in the nature of their outcomes, desired goals, and the constraints imposed on the generated sentences. Across these differences, we show that GOLI can successfully and easily model the various constraints posed by these tasks and efficiently generate novel stimuli, outperforming other methods currently used to prepare stimuli for such tasks. We will make all the code and data related to this work publicly available.

2 Problem description

In this section, we state the assumptions underlying GOLI. We introduce notation that we use in the rest of this work and then state the problem we solve.

GOLI assumes an experiment uses a set of linguistic stimuli to stimulate either a language property or a cognitive outcome. Further, it assumes the property or outcome is quantifiable, and a statistical model can predict its values corresponding to an input linguistic stimulus. For example, if an experiment outcome measures logical inaccuracy (linguistic outcome) or reading times (cognitive outcome) in a sentence, then GOLI assumes a

model which maps a random sentence stimulus $\mathcal{S}$ to a quantifiable measure of either the extent of inaccuracy (former) or the time taken to read a sentence (latter). We denote the mapping model as $\boldsymbol{\theta}_{\text{map}}$- the weights it is parameterized by, and the linguistic property or cognitive outcome as $y \in \mathbb{R}$. If the model does not initially exist, it can be learned as a preliminary step. We discuss in Section 6 the case where learning such a mapping model is not feasible.

Generally, $\boldsymbol{\theta}_{\text{map}}$ is trained to predict $y \in \mathbb{R}$ from representations $\boldsymbol{r} \in \mathbb{R}^d$ of the input stimulus $\mathcal{S}$. We use a large language model $\boldsymbol{\theta}_{\text{LLM}}$ without loss of generality, *i.e.* $\boldsymbol{r} = \boldsymbol{\theta}_{\text{LLM}}(\mathcal{S})$. Any alternate representation which is similarly differentiable can also be used, and this is yet another strength of GOLI.

$\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ denotes a sentence stimulus consisting of n tokens $\boldsymbol{x}_i \in \{0,1\}^{|V|}$, where $\boldsymbol{x}_i$ is a token from the set V, the vocabulary of permissible tokens which the LLM processes. We provide concrete examples describing the structure of $\boldsymbol{x}_i$ in Section 3. Following this notation, we have $y^{\text{pred}} = \boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}(\{\boldsymbol{x}_i\}_{i=1}^n)$, where $\circ$ denotes model composition, *i.e.* $\mathcal{S}$ is first input to $\boldsymbol{\theta}_{\text{LLM}}$, whose output is then input to $\boldsymbol{\theta}_{\text{map}}$. GOLI allows any number of and any kind of such models to be composed together.

We study the problem of generating a sentence $\mathcal{S}^{\text{gen}}$ by transforming $\mathcal{S}$ into $\mathcal{S}^{\text{gen}}$ in a way such that y^{pred}, the prediction of $\boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}$, is *close* to y^{desired}, an outcome specified by the experimenter.

3 Method

In this section, we motivate our method using an example. We then show how the problem of generating novel linguistic stimuli can be cast and solved as a problem in first-order (gradient-based) optimization. Consider the sentence:

`Running slow makes me very happy.`

which when input to the model $\mathcal{M} = \boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}$ predicts the sentiment $y^{\text{pred}} = \text{positive}$ ($\boldsymbol{\theta}_{\text{map}}$ is a binary sentiment classifier). Further, let the vocabulary V consist of the tokens:

$$V = \left\{ \begin{array}{c} \texttt{Run}, \texttt{Sit}, \texttt{Stand}, \texttt{ing}, \texttt{ed}, \texttt{slow}, \texttt{happy}, \\ \texttt{car}, \texttt{me}, \texttt{you}, \texttt{very}, \texttt{makes}, \texttt{well}, \cdot, ? \end{array} \right\} \tag{1}$$

The sentence has six space separated words with a terminating period symbol in it. Let's assume tokenizing this sentence generates the following eight tokens:

$$\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^8 = \{\text{Run}, \text{ing}, \text{slow}, \text{makes}, \text{me}, \text{very}, \text{happy}, \cdot\} \tag{2}$$

Further assume we desire a novel sentence whose prediction is $y^{\text{desired}} = \text{negative}$. The *generation* of sentences that we describe in this work involves modifying a subset of the eight tokens in $\mathcal{S}$ in a way that results in the model $\mathcal{M}$ predicting a value that is *closer to* y^{desired} *i.e.* is transformed to a negative sentiment sentence.

Two important questions need to be addressed to generate such sentences. First, which tokens or *sites* in the sentence should be modified? Of the n sites, if we are allowed to choose at most k sites, which set of $\leq k$ sites would best guide $\mathcal{M}$ to the desired prediction. We call this the **site selection problem**. Second, how should a token at a given site be modified, and what should the modified token be? We call this the **site perturbation problem**.

Site selection. The benefit of isolating site selection as a distinct sub-problem is it supports complex formulations, such as constraining and optimizing specific sites. For instance, site selection and site perturbation can be jointly optimized: an optimal site can depend on the optimal token found by the site perturbation sub-problem and vice versa.

We employ a simple site selection strategy in this work. To select k specific sites from the available n sites, we follow the gradient-based word importance method from Wallace et al. (2019). The method first sorts the tokens $\boldsymbol{x}_i$ in decreasing order of the magnitude of the gradient on the output y with respect to $\boldsymbol{x}_i$. The top k magnitude tokens are selected as the sites to perturb, since they impact the output the most.

Site perturbation. At a given site, there are three operations which would modify the token: *replace* an existing token with another token, *insert* another token at the site–either before or after the token present at the site, or *retain* the token at the site unaltered (this is equivalent to not selecting a site to carry out a modification operation).

We discuss only replace modifications, since deletion and insertion reduce to replace modifications. Deletion is replacing with an empty token, and an insertion is a replace modification applied to a dummy token inserted at a site.

A replacement token modification strategy requires a replacement token $\boldsymbol{u}_i \in \{0,1\}^{|V|}$ to be

identified from the set of tokens V. For example, if the selected site for replacement in (2) is 3: `slow`, then a possible sentence could result from replacing `slow` with `car` (token 8 sampled from the vocabulary V in (1)), resulting in the previously unseen sentence: `Running car makes me very happy.` Increasing the number of sites to be perturbed results in a sentence that is very different from the original sentence. Similarly, inserting new tokens can introduce new words and phrases.

Selecting an appropriate token from a vocabulary is a combinatorially expensive problem: it takes $O(|V|^k)$ time to select tokens at k sites from the vocabulary V, since each site offers $|V|$ possible tokens to choose from. The aim is to thus tractably select a replacement token $\boldsymbol{u}_i$ at a site i such that the predicted activation y^{pred} matches y^{desired}. We set this up as a combinatorial optimization problem and solve for $\boldsymbol{u}_i$.

3.1 Solution formulation

Based on the site selection perturbation formulation, we formally define the described replacement operation.

For a sentence $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$and a set K of k site indices to perturb, wherein each site index j satisfies $(1 \leq j \leq n)$, we formalize site perturbation in the following way: we introduce a one-hot vector $\mathbf{u}_i \in \{0,1\}^{|V|}$ to encode the selection of a token from V which would serve as the replaced token for at a chosen site.

If the j^{th} entry $[\mathbf{u}_i]_j = 1$ and $i \in K$, then the j^{th} token in V is used as the modified token which will replace $\boldsymbol{x}_i$ at the site i. We also impose the constraint $\mathbf{1}^T\mathbf{u}_i = 1$, implying that only one perturbation is performed at $\boldsymbol{x}_i$.

Let vector $\mathbf{u} \in \{0,1\}^{k\times|V|}$ denote k different $\mathbf{u}_i$ vectors, one for each token $i \in K$, where $|K| = k$. We then define a newly generated or transformed sentence $\mathcal{S}^{\text{gen}}$ as comprising tokens $\{\boldsymbol{x}_i^{\text{gen}}\}_{i=1}^n$, where each $\boldsymbol{x}_i^{\text{gen}}$ is defined as:

$$\boldsymbol{x}_i^{\text{gen}} = \begin{cases} \boldsymbol{u}_i, & \forall i \in K, \text{ where } \mathbf{1}^T\mathbf{u}_i = 1,\ \mathbf{u}_i \in \{0,1\}^{|V|} \\ \boldsymbol{x}_i, & \forall i \notin K \end{cases} \tag{3}$$

We solve the following objective to obtain $\boldsymbol{u}$:

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{minimize}} \quad & \ell(\boldsymbol{u}; \boldsymbol{x}, \boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}) \\ \text{subject to} \quad & \text{constraints in (3)} \end{aligned} \tag{4}$$

where ℓ denotes an appropriate loss function which encodes the desired cognitive outcome. Algorithm 1 in Appendix 1 describes how GOLI solves this optimization problem.

Incorporating additional constraints. A variety of constraints on $\mathcal{S}^{\text{gen}}$ can be imposed by using appropriate loss functions and vocabulary subsets to find candidate replacement tokens from. Section 4.1 discusses how a loss function can be modified to generate stimuli that are grammatically likely. Similarly, other site-specific constraints like capitalizing the first word of a sentence or the last token being a punctuation can be ensured by assigning different subsets of naturalistic vocabularies when solving the site perturbation problem. See notes in Appendix A for details.

4 Experiments & Results

We demonstrate and assess GOLI on two tasks—constructing minimal pairs of counterfactual sentences for sentiment analysis, and an fMRI-based targeted brain response task. These two tasks differ in the questions they ask, the architectures used to encode sentences (BERT vs. GPT2-XL), the outcome of $\boldsymbol{\theta}_{\text{map}}$ (sentiment-class classification vs. brain region response predictions), the set of constraints imposed on the generated sentences, and consequently the loss functions needed to generate sentences. We describe these details below.

4.1 Counterfactual minimal-pair task

Training-data augmentation with *counterfactuals* (CFs) has been proposed as a way to mitigate out-of-domain generalization of NLP models (Levesque et al., 2012; Kaushik et al., 2020). Rooted in causal learning, a CF in the context of NLP models is designed to study the change in an NLP model's prediction following an intervention to its input text, generally implemented as minimal edits to the text. Such minimal changes to different input features help ascertain the causal role of these features in a model's prediction. Producing such CF stimuli though can be challenging, and resembles the process of developing minimal-pair stimuli in psycholinguistics experiments discussed in Section 1, Introduction.

Recent work however has explored automated generation algorithms for such CFs (Wang and Culotta, 2021; Yang et al., 2021; Howard et al., 2022). Notably, Howard et al. (2022), the state-of-the-art, propose a system to generate CFs for sentiment analysis on the SST-2 IMDB movie reviews dataset (Socher et al., 2013). The CF reviews they generate have the opposite sentiment as the original stimulus, while being *natural* in a way that would resemble

CFs generated by human experts (Kaushik et al., 2020; Gardner et al., 2020). We demonstrate how GOLI can be setup for this task by appropriately customizing the loss function and constraints in the formulation in Eq. (4).

Objective. We use a BERT-based sentiment classifier fine-tuned on the SST-2 task (binary classification) as our mapping model, $\boldsymbol{\theta}_{\text{map}}$. In this case, BERT serves as $\boldsymbol{\theta}_{\text{LLM}}$. Our objective then is to generate modifications to a given sentiment review such that $\boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}$ flips its prediction on the modified sentence and the modified sentence is *close* to the original sentence.

Loss, Constraints. We use the standard binary cross entropy (Loss_{BCE}) as our loss function as it allows us to specify the desired class we want $\boldsymbol{\theta}_{\text{map}}$ to predict in the binary sentiment classification task:

$$\ell(\boldsymbol{u}) = \text{BCE}(\boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}(\boldsymbol{u}), y^{\text{desired}}) \quad (5)$$

where y^{desired} is 0 for the negative sentiment class and 1 for positive. An alternate loss function which we do not try and defer to future work is ensuring general fluency and grammaticality of the generated sentences (Goswamy et al., 2020) by introducing two additional loss terms:

$$\ell(\boldsymbol{u}) = \text{Loss}_{\text{BCE}} + \text{Loss}_{\text{BOW}} + \text{KL}(H(\boldsymbol{u}), H(\boldsymbol{x})) \quad (6)$$

where $\text{Loss}_{\text{BOW}} = -\log(\sum(p_i u_i))$ penalizes selecting a u_i whose bag-of-words probability p_i is low or unlikely, and $\text{KL}(\cdot)$ is the KL-divergence between the intermediate decoder representation $H(\cdot)$ of the modified input $\boldsymbol{u}$ and the unmodified, original input stimuli $\boldsymbol{x}$. The KL-term ensures the distribution of each generated token u_i is similar to the original token x_i. To ensure minimal pairs, we select a maximum of two sites to be modified in each original sentence.

Evaluation. We evaluate our generated stimuli against the CF-generation method introduced in Howard et al. (2022). They work with a subset of the IMDB dataset (training set, N=8173). For each sentence in the training set, they generate a CF using the following complex setup: first, they provide a prompt (a part of the original stimulus) and a desired sentiment (positive or negative) to a pre-trained adaptation of the GPT-2 model (first proposed by Gururangan et al. (2020)). The model is optimized to generate reviews which complete the prompt and are of the desired sentiment polarity. Second, they use a constrained decoding algorithm (first proposed by Lu et al. (2021)) with the adapted GPT-2 model to ensure the generated sentence remains a minimal pair to the given input sentence. They augment the generated CFs with the training set and fine-tune a RoBERTa-based classifier. They augment with CFs generated from two settings–**NeuroCF-1g**: where they provide just the first word in the original stimulus as a prompt to their system, and **NeuroCF-np**: which selects a subset of the original sentence as a prompt. They evaluate the fine-tuned model on three out-of-distribution test sets - **Test-set 1**: another subset from the IMDB dataset (N=2245), **Test-set 2**: a dataset from Kaushik et al. (2020) (N=488), **Test-set 3**: a dataset from Gardner et al. (2020) (N=488). Accuracy on the test set and descriptive metrics of the generated sentences (described below) serve as merit indicators for the effectiveness of the generated CFs.

To evaluate GOLI, we generate minimal-pairs for each sentence in their training set, and fine-tune and evaluate their RoBERTa model using an augmented dataset containing GOLI-generated sentences. We compare the accuracy of the RoBERTa model against the CFs generated by NeuroCF-1g and NeuroCF-np. Expert-crafted CFs generated by Kaushik et al. (2020) serves as an upper bound for performance in our evaluation.

Results. Table 2, *Top* shows the accuracy of the CF-augmented RoBERTa models on the three test sets, averaged over 10 random runs. Across the test sets, we see that GOLI consistently outperforms NeuroCF-1g, and is comparable in its performance to NeuroCF-np. We highlight that GOLI uses a very generic formulation for transforming an input sentence to meet a desired goal, and despite the generality, is capable of matching the performance of a bespoke solution like NeuroCF.

Further, to evaluate the quality of the generated sentences, Howard et al. (2022) use two metrics (Table 2, *Bottom*): MoverScore (Zhao et al., 2019) and perplexity. A MoverScore is computed between the generated counterfactual and the original sentence. A low score suggests the two sentences are similar. We see that GOLI generates sentences of similar MoverScores to NeuroCF-1g and comparable to NeuroCF-np. However, the average counterfactuals created by experts seem to be fairly farther off from their respective original sentences, suggesting room for automated stimuli generation methods to improve.

	Test-set 1	Test-set 2	Test-set 3
GOLI	$\mathbf{93.37_{0.01}}$	$\mathbf{94.94_{0.53}}$	$\mathbf{92.14_{0.05}}$
NeuroCF-1g	$92.75_{0.03}$	$93.10_{0.06}$	$89.27_{0.04}$
NeuroCF-np	$93.10_{0.05}$	$\mathbf{94.74_{0.08}}$	$\mathbf{91.18_{1.17}}$
Expert-crafted	$92.63_{0.48}$	$97.34_{0.37}$	$95.22_{0.45}$

	MoverScore	Perplexity
GOLI	0.45	39.2
NeuroCF-1g	0.46	14.1
NeuroCF-np	**0.20**	**12.7**
Expert-crafted	0.70	19.3

Table 2

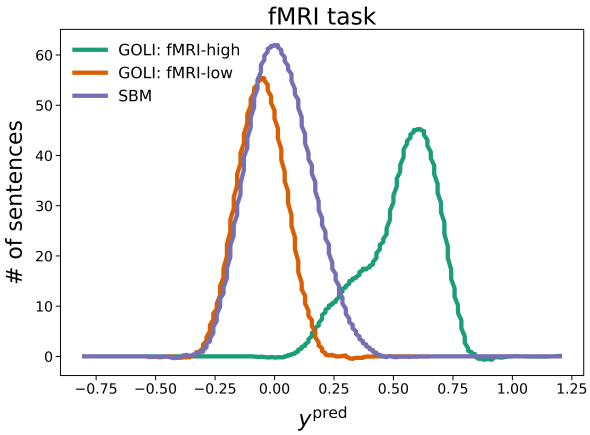

Figure 2

Results. Table 2: GOLI-generated counterfactual sentences vs. NeuroCF (Howard et al., 2022) vs. expert-crafted CFs (serves as an upper bound; Kaushik et al. (2020)), augmented with training data to improve the robustness of a RoBERTa-based sentiment analysis classifier. *Top.* Accuracy (percent) on three unseen test sets. Std. dev. across 10 runs mentioned as subscripts. *Bottom.* MoverScore and Perplexity, two quality measures of the generated sentences. **Figure 2**: A histogram of y^{pred} from sentences sampled using a search-based method (SBM), and those generated by GOLI optimized on two objectives: fMRI-high and fMRI-low.

Perplexity is a measure of how likely a given sentence is, which we evaluate on GPT-J, a domain-agnostic model. A lower score suggests a higher likelihood of the sentence. Table 2 shows that GOLI produces sentences with comparatively higher perplexity. This was expected since we do not incorporate felicity-related loss terms as described in Eq. (6). As seen in previous works *e.g.* Goswamy et al. (2020), a modified loss incorporating felicity should improve perplexity, making it comparable to NeuroCF. We defer this verification to future work.

4.2 fMRI task

A characterization of the sentences that activate the language network in the human brain (Fedorenko et al., 2010) remains an open question. The fMRI task is to thus generate sentences that predict a desired brain response in the language network. To do so, we set up an fMRI experiment where we first collect brain responses of participants reading random sentences. We then fit a linear model $\boldsymbol{\theta}_{\text{map}}$ to predict these brain responses from LLM representations of the sentences. Given a trained $\boldsymbol{\theta}_{\text{map}}$, we use GOLI to generate novel sentences using a separate dataset of seed sentences.

Objectives. GOLI is provided two separate objectives: to generate sentences that predict high responses in the language network (fMRI-high; $y^{\text{pred}} \geq +0.4$) and to generate another set of sentences which predict low brain responses (fMRI-low; $y^{\text{pred}} \leq -0.3$).

Setup. We invited participants (N=5) to passively read a set of 1000 diverse, corpus-extracted 6-word sentences in an event-related design (referred henceforth as training set, see Appendix C.2 for details). We pre-process and select responses from language-selective areas of the brain (Appendix C.4). Voxels (3D pixels) from these areas were averaged within and across each participant to yield a scalar language network response value associated with each sentence stimulus. The range of brain responses values predicted by $\boldsymbol{\theta}_{\text{map}}$ on the training set across participants was $[-0.47, +0.54]$. These values represent z-scores of brain responses and hence are both positive and negative–they represent relative magnitudes of brain responses (see Appendix C.3). Based on the training set, we interpret negative values ≤ -0.3 as a low response and $\geq +0.4$ as high. A linear model $\boldsymbol{\theta}_{\text{map}} \in \mathbb{R}^d$ was learned to predict these average brain responses across participants from GPT2-XL representations $r \in \mathbb{R}^d$ of sentences ($d = 1600$; details in Appendix C.5).

Loss, Constraints. We model the fMRI-high and fMRI-low objectives with a squared-loss function:

$$\ell(\boldsymbol{u}) = (y^{\text{desired}} - \boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}(\boldsymbol{u}))^2 \quad (7)$$

To generate sentences with high positive and high negative desired predicted responses, we set y^{desired} to $+1.2$ and -0.8 respectively, values slightly beyond the maximum and minimum predicted values seen on the training set.

The number of words in each sentence in $\mathcal{S}^{\text{gen}}$ was constrained to contain six space separated

words, terminated by a punctuation, with the first word capitalized. These constraints ensure avoiding confounding effects of sentence length and unusual orthography (*e.g.* lack of capitalization, no end-of-sentence punctuation) in fMRI recordings.

Search-based method (SBM). We compare GOLI to a search-based approach which is routinely used to assemble language stimuli for such a task: exhaustively searching a large, unseen naturalistic corpus of text. Each sentence from such a corpus is individually tested against the desired goal. Unlike GOLI, the search-based method does not modify any sentences in the set of sentences it searches through–it just filters and selects those that achieve the desired goal. A key drawback of this method is that a prohibitively large corpus may then need to be sampled from should a small proportion of natural sentences meet the desired goals (fMRI-high, fMRI-low objectives for this task). Further, as discussed in Section 1, natural sentences may not necessarily meet the desired goals for this task—the brain may well be responsive to a very particular subset of sentences and sentence structure patterns. SBM over a naturalistic corpus then threatens the discovery of such patterns.

Evaluation criteria. We select 1500 sentences extracted from various, diverse text corpora (referred henceforth as test set; see Appendix C.1) to evaluate SBM and GOLI. We demonstrate the utility of GOLI over SBM along two dimensions: **sample efficiency:** the number of sentences needed in a corpus which when sampled results in the desired number of linguistic stimuli which satisfy the desired goal, and **solution diversity:** whether the sentences generated by GOLI achieves (or outperforms) the desired goal in both quality and quantity.

Results. Figures 2 summarizes our results. We plot the distribution of y^{pred}–predictions made by $\boldsymbol{\theta}_{\text{map}}$—on processing sentences produced by GOLI and by SBM on the test set (N=1500). We see that while most randomly sampled sentences (marked in blue, SBM) in the test set elicit average brain responses (around the z-score 0 of y^{pred}), 0.2% ($\frac{3}{1500}$) sentences elicit high brain responses ($y^{\text{pred}} \geq 0.4$). In sharp contrast, we find that GOLI, when optimized for fMRI-high (green curve in Figure 2), generates 80% ($\frac{838}{1049}$) high-response prediction sentences. We work with 1049 GOLI-generated sentences because 451 (1500–1049) of those failed the automated filters (Appendix C.7).

Comparing the two methods on the fMRI-low objective, we see 0.2% ($\frac{4}{1500}$) sentences in SBM elicit low brain responses ($y^{\text{pred}} \leq 0.3$). GOLI sentences optimized for fMRI-low (red curve, Figure 2) yield an interesting observation: despite the sentences being optimized to minimize their predictions, we find that, unlike in the fMRI-high objective, GOLI is unable to generate sentences that predict values significantly lower than those found on the test set. GOLI generates 0.8% ($\frac{8}{990}$) low-response sentences, although the overall average y^{pred} drops to -0.05 in fMRI-low, from $+0.02$ in the SBM setting.

These results suggest that in order to assemble a total of 500 high or low activity sentences (a reasonable estimate of the number of unique sentences needed in an fMRI experiment), one would have to significantly increase the number of sentences to sample from when using SBM, which increases the compute and data-needs to run such experiments. For the fMRI-high objective especially, we see that the number of sentences required to sample from may be significantly higher than 20x since we never see sentences greater than 0.40 on the training set, while GOLI reveals that perhaps high-response sentences are those that predict ≥ 0.65.

Further, we find that GOLI generates fairly unusual sentence structures for the fMRI-high objective (Fig 1 and Appendix C.9), while generating more 'regular-looking' sentences for the fMRI-low objective. This is an interesting result which would not have been discovered had we sampled from regular text corpora via SBM. Collecting brain data for the GOLI-generated sentences, and analyzing the implication of these *unusual* fMRI-high sentences on the neuroscience of language-responsive brain regions is left for future work. We discuss this more in Section 6.

5 Related work

Automated methods to create linguistic stimuli have been explored in the context of comprehension difficulty (Boyce et al., 2020), probing language models (Warstadt et al., 2020) and generating counterfactuals (Wang and Culotta, 2021; Yang et al., 2021; Howard et al., 2022). Boyce et al. (2020) present a use-case that GOLI directly addresses. They use a language model to find high surprisal words to use as distractors in their stimuli. However, their setup does not generalize to the use-cases we cover in this work and is limited to finding high surprisal words, a trivial objective in

our formulation. We discuss the sentence templates from Warstadt et al. (2020) in Section 1 in detail.

Methods used in counterfactual generation resemble methods used for adversarial attacks (Gao et al., 2018; Ribeiro et al., 2018; Pezeshkpour et al., 2019; Ren et al., 2019; Yoo et al., 2020), which come closest to the stimuli generation algorithm (Algorithm 1) we propose in this work. Other works have proposed using a GAN (Zhao et al., 2018) and controlling the output of model decoders (Lu et al., 2021; Dathathri et al., 2020; Goswamy et al., 2020) to generate naturalistic sentences. GOLI's formulation subsumes the methods used in these works. None of these works formulate the problems of site-selection and site-perturbation we introduce (Section 3). Works on adversarial attacks end up implicitly solving just the site-selection problem using gradients, replacing the selected sites with randomly selected words. Their primary objective is to flip a model's prediction while maintaining the semantic meaning of a sentence, which is just one of the many desired goals that can be accomplished using GOLI. Similarly, prior works on constrained decoding do not allow token-level constraints to be imposed and optimized for, rendering them inadequate for modeling a large class of tasks, including the fMRI task we evaluate in this work.

6 Discussion

We demonstrate the effectiveness of GOLI in generating stimuli in two distinct experiment settings. The minimal pairs task demonstrates how easily GOLI can be employed to generate a tightly constrained set of stimuli. It is infeasible to generate such stimuli pairs using either templates or by looking in naturalistic corpora.

In the fMRI task, GOLI helped generate stimuli that are predicted to elicit high or low responses in the language network in the human brain. Knowing which stimuli elicit maximal activity in neurons can provide useful insight into the representations and computations that brain areas perform (Hubel and Wiesel, 2009; Bashivan et al., 2019; Xiao and Kreiman, 2020). The task—of predicting specific brain responses—is unique, and we demonstrate how such goals can successfully be encoded in GOLI, which even handcrafting does not support. We highlight the innovative use of $\boldsymbol{\theta}_{\text{map}}$ as a surrogate model to quantify and predict the goal, which GOLI then uses to guide stimuli generation. It is possible that the *unusual* fMRI-high sentences generated by GOLI are high on surprisal, since only a few words are abruptly modified in $\mathcal{S}^{\text{gen}}$. This hypothesis can be confirmed in a follow-up fMRI study where participants' brain responses to GOLI-generated sentences are compared to their responses to other meaningful sentences with one or two words randomly swapped out. We will investigate this in future work.

GOLI in other domains. GOLI can potentially be used to generate inputs in domains beyond language, such as tasks in memory (Barr et al., 2016), motor-control (Srivastava et al., 2022), planning in robotics (Aznan et al., 2019), and AI (Chollet, 2019) or in engineering such as circuit design (Liu et al., 2018), processor design (Ritter and Hack, 2020) and electric machine design (Wang et al., 2017). In each of these works, the authors attempt to generate hand-crafted stimuli or inputs in a discrete domain (similar to linguistic stimuli) that are required to satisfy a suite of constraints their respective problem domains pose.

7 Limitations

Not all tasks in psycholinguistics will readily benefit from GOLI. Many such tasks imaginably may want to enforce constraints which GOLI may not facilitate. For instance, state of the art in constrained decoding performs poorly for long strings. Hence, generating stimuli which are long, meaningful while being constrained may be a challenge for GOLI.

Moreover, it may not be easy to learn $\boldsymbol{\theta}_{\text{map}}$ for every problem. Access to a differentiable $\boldsymbol{\theta}_{\text{map}}$ which predicts the desired outcome well is essential. The task of learning $\boldsymbol{\theta}_{\text{map}}$, which requires a labeled data-set, could end up being as expensive as handcrafting the stimuli in some cases. One way to alleviate this limitation is through black-box optimization (Liu et al., 2020). It enables optimizing Eq. (4) even in the absence of $\boldsymbol{\theta}_{\text{map}}$ and gradients.

GOLI is not immune to experimenter biases. The choice of a loss function and its components, such as the one described in Eq. (6), biases the generated stimuli. Moreover, optimal results obtained by the site selection and site perturbation sub-routines are affected by the inductive biases used to train $\boldsymbol{\theta}_{\text{map}}$ and $\boldsymbol{\theta}_{\text{LLM}}$. Such data-driven generation methods only ensure the stimuli are not limited by an experimenter's word choice and other top-down assumptions of sentence structures.

References

John Ashburner and Karl J. Friston. 2005. Unified segmentation. *NeuroImage*, 26:839–851.

Nik Khadijah Nik Aznan, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. 2019. Using variable natural environment brain-computer interface stimuli for real-time humanoid robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4889–4895. IEEE.

Rachel Barr, Alecia Moser, Sylvia Rusnak, Laura Zimmermann, Kelly Dickerson, Herietta Lee, and Peter Gerhardstein. 2016. The impact of memory load and perceptual cues on puzzle learning by 24-month olds. *Developmental Psychobiology*, 58(7):817–828.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. 2019. Neural population control via deep image synthesis. *Science*, 364.

Douglas Knox Bemis and Liina Pylkkänen. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, 31:2801 – 2814.

Veronica Boyce, Richard Futrell, and Roger P Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Joan Bresnan and Jonni M Kanerva. 1989. Locative inversion in chichewa: A case study of factorization in grammar. *Linguistic inquiry*, pages 1–50.

Rui Pedro Chaves and Jeruen E. Dery. 2018. Frequency effects in subject islands. *Journal of Linguistics*, 55:475 – 521.

François Chollet. 2019. On the measure of intelligence. *CoRR*, abs/1911.01547.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Lit. Linguistic Comput.*, 25:447–464.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94.

Karl J. Friston, John Ashburner, Chris D. Frith, J. B. Poline, Jon D. Heather, and Richard S. J. Frackowiak. 1995. Spatial registration and normalization of images. *Human Brain Mapping*, 3.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Asher Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2020. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63 – 77.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2022. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119.

Sebastian Hoffmann. 2004. Are low-frequency complex prepositions. *Corpus approaches to grammaticalization in English*, 13:171.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

David H. Hubel and Torsten N. Wiesel. 2009. Republication of the journal of physiology (1959) 148, 574-591: Receptive fields of single neurones in the cat's striate cortex. 1959. *The Journal of physiology*, 587 Pt 12:2721–32.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Alan Kennedy, Joël Pynte, Wayne S. Murray, and Shirley-Anne S. Paul. 2013. Frequency and predictability effects in the dundee corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, 66:601 – 618.

Arnold R Kochari, Ashley Glen Lewis, Jan-Mathijs Schoffelen, and Herbert Schriefers. 2018. Semantic and syntactic composition of minimal adjective-noun phrases in dutch: An meg study. *Neuropsychologia*, 155.

Vicky Tzuyin Lai, Roel M. Willems, and Peter Hagoort. 2015. Feel between the lines: Implied emotion in sentence comprehension. *Journal of Cognitive Neuroscience*, 27:1528–1541.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. 2020. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54.

Xiaoping Liu, Jihong Ren, Wendem Beyene, Simon Ku, Chin Hong Heah, and Sherman Hsu. 2018. Design optimization and accurate extraction of on-die decoupling capacitors for high-performance applications. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*, pages 1712–1719. IEEE.

Beth L Losiewicz. 1992. *The effect of frequency on linguistic morphology*. The University of Texas at Austin.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.

Manuel Martín-Loeches, Anabel Fernández, Annekathrin Schacht, Werner Sommer, Pilar Casado, Laura Jiménez-Ortega, and Sabela Fondevila. 2012. The influence of emotional words on sentence processing: Electrophysiological and behavioral evidence. *Neuropsychologia*, 50:3262–3272.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher John Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher A. Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily T. Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8.

Alfonso Nieto-Castanon. 2020. Handbook of functional connectivity magnetic resonance imaging methods in conn.

Richard Charles Oldfield. 1971. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9 1:97–113.

Alicia Parrish and Liina Pylkkänen. 2021. Conceptual combination in the latl with and without syntactic composition. *Neurobiology of Language*, 3:46–66.

Douglas B. Paul and Janet M. Baker. 1992. The design for the wall street journal-based csr corpus. *2nd International Conference on Spoken Language Processing (ICSLP 1992)*.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3336–3347.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick Norris Kay. 2022. Improving the accuracy of single-trial fmri response estimates using glmsingle. *eLife*, 11.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Association for Computational Linguistics (ACL)*, pages 856–865.

Fabian Ritter and Sebastian Hack. 2020. Pmevo: portable inference of port mappings for out-of-order processors by evolutionary optimization. In *Proceedings of the 41st ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 608–622.

Daniel Ross. 2018. Small corpora and low-frequency phenomena: try and beyond contemporary, standard english. *Corpus*, (18).

Cory Shain, Idan Asher Blank, Marten van Schijndel, Evelina Fedorenko, and William Schuler. 2019. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Matthew Siegelman, Idan A Blank, Zachary Mineroff, and Evelina Fedorenko. 2019. An attempt to conceptually replicate the dissociation between syntax and semantics during sentence comprehension. *Neuroscience*, 413:219–229.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Megha Srivastava, Erdem Biyik, Suvir Mirchandani, Noah Goodman, and Dorsa Sadigh. 2022. Assistive teaching of motor control tasks to humans. In *Advances in Neural Information Processing Systems*.

Stefan Thesen, Oliver Heid, Edgar Mueller, and Lothar Rudi Schad. 2000. Prospective acquisition correction for head motion with image-based tracking for real-time fmri. *Magnetic Resonance in Medicine*, 44.

Chris Turner. 2020. Towards a new pedagogical approach to some and any based on large-scale corpus analysis.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Yawei Wang, Giacomo Bacco, and Nicola Bianchi. 2017. Geometry analysis and optimization of pm-assisted reluctance motors. *IEEE Transactions on Industry Applications*, 53(5):4338–4347.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14024–14031.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leila Wehbe, Idan Asher Blank, Cory Shain, Richard Futrell, Roger Levy, Titus von der Malsburg, Nathaniel Smith, Edward Gibson, and Evelina Fedorenko. 2021. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*, 31(9):4006–4023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Will Xiao and Gabriel Kreiman. 2020. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS Computational Biology*, 16.

Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. 2019. Topology attack and defense for graph neural networks: An optimization perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3961–3967. International Joint Conferences on Artificial Intelligence Organization.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. 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)*, pages 306–316, Online. Association for Computational Linguistics.

Jin Yong Yoo, John Morris, Eli Lifland, and Yanjun Qi. 2020. Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 323–332, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Algorithm

Algorithm 1 GOLI: A gradient-based sentence transformation method

1: **Input:** Random $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, model $\mathcal{M} = \boldsymbol{\theta}_{\text{map}} \circ \boldsymbol{\theta}_{\text{LLM}}$; Learning rate α; Loss function ℓ; Perturbation iterations N; Number of sites to perturb k
2: ▷ **Site selection**
3: $\mathcal{T}$ = OrderByImportance($\{\boldsymbol{x}_i\}_{i=1}^n$)
4: ▷ From Wallace et al. (2019); see Section 3
5:
6: ▷ **Site perturbation**
7: $\boldsymbol{u} = \boldsymbol{x}$
8: **for** j in N **do**
9: **for** $\boldsymbol{x}_i$ in $\mathcal{T}$ **do**
10: **if** $k > 0$ **then**
11: $\boldsymbol{u}_i^{\text{soft}}$ = Softmax($\boldsymbol{u}_i$)
12: $\boldsymbol{u}_i$= Multinomial($\boldsymbol{u}_i^{\text{soft}}$)
13: $k = k - 1$
14: **end if**
15: **end for**
16: $y^{\text{pred}} = \mathcal{M}(\boldsymbol{u})$ ▷ Forward pass
17: $\nabla = \frac{\partial}{\partial \boldsymbol{u}} \ell(y^{\text{pred}})$ ▷ Backward pass
18: $\boldsymbol{u} = \boldsymbol{u} - \alpha.\nabla$
19: **end for**
20: $\mathcal{S}^{\text{gen}} = \boldsymbol{u}$
21: **return** $\mathcal{S}^{\text{gen}}$

Summary. A backward pass (line 17) allows gradients with respect to the input $\boldsymbol{u}$ to be propagated from the loss function ℓ. The input is then modified in the direction of these gradients (line 18), with the modified input passed to $\mathcal{M}$ (line 11-16).

To solve Eq (4) effectively, we relax $\mathbf{u}_i \in \{0,1\}^{|V|}$ to $\mathbf{u}_i \in [0,1]^{|V|}$. This continuous relaxation of binary variables is a commonly used trick in combinatorial optimization to boost the stability of learning procedures in practice (Boyd et al., 2004).

See Jang et al. (2017); Maddison et al. (2017) for details on how the softmax (line 11, Algorithm 1) aids in reparametrization of the `argmax` functionality in the categorical case. This is theoretically equivalent to the Gumbel softmax trick.

Once the continuous optimization problem Eq (4) is solved, a hard thresholding operation or a randomized sampling method can be used to map

a continuous solution to its discrete domain. For the randomized sampling method, we consider $\mathbf{u}$ as probability vectors with elements drawn from a Multinomial distribution. A Multinomial distribution models selecting one of the $|V|$ classes when selecting a token from the vocabulary. We use the randomized sampling method in our experiments and follow the setup described in Algorithm 1 in Xu et al. (2019). See also Xu et al. (2019) for a proof of convergence of the randomized sampling method.

When incorporating additional constraints, such as capitalizing the first word or ensuring the last word is a punctuation, we sample from a subset of u_i indices. Originally, $|u_i| = |V|$. To sample from a subset of the vocabulary, say capitalized letters, we identify the set of indices C in the vocabulary corresponding to capitalized letters. When sampling from u_i, we mask out all those indices not in C, and sample only from those present in C.

B Minimal-pair counterfactuals task

We reuse all models and hyperparameters described in the codebase release by (Howard et al., 2022). Link: https://github.com/IntelLabs/NeuroCounterfactuals

We do not run NeuroCF from scratch since the CFs corresponding to their trainsets are publicly available on their repository: https://github.com/IntelLabs/NeuroCounterfactuals/blob/main/output/NeuroCFs-1g/counterfactuals.pkl

Instead we train their base RoBERTa model by running `evaluate_counterfactuals.py` and augmenting CFs generated by GOLI on their training set.

To calculate MoverScore distances, we use the implementation by Zhao et al. (2019), by installing their codebase from source (commit ID 9c362cc5aea61270e988ea0870bf5ae495cc80a3): https://github.com/AIPHES/emnlp19-moverscore The version on PyPI is outdated.

We calculate Perplexity scores computed by GPT-J (Available at https://huggingface.co/docs/transformers/model_doc/gptj).

Howard et al. (2022) do not publish the original set of sentences they generate CFs for. We hence reuse the Moverscore and Perplexity scores for NeuroCF-1g and NeuroCF-np from Table 4 of Howard et al. (2022).

B.1 Sentence representations

We obtain sentence representations from the same fine-tuned BERT SST2 model that we use as $\boldsymbol{\theta}_{\text{map}}$.

BERT-uncased, the transformer model on which the SST model was fine-tuned, has 24 layers (*i.e.*, Transformer blocks) and an embedding dimension of 768. We obtained model representations by tokenizing each sentence using the model's standard tokenizer (BertTokenizerFast) and passing each sentence through the model. We retrieved model representations for each model layer (*i.e.*, at the end of each Transformer block). The fine-tuned model uses a sequence summary representation of each sentence using the special classification, [CLS], token which is prepended to the sequence and is standardly used as a token for classification output (Devlin et al., 2019). We used the pretrained model available via the HuggingFace library (Wolf et al. (2020), Transformers version 4.11.3) for our implementation.

B.2 GOLI settings.

We generate CFs using GOLI by using a BERT-based fine-tuned SST-2 classifier (available at https://huggingface.co/gchhablani/bert-base-cased-finetuned-sst2). We use the following settings for GOLI:

- Loss function: BCE
- Desired objective: The opposite of the class currently predicted.
- Learning rate: 0.01
- Number of perturbation iterations: 10
- Number of multinomial samples per iteration: 20
- Number of sites: Randomly sampled from $\{1, 2\}$
- Max. time taken per sentence: 30 minutes. We abort processing a sentence if it takes more than 30 minutes.
- $\boldsymbol{\theta}_{\text{map}}$: Fine-tuned classification head in SST2 BERT.
- $\boldsymbol{\theta}_{\text{LLM}}$: Fine-tuned SST2 BERT.

The runs took an average of 2 minutes per sample.

C fMRI experiment

We present details of the fMRI experiment in this section.

C.1 Dataset

We curated a large set (n=1500) of naturally occurring, diverse sentences (note that although not every stimulus constituted a sentence by some definitions, we use the term 'sentence' throughout for convenience) The sentences were obtained by randomly sampling from N=5 diverse text corpora spanning three main categories: 1) Published written text (The Wall Street Journal Corpus: Paul and Baker (1992), The Toronto Book Corpus (genres: fantasy, mystery, adventure): Zhu et al. (2015)), 2) Web media text (Common Crawl C4: Raffel et al. (2020)), and 3) Transcribed spoken text (The Cornell Movie-Dialogs Corpus: Danescu-Niculescu-Mizil and Lee (2011), the spoken section of The Corpus of Contemporary American English: Davies (2010)). Prior to sampling, the corpora were filtered to only include 6-word sentences with printable ASCII characters and had a letter as the first character. The corpora were preprocessed to remove repeated/leading/trailing whitespace, strip whitespace before common punctuation characters (?.!,:;), append a final period if the last sentence character was not a punctuation character, and uppercase the first letter. To obtain a diverse sentence set, 1000 sentences were randomly sampled from each of the three main text categories. These 3000 sentences were filtered according to the filtering criteria specified in Appendix C.7. To obtain a set of 1500 sentences, 500 sentences were randomly sampled from each main category to ensure diversity in final sentence set.

C.2 Participants and data acquisition

A total of 5 neurotypical adults (4 female), aged 21 to 30 (mean: 25; 3.16 std), participated for payment between October 2021 and April 2022. All participants completed two scanning sessions where each session consisted of 10 runs of a sentence reading experiment (sentences presented on the screen one at a time for 2s with an inter-stimulus interval of 4s, 50 sentences per run) along with additional tasks. The participants were exposed to the same set of 1000 6-word, corpus-extracted sentences (no repetitions), but in fully randomized order. All participants had normal or corrected-to-normal vision, and no history of neurological, developmental, or language impairments. All participants were right-handed, as determined by the Edinburgh handedness inventory (Oldfield, 1971). All participants were native speakers of English. All participants gave informed written consent in accordance with the requirements of an institutional review board (to be made public after the anonymous review phase).

Structural and functional data were collected on the whole-body, 3 Tesla, Siemens Trio scanner with a 12-channel (G1; N=18) or 32-channel (G2; N=788) head coil. T1-weighted structural images were collected in 176 sagittal slices with 1 mm isotropic voxels (TR = 2530 ms, TE = 3.48 ms). Functional, blood oxygenation level dependent (BOLD), data were acquired using an EPI sequence (with a 90 degree flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: 33 (G1) or 31 (G2) 4 mm thick near-axial slices acquired in the interleaved order (with 10% distance factor), 3.0 mm × 3.0 mm (G1) or 2.1 mm × 2.1 mm (G2) in-plane resolution, FoV in the phase encoding (A » P) direction 192 mm (G1) or 200 mm (G2) and matrix size 64 mm × 64 mm (G1) or 96 mm × 96 mm (G2), TR = 2000 ms and TE = 30 ms. Prospective acquisition correction (Thesen et al., 2000) was used to adjust the positions of the gradients based on the participant's motion from the previous TR. The first 10s of each run were excluded to allow for steady state magnetization.

C.3 fMRI data preprocessing and first-level modeling

fMRI data were analyzed using SPM12 (release 7487). Each participant's functional and structural data were converted from DICOM to NIfTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session (Friston et al., 1995). Potential outlier scans were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using default thresholds in CONN preprocessing pipeline (5 standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm; (Nieto-Castanon, 2020)). Functional and structural data were independently normalized into a common space (the MontrealNeurological Institute [MNI] template; IXI549Space) using SPM12 unified segmentation and normalization procedure

(Ashburner and Friston, 2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between MNI-space coordinates (-90, -126, -72) and (90, 90, 108), using 2 mm isotropic voxels and 4th order spline interpolation for the functional data, and 1 mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were smoothed spatially using spatial convolution with a 4mm FWHM Gaussian kernel.

A General Linear Model (GLM) was used to estimate the beta weights representing the blood oxygenation level dependent (BOLD) response amplitude evoked by each sentence trial. Specifically, the data were modeled using the GLMsingle framework (Prince et al., 2022) using a fixed number of noise regressors (5) and a fixed ridge regression fraction (0.05). By default, GLMsingle returns beta weights in units of percent signal change by dividing by the mean signal intensity observed at each voxel in the brain and multiplying by 100. Hence, the beta weight for each voxel can be interpreted as a signal change for a given sentence trial relative to fixation baseline. To mitigate the effect of collecting data across two separate scan sessions, the betas from each session were z-scored (i.e., mean 0, std 1) for each voxel in each participant.

C.4 Language-selective brain regions

We were interested in brain responses from language-selective areas of the brain, and hence we identified the language network functionally in each participant using a well-validated fMRI language localizer task contrasting *sentences* with *nonwords* (defined as the top 10% language-selective voxels in a set of 5 pre-defined anatomical parcels: LIFGorb, LIFG, LMFG, LAntTemp, and LPostTemp) (Fedorenko et al., 2010). Voxels from these language-selective areas were z-scored and averaged for each participant, and finally averaged across partipants to yield a scalar response value for the average language network associated with each sentence trial.

C.5 Sentence representations

We obtained sentence representations from GPT2-XL. The other natural choice was bidirectional-attention Transformer model BERT-large-cased (Devlin et al., 2019). We prefer GPT2 over BERT because it is an auto-regressive model, pre-trained with the context of words that appear before it as opposed to the whole sentence. This is considered to be similar to human behavior. GPT-XL has 24 layers (*i.e.*, Transformer blocks) and an embedding dimension of 768. We obtained model representations by tokenizing each sentence using the model's standard tokenizer (GPTTokenizer) and passing each sentence through the model. We retrieved model representations for each model layer (*i.e.*, at the end of each Transformer block). We obtained a sequence summary representation of each sentence by selecting layer 21 (determined after cross-validation). We used the pretrained model available via the HuggingFace library (Wolf et al. (2020), Transformers version 4.11.3).

C.6 Training $\boldsymbol{\theta}_{\text{map}}$

We learned $\boldsymbol{\theta}_{\text{map}}$ using ridge regression (Scikit learn RidgeCV). We evaluated hyperparameter α in the range $[10^{-30}, 10^{29}]$. We validated $\boldsymbol{\theta}_{\text{map}}$ using 5-fold cross-validation. We obtained a Pearson R score per fold, and took the mean of those scores. The predicted values of $\boldsymbol{\theta}_{\text{map}}$ were correlated with the actual brain responses (5-fold cross validation) with a Pearson correlation of 0.38 (std across folds: 0.03).

C.7 Sentence filtering criteria

The filtering criteria were defined *prior to* setting up our experiments. In the following, the term "token" refers to groups of alphanumeric characters separated by whitespace.

C.7.1 Automatic filtering criteria

- More than 50% of tokens in the sentence contain numeric characters.
- More than 50% of all characters in the sentence are uppercase.
- The sentence contains one or more tokens that are longer than 20 characters.
- The sentence contains more than 4 consecutive punctuation characters.
- The sentence contains non-ascii characters (unicode index larger than 127) or a character in the following set: $*$, @ , [, \ ,] , / , $<$, $=$, $>$, ˆ , _ , ‘ , { , } , | , $\sim$

C.7.2 Manual filtering criteria

- The sentence is inappropriate/offensive (e.g., contains a racial slur or a taboo word).

- All tokens in the sentence are not English.
- The sentence contains some type of emoji face, e.g., ":)".
- All tokens in the sentence are trademarks, website names, brand names, product names, person names, journal identifiers, journal footers, date and geographical locations.
- The sentence contains more or fewer than 6 words due to tokenization/punctuation errors.

Referring to the example introduced in Section 3, had the site selection index been 1 instead of 3, thus selecting the token `Run` to be replaced, and had the optimal replaced token been `you` for this site, the replaced sentence would then be:

`youing slow suits me very well.`

This sentence makes little sense despite perhaps having a predicted neural response close to y^{desired}. While we do not impose any constraints on the generation algorithm itself to handle such cases, we post-process the generated sentences using the rule-based filters described above and discard sentences which fail the imposed filters. We do not filter any other sentences.

C.8 GOLI settings.

We use the following settings for GOLI:

- Loss function: Squared loss
- Desired objective: $+1.2$ for fMRI-high, -0.8 for fMRI-low.
- Learning rate: 0.01
- Number of perturbation iterations: 40
- Number of multinomial samples per iteration: 20
- Number of sites: Randomly sampled from $\{1, 2, 3, 4\}$
- Max. time taken per sentence: 30 minutes. We abort processing a sentence if it takes more than 30 minutes.
- $\boldsymbol{\theta}_{\text{map}}$: Trained, regularized linear regression
- $\boldsymbol{\theta}_{\text{LLM}}$: GPT2-XL

The runs took an average of 15 minutes per sample.

C.9 Examples of GOLI-generated stimuli

Appended in the following page spanning both columns.

Seed sentence	**fMRI-low**	**fMRI-high**
I don't find that particularly funny.	This don't find that particularly funny.	Pear don't find that particularly Cance.
Nice bike 24 fit for teenagers.	SourceFile bike 24 fit for teenagers.	Trump bike 24 rescuing for condem.
I seem to have been mistaken.	Certain hast to have been mistaken.	I ecosystems to have been IND.
But the shooting didn't really improve.	Our the shooting didn't really improve.	Edited the shooting didn't really IND.
This time two men got out.	Some time two men got alone.	Courtesy sentence two men got INCLUD.
But he may be suffering insomnia.	Sure he may be suffering pain.	But he Adapter be suffering SER.
I nodded and faked a smile.	Members nodded and faked a smile.	I nodded and superiors a avg.
That's been so important for me.	He's been so important for me.	ULAR's been so Huff for Ask.
We thought that might be interesting.	Miss thought that might be interesting.	Starting fre that happiness be DEC.
Need a case for my Epic!	Need a creature for my laughs!	ESPN incomp Greenberg for my Said!
Take a look at this, sir.	Come a look at this, brave.	BILL a look at this, pred.
Privacy should be provided for dressing.	She should be provided for dressing.	Privacy islands be Dig for pred.
Toomey is up by double digits.	Toomey is up by double tackling.	Toomey is up by double Fr.
The Blue pointed off to town..	The Blue pointed sped to town..	The Blue pointed rematch to Recommend..
Pay particular attention on her glasses.	Mother particular fingertips on her glasses.	Pay particular attention Links equate DEN.
And there should be more wood!	My there should be more wood!	And consumes savvy be more prev!
I think with intelligence and humor.	Wild think with intelligence and humor.	I Debor with Afghan and compr.
Thy will is now my will.	Thy will is now granted initiative.	Thy intercourse is now syll contr.
And their coffee can't be beat!	And their composure can't be beat!	And Administration conceptions can't be Deb!

Table 3: Examples of GOLI-generated sentences for the fMRI task