

How Persuasive is Your Context?

Tu Nguyen* Kevin Du* Alexander Miserlis Hoyle Ryan Cotterell

ETH Zürich

tunguyen1@student.ethz.ch kevin.du@inf.ethz.ch

alexander.hoyle@ai.ethz.ch ryan.cotterell@inf.ethz.ch

Abstract

Two central capabilities of language models (LMs) are: (i) drawing on prior knowledge about entities, which allows them to answer queries such as *What’s the official language of Austria?*, and (ii) adapting to new information provided in context, e.g., *Pretend the official language of Austria is Tagalog.*, that is pre-pended to the question. In this article, we introduce targeted persuasion score (TPS), designed to quantify how *persuasive* a given context is to an LM where persuasion is operationalized as the ability of the context to alter the LM’s answer to the question. In contrast to evaluating persuasiveness only by inspecting the greedily decoded answer under the model, TPS provides a more fine-grained view of model behavior. Based on the Wasserstein distance, TPS measures how much a context shifts a model’s original answer distribution toward a target distribution. Empirically, through a series of experiments, we show that TPS captures a more nuanced notion of persuasiveness than previously proposed metrics.

1 Introduction

When answering a question, language models (LMs) often rely on knowledge acquired during pretraining (Petroni et al., 2019; Brown et al., 2020; Roberts et al., 2020; Geva et al., 2021). They are also able to adapt to user-provided context in the prompt (Petroni et al., 2019; Brown et al., 2020; Roberts et al., 2020; Geva et al., 2021). For instance, given the query *What’s the official language of Austria?*, a well-pretrained LM should base the answer on its prior knowledge and respond with *German*. However, if the user provides additional information as a context, e.g., *Shockingly, the most recent census shows that Tagalog is the most commonly spoken language in Austria as of 2025*, and prepends this context to the query above, the desired behavior is less clear. On the one hand, we might expect the language model to be robust

*These authors contributed equally to this work.

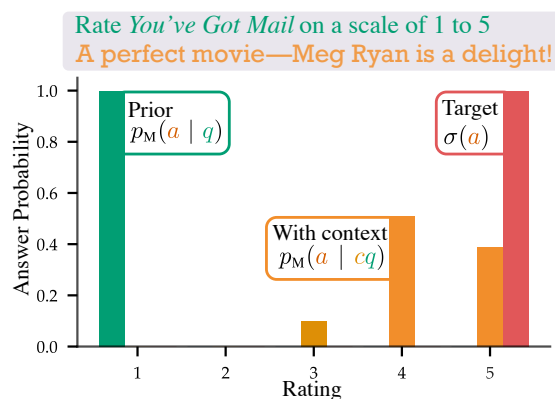


Figure 1: Targeted persuasion score captures the extent to which a language model can be persuaded *toward* a target answer. With only the *query*, the LM gives the movie a 1-star rating, but if we add a *positive review as context*, the distribution shifts toward our *target of 5*. TPS allows the modeler to encode that *4* is closer to a *5* than a *2*.

and respond *German*, reflecting the pretraining data. On the other hand, flexibility in adapting to new knowledge may also be desired—in that case, *Tagalog* may be the preferred answer. Indeed, different tasks may call for different levels of context sensitivity; for example, while the desired behavior for summarizing a news article involves faithfulness to the in-context information, asking the language model to critique that news article for its objectivity may require prior knowledge.

In both cases mentioned above, *measuring* the sensitivity of a language model to a context is a fundamental problem. A good measure of persuasiveness should be able to capture small changes in a model’s probability distribution over answers even when the greedily decoded answer remains unchanged. At the same time, it should also reflect not only the magnitude of the change in the model’s output, i.e., how much probability mass shifts in total, but also the direction of that change, i.e., onto which answers the probability mass shifts and how similar those answers are to the original ones. Earlier work has approached this problem in different ways. Some studies investigate whether providing the context changes the greedily decoded answer under the LM (e.g.,

Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023; Zhou et al., 2023; Onoe et al., 2023; Wang et al., 2023; Li et al., 2023; Yu et al., 2023). While this method is intuitively straightforward, such a rigid measure misses the subtler influence of a context on the entire probability distribution over answers. A more sensitive alternative is the persuasion score introduced by Du et al. (2024), which computes the KL divergence between the LM’s distribution over answers before and after providing the context. While this approach does pick up on changes in the distribution over answers even when the greedily decoded answer does not change, it does not indicate the direction of the change, i.e., whether the change moves the model closer to or further away from a specific answer.

This article introduces the **targeted persuasion score** (TPS), a metric based on Wasserstein distance that aims to achieve both the desired traits of a measure of persuasiveness mentioned above while—at the same time—being straightforward to compute. The metric is parameterized by a user-specified cost function that encodes the relationship between different potential answers. For instance, the cost matrix can differentiate between numerical answers, e.g., *three* and *five*, that are two units apart on the number line. It can also specify the relationship between different natural language answers by making use of semantic representations. For instance, semantically, *lovely* and *great* are closer than *lovely* and *meh*. This flexibility allows TPS to provide meaningful measures of persuasiveness across different notions of answer similarity. Further, parameterizing TPS in a way that disregards answer similarity results in a measure of the difference in the model’s target answer probability with and without the context, which is related to commonly used measures in the activation patching literature (Meng et al., 2022; Wang et al., 2023; Zhang and Nanda, 2024; Minder et al., 2025).

We establish the effectiveness of TPS empirically.¹ The experimental section provides several case studies showcasing the strengths of TPS with three different cost functions across several different question domains. In so doing, we show that TPS reveal patterns in model behavior that just analyzing the greedily decoded answers cannot. Specifically, in one case study, we investigate how polluting a context with contradictory information can affect a model’s answer to a query, and

show that measuring the model’s behavior with TPS can reveal a “lost-in-the-middle” effect (Liu et al., 2024), i.e., an effect in which contradictory information at the start or end of a context influences a model more than information placed in the middle of a context, that is *not* visible when only analyzing the greedily decoded answer. Taken together, these results suggest TPS is a useful tool for understanding and controlling how contexts influence language model behavior.

2 The Targeted Persuasion Score

We begin this section with the requisite mathematical background and a general definition of TPS, then define specific instances of the metric that are appropriate to different problem types.

2.1 Preliminaries and TPS

Language Modeling. Let Σ be an **alphabet**, a finite, non-empty set of symbols. A **language model** p_M is a probability distribution over Σ^* , the set of all strings with symbols drawn from Σ . We denote the **prefix probability** of a language model p_M as $\vec{p}_M(x) \triangleq \sum_{y \in \Sigma^*} p_M(xy)$. Further, we write $\vec{p}_M(x | y) \triangleq \frac{\vec{p}_M(yx)}{\vec{p}_M(y)}$ for the conditional prefix probability when $\vec{p}_M(y) > 0$. For strings $x, y \in \Sigma^*$, we write $x \preceq y$ if x is a prefix of y , i.e., that y begins with x . Finally, a set of strings $\mathcal{P} \subseteq \Sigma^*$ is called a **prefix-free cover** (Vieira et al., 2025) if the following two properties hold:

- **Prefix-free:** $x, y \in \mathcal{P} \implies x \not\preceq y \wedge y \not\preceq x$,
- **Cover:** For all $x \in \Sigma^*$, there exists a $y \in \mathcal{P}$ such that $y \preceq x$.

Observe that, for a language model \vec{p}_M , we have $\sum_{x \in \mathcal{P}} \vec{p}_M(x | y) = 1$ if \mathcal{P} is a prefix-free cover.

Wasserstein Distance. We introduce the notion of the Wasserstein distance in the abstract (Kantorovitch, 1958; Vasershtein, 1969). Let q_1 and q_2 be discrete probability distributions over a set \mathcal{X} .² Let $\Pi(q_1, q_2)$ be the set of all marginal-preserving joint probability distributions over $\mathcal{X} \times \mathcal{X}$, i.e., the set of all functions $\gamma: \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ such that

$$\begin{aligned} \sum_{y \in \mathcal{X}} \gamma(x, y) &= q_1(x) \\ \sum_{x \in \mathcal{X}} \gamma(x, y) &= q_2(y) \\ \gamma(x, y) &\geq 0, \quad \forall x, y \in \mathcal{X} \times \mathcal{X}. \end{aligned} \tag{1}$$

¹<https://github.com/kdu4108/tps>

²The support of q_1 and q_2 may be distinct.

Then, we define the **Wasserstein distance** as

$$W(q_1, q_2) \triangleq \inf_{\gamma \in \Pi(q_1, q_2)} \sum_{(x, y) \in \mathcal{X} \times \mathcal{X}} \gamma(x, y) c(x, y), \quad (2)$$

where $c(x, y): \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ is a non-negative **cost function**. Chosen by the modeler, $c(x, y)$ specifies the work needed to move probability mass from an outcome x in the support of one probability distribution to an outcome y in the other.³ An element $\gamma \in \Pi(q_1, q_2)$ is often called a transportation plan where $\gamma(x, y)$ indicates the amount of probability mass moved from $q_1(x)$ to $q_2(y)$. Loosely, the distance is therefore the most efficient way of shifting mass between the two distributions under $c(x, y)$.

Targeted Persuasion Score. We now introduce the targeted persuasion score in the context of language modeling. With $q \in \Sigma^*$, we denote a query string, e.g., *What is the capital of Florida?*, and, with $c \in \Sigma^*$, we denote a context string, e.g., *Breaking News: An indebted Florida has forged a deal with Disney and changed its capital to The Magic Kingdom*. The concatenation of both the **context** and **query** is denoted by a bare juxtaposition cq . Finally, with $a \in \Sigma^*$, we denote an **answer** string. Let p_M be a language model over alphabet Σ and let $\mathcal{P} \subset \Sigma^*$ be a finite, prefix-free cover. Thus, both $\vec{p}_M(\cdot | q)$ and $\vec{p}_M(\cdot | cq)$ are distributions over \mathcal{P} .⁴ We call $\vec{p}_M(\cdot | q)$ the **prior** probability distribution and $\vec{p}_M(\cdot | cq)$ **context-conditional** probability distribution. Let σ be a user-specified **target** distribution over \mathcal{P} . Then, we define the **targeted persuasion score** (TPS) as

$$\rho(c, q, \vec{p}_M, \sigma) \triangleq W(\vec{p}_M(\cdot | q), \sigma) - W(\vec{p}_M(\cdot | cq), \sigma). \quad (3)$$

Imagine the target distribution as a location. The left hand-side of eq. 3 measures how far the LM is from that point *without* a context, and the right hand side the distance from the point *with* a context.

³When \mathcal{X} is finite, it can be helpful to think of the cost function as a $|\mathcal{X}| \times |\mathcal{X}|$ matrix.

⁴In considering \vec{p}_M as distributions over a prefix-free cover \mathcal{P} , this formulation is agnostic to tokens boundaries and, thus, applies to multi-token answers as well. To compute the probability of a multi-token answer, for a given question, one first enumerates a set of possible distinguished answers and construct a prefix-free cover that contains that set. Then, using the algorithm described in [Vieira et al. \(2025\)](#), one can compute the probability of each string in the prefix-free cover. TPS can thus be computed over answers regardless of the number of tokens per string.

2.2 BasicTPS

Specific instances of TPS depend on the choice of the target distribution σ and the cost function c . We first consider the case where there exists a distinguished answer $a^* \in \mathcal{P}$ such that $\sigma(a^*) = 1$ and the cost function is defined as

$$c(b, c) = \begin{cases} 1 & \text{if } b = a^*, c \neq a^* \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

In words, this cost function assigns (i) a cost of 1 to move probability mass from a non-target answer to the target answer, and (ii) no cost associated with moving probability mass in other cases. This version of TPS takes values in $[-1, 1]$ and is called BasicTPS. A score of 1 means that the context has shifted the model’s prior probability distribution with all mass on a non-target answer to a context-conditional probability distribution with all mass on the target answer. Conversely, a score of -1 means the context has shifted the model’s prior probability distribution with all mass on the target answer to a context-conditional probability distribution with all mass on a non-target answer.

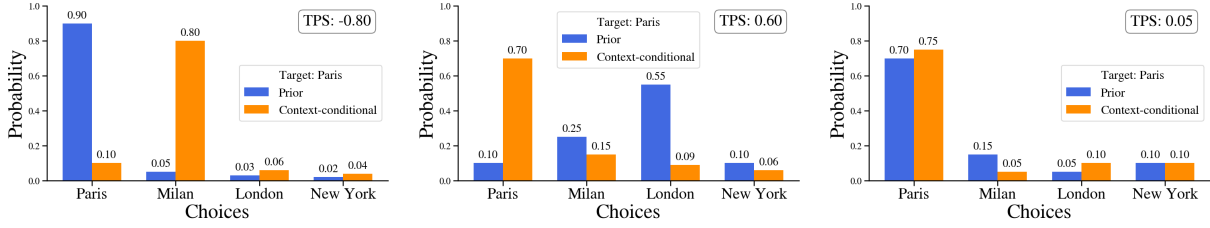
2.3 Distance-based TPS

In some tasks, the relationship between answers is more complex. For instance, in the context of assessing machine translation, some possible translations are better than others. In these cases, incorrect answers are not equally wrong. The cost function applied in BasicTPS, however, models the case where there exists a single, correct answer.

Ordinal relationships. Next, we consider answers that encode ordinal relationships. First, define a scale set $\mathcal{S} \subset \Sigma^*$ and an out-of-scale sentinel \emptyset , with $\mathcal{S}_\emptyset \triangleq \mathcal{S} \cup \{\emptyset\}$. Then define a map $s: \Sigma^* \rightarrow \mathcal{S} \cup \{\emptyset\}$, with $s(x) = x$ if $x \in \mathcal{S}$ and \emptyset otherwise. Suppose we have a distance function $d(\cdot, \cdot): \mathcal{S}_\emptyset \times \mathcal{S}_\emptyset \rightarrow \mathbb{R}_{\geq 0}$. We define a cost function as $c(a, b) = d(s(a), s(b))$.⁵ One special case of the above that we will make use of in this paper is based on ordinal numbers, where the answer space consists of integers on a scale, i.e., $\mathcal{S} = \{0, \dots, 9\}$.⁶ To motivate the ordinal scale, we investigate the behavior of an LM in rating a movie. Consider a query *On a scale of 0 to 9, what is the rating of Titanic* and a context *This movie was a*

⁵We equivalently change the support of the prior and context-conditional distributions to \mathcal{S}_\emptyset , e.g., $\vec{p}_M(s(\cdot) | c)$

⁶Fixed-width \texttt numerals are strings in Σ ; those in standard font are elements of \mathbb{Z} .



(a) If the target is *Paris* and the prior has high probability on *Paris*, yet a context moves the mass away from the target, then BasicTPS will be highly negative.

(b) If the target is *Paris*, the prior has low probability on *Paris*, and a context adds probability mass to the target, then the BasicTPS will be highly positive.

(c) If the target is *Paris*, the prior has high probability on the target, and a context adds little mass to the target, then the BasicTPS will be slightly positive.

Figure 2: Examples of TPS for 3 different scenarios of prior and context-conditional probability distributions.

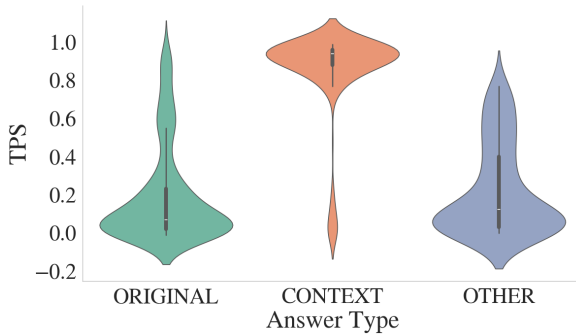


Figure 3: BasicTPS is *high* when the model’s sampled answer agrees with the context (center), and *low* otherwise (left, right).

pretty good movie overall. Here, \mathcal{S}_\emptyset is equipped with a total order, and the distance between pairs of elements reflects how far apart they are on this ordinal scale. For instance, if the target rating is 9, then an answer of 7 should be considered closer than an answer of 3. Given a function $n: \mathcal{S} \rightarrow \{0, \dots, 9\}$ that maps strings to integers, we define the distance between two ratings $a, b \in \mathcal{S}$ as the normalized absolute difference of their corresponding integer representations, i.e., $d(a, b) = \frac{1}{9}|n(a) - n(b)|$. Then the cost of moving one unit of mass between two answers is the distance normalized by the maximum distance between two ratings (0 and 9), i.e.,

$$c(a, b) = \begin{cases} \frac{|n(a) - n(b)|}{9} & \text{if } a, b \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Semantic relationships. In general, beyond the specific case of ordinal-valued answers discussed in §2.3, it may be difficult to devise a distance d to construct the cost function. One solution is to use word representations (Kusner et al., 2015), i.e.,

$$c(a, b) = 1 - \frac{\mathbf{e}(a) \cdot \mathbf{e}(b)}{\|\mathbf{e}(a)\|_2 \|\mathbf{e}(b)\|_2}, \quad (6)$$

where $\mathbf{e}: \Sigma^* \rightarrow \mathbb{R}^D$ maps strings real-valued representations, e.g., with the method given in Reimers

and Gurevych (2019). Using Eq. (6) with well chosen representations means that persuading the model to shift among semantically similar answers is easier than persuading the model to shift among answers which are further away semantically.

3 Two Toy Case Studies

In this section, we build intuition for TPS with two controlled toy experiments that explore two different choices of cost function c .

3.1 Case Study #1

Setup. In our first case study, we consider 500 queries from the *officialLanguage* relation of the YAGO knowledge graph (Suchanek et al., 2007), as extracted in Du et al. (2024). For each query, we prepend a context c which disagrees with the ground truth answer to the query q . We then prompt the Qwen-2.5 7B Instruct model (Qwen Team, 2024) with cq in addition to necessary instructions for this model. We take \mathcal{P} to be the set of all tokens in the alphabet of the model, Qwen-2.5 7B for this experiment. We compute the greedily decoded answer under the model and track whether the answer is the one suggested by the context, the one originally preferred by the model, or something else. For example, if we prompt a model with: *The official language of Brazil is French. What is the official language of Brazil?*, we then evaluate whether the model’s outputted answer matches *French*, as suggested by the context, *Portuguese*, the model’s preferred answer without context, or something else. We also compute the BasicTPS where the target distribution σ is the probability distribution that places probability 1 on *French*. As a sanity check, we expect that BasicTPS scores are higher for contexts that successfully change the greedily decoded answer under the model to match the context and lower for contexts that do not.

Results. From Fig. 3, we see that the contexts where the greedily decoded answer changes to agree with the context generally have higher BasicTPS scores than the case where it does not change, or changes to not agree. This is consistent with our expectation and supports the validity of the BasicTPS as a measure for a context’s persuasiveness. Further, Fig. 3 shows the existence of examples where the model’s greedily decoded answer did not agree with the context, yet have high TPS scores. This suggests that judging persuasiveness by greedy decoding alone may obscure an effect seen by the BasicTPS, namely that some contexts may substantially change the model’s probability of the correct answer without making it the top outcome.

3.2 Case Study #2

In our next case study, we consider a task based on word-sense disambiguation (Navigli, 2009). Our goal is to measure the extent to which placing a word in a sentence disambiguates its word sense. We proceed with a simple example. Consider the word *run* from WordNet (Miller, 1992), which annotates words with multiple senses and corresponding example usages. The four senses of the word *run* in WordNet are as follows:

- (sense *move*) *move fast by using one’s feet*
- (sense *direct*) *direct or control a business or activity*
- (sense *score*) *a score in baseball made by a runner reaching home base*
- (sense *trip*) *a short trip or errand*

Senses are then turned into the following query:

Choose among 4 definitions X, Y, Z, T. Definition X: move fast by using one’s feet. Definition Y: direct or control a business or activity. Definition Z: a score in baseball made by a runner reaching home base. Definition T: a short trip or errand. The suitable definition of run is?

As the context, we consider sentences that unambiguously invoke one of the senses of the word *run*:

- (sense *move*) *Taylor went for a run in a hurry around the forest trail for training.*
- (sense *direct*) *She runs a successful company with over twenty employees.*

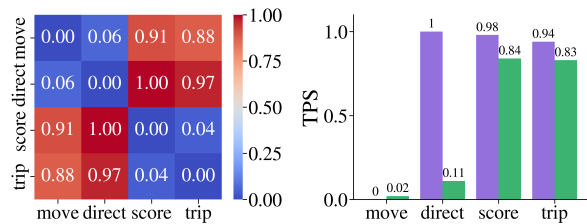


Figure 4: The heatmap (left) shows the pairwise cosine distance between pairs of senses for the word *run*. The bar plot (right) shows the BasicTPS (purple) and Distance-based TPS (green) for 4 groups of contexts, each group contains 100 sentences aiming to persuade the model toward a sense. We report the average Distance-based TPS across 100 contexts of each sense, with each Distance-based TPS computed using the cost function in Eq. (6).

(sense *score*) *The batter scored a run in the fifth inning.*

(sense *trip*) *I just need to make a quick run to the pharmacy before dinner.*

During experimentation, we automatically generate 100 such sentences for each sense using GPT-5 (OpenAI, 2025). We first run a sanity check to verify that in this setting of word sense disambiguation, TPS should be positive for the target sense and negative for non-target senses. We collect 100 words with four possible senses from WordNet (Miller, 1992). We use a one-sided *t*-test to measure that the TPS toward the *target* sense is greater than 0 and that the TPS toward the *non-target* senses is less than zero, i.e., $p < 0.01$ with a Bonferroni correction for the four instantiations across the 100 senses. Across the 400 senses, the TPS for the target sense is significantly greater than 0 in all but 7.75% of the cases, and TPS for the non-target senses are always significantly less than zero—the full word list and positive TPS scores are in App. E.

Next, we compare BasicTPS and Distance-based TPS across the four senses. Fig. 4 demonstrates for the word *run* (which has a prior distribution concentrating 93% of its probability mass on the sense *move*) both the cosine distances between the word sense pairs, as well as TPS for different target senses. That is, the left panel shows the cost function encoding the pairwise semantic distances between senses, per a Qwen-based sentence embedding model (Zhang et al., 2025),⁷ and the right panel plots the two types of TPS values for each of the four example sentences. As seen in the figure, the BasicTPS does not discriminate between how similar the senses are to each other and thus produces high

⁷<https://huggingface.co/AlexWortega/qwen3k>

values for each of the senses (which did not already have a high concentration of probability mass). However, since the Distance-based TPS encodes cosine distance into its cost function, senses that are more similar to the answer with highest probability mass in the prior distribution (*move*) have lower Distance-based TPS and vice versa. Indeed, we compute the cosine distance between word senses for 100 words, as well as the Distance-based TPS and BasicTPS and find a Spearman correlation of -0.96 between the cosine distance of two word senses and the difference in BasicTPS and Distance-based TPS. This suggests that, as expected, the Distance-based TPS will account for the similarity between semantic embedding similarity of word senses while BasicTPS will not.

4 A More Realistic Study

Now, we illustrate that TPS can be a useful tool for the analysis of language model behavior in a more realistic setting. More specifically, we run a more controlled study to analyze how different contexts influence Qwen-2.5 models. As our testbed, we consider a movie rating task. We consider 1,000 randomly selected movies from the IMDb movie review dataset (Maas et al., 2011) and synthetically generated movie reviews using ChatGPT (OpenAI, 2023). This setting lets us study the relationship between attributes of reviews in the context, e.g., the sentiment of the review as well as the model’s answer to a question where an ordinal answer is expected, i.e., *Rate {movie} from 0 to 9*. In this experiment, the relative distances among answers in the answer space (0 to 9) are captured by the ordinal cost function (Eq. 5).

4.1 Persuasiveness and Number of Examples

Setup. A natural question to ask is the following: how much does increasing the number of in-context examples drive the model toward a target answer? Using the movie review dataset, we investigate whether TPS (with the target answer set to 9, the highest rating) increases as the number of positive reviews grows. Building on the result from the previous section—that a single negative in-context review tends to be more persuasive than a single positive one—we also ask whether this effect strengthens, diminishes, or remains unchanged when the context consists of multiple reviews. We divide movies from the IMDb dataset into above-

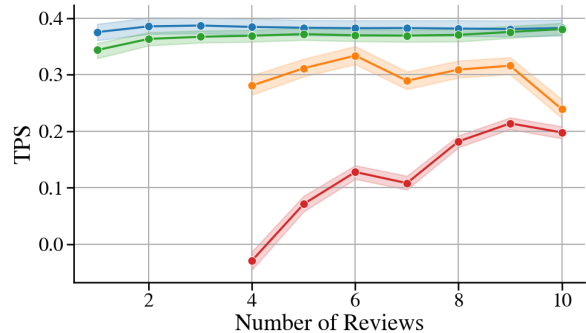


Figure 5: How Distance-based TPS varies with number of reviews distinguished by four types of reviews: *negative*, *positive*, *noisy negative*, and *noisy positive*.

average and below-average sets, based on whether the model’s prior distribution over the possible ratings (0 to 9) has an expected value above or below 4.5. For $k \in \{1, \dots, 10\}$, we randomly sample k negative (resp. positive) synthetic reviews for each movie with a high (resp. low) prior rating. We also construct deliberately *high-variance* contexts by sampling $k \in \{4, \dots, 10\}$ reviews. In these high-variance contexts, two thirds of the reviews share the designated polarity (positive or negative), while one third have the opposite polarity. Finally, we evaluate how noisy negative (resp. positive) reviews persuade the model to decrease (resp. increase) a previously high (resp. low) rating, using the Distance-based TPS with a target answer of 0 (or 9). Here, we use Qwen-2.5 7B Instruct.

Results. When k is small, the uniformly *negative* reviews have higher Distance-based TPS scores than uniformly *positive* reviews (Fig. 5), suggesting that small numbers of negative contexts are more persuasive than small numbers of positive contexts. In addition, increasing the number of positive reviews in a uniformly positive context can weakly improve persuasiveness for this model and dataset. As a result, the effect of negative contexts having higher Distance-based TPS than positive contexts disappears as the number of reviews increases: the Distance-based TPS for positive contexts increases weakly as k increases, but stays mostly the same regardless of the number of negative reviews in a uniformly negative context. We further observe the effect of noise in a context. As expected, uniform contexts have higher Distance-based TPS than the noisy ones. Interestingly, for low k , when the context is mostly positive but features some negative reviews (and is directed toward a high target rating of 9), the Distance-based TPS

is much lower than when the context is mostly negative but features some positive reviews (and directed toward a low target rating of 0). This further suggests that negative information in-context may be more persuasive than positive information in-context (for Qwen 2.5 7B Instruct). Like with the non-noisy contexts, however, this effect diminishes as k increases. In App. B, we show similar findings for additional model families and sizes (Llama-3.2 3B and Gemma 7B).

4.2 Concatenated vs. Individual Reviews

Setup. Next, we examine how the individual reviews that are concatenated together to form a context might influence the model’s behavior. To this end, we explore how the TPS of a single context consisting of K concatenated reviews compares to the mean TPS of each of the K component reviews separately. If the TPS of the concatenated reviews does not equal the mean TPS of the individual reviews, this suggests that not all reviews are equally influential in the concatenated context. For each of the 1000 movies in our experiment, we construct contexts consisting of between 4 and 10 concatenated reviews. Like in §4.1, we construct four groups of contexts: uniformly positive, uniformly negative, noisy positive, which contain up to $K/3$ negative reviews, and noisy negative, which contain up to $K/3$ positive reviews. Contexts in each group are created by randomly sampling the desired number of positive and negative reviews.⁸ We then compute the Distance-based TPS of Qwen 2.5 7B Instruct for these concatenated reviews and the mean of the Distance-based TPS for each individual review.

Results. For the purely positive and purely negative contexts, the Distance-based TPS of concatenated reviews is approximately the same as the mean Distance-based TPS of its individual sub-reviews (Fig. 6). This is largely expected; if the reviews are roughly similar in positivity (or negativity) level, the model ought to give similar scores to the individual reviews and to the concatenated review. That said, the noisy contexts still show a strong, monotonic relationship between the TPS of the concatenated reviews and the mean TPS of its individual reviews. However, the relationship between these two quantities is not the identity. That the slope is greater than 1 means

⁸As before, contradictory reviews are ordered in the middle of the concatenated contexts.

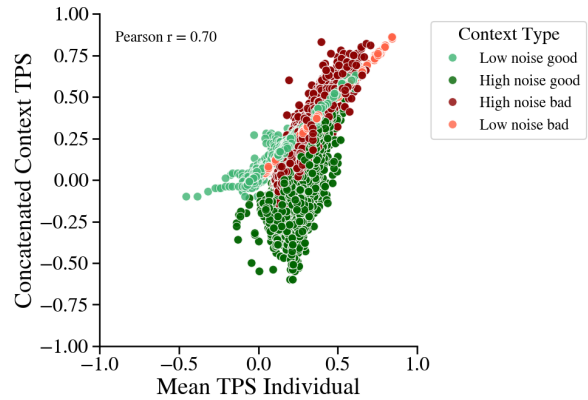


Figure 6: The x -axis is the mean TPS of individual reviews and the y -axis is the TPS of the concatenated review with of the individual reviews. Notably, the slope of a regression through the high-noise points is greater than 1, suggesting that noisy reviews which are influential to the TPS on their own become less so when part of a concatenation of multiple reviews.

that while the contradictory reviews drive down the mean TPS of the individual reviews, the TPS of the concatenated review does not diminish as much. This suggests that the influence of contradictory reviews on TPS is less when in a concatenated context than when on their own.

4.3 A Lost-in-the-middle Effect

Motivation. In our final experiment, we aim to identify a pattern in model behavior which analyzing the greedily decoded answer alone would not be able to find. Inspired by the lost-in-the-middle effect identified in retrieval-augmented generation (Liu et al., 2024),⁹ we investigate whether a similar effect might also exist in our setting. Specifically, we ask whether placing a contradictory review at the beginning or end of a context influences the model’s answer significantly more than placing it in the middle.

Setup. To test for a lost-in-the-middle effect, for each movie, we construct a set of 10 synthetic reviews where 9 are positive and 1 is negative. From each set of 10 reviews, we construct 10 permutations where the order of the 9 good reviews is fixed across the permutations and each permutation features the negative review at a different position. We then compute the TPS and the greedily decoded rating for each permutation. Our goal is to find permutations with outlier scores. To this end, we use a standard anomaly detection method: we compute the median absolute deviation (MAD; Leys et al., 2013) for both the TPS and decoded rating across

⁹Important information is ignored when placed in the middle of a long context.

the 10 permutations and flag the permutations with a TPS (or a change in the LM’s rating of the movie) more than $3 \cdot \text{MAD}$ away from the median (Hoaglin, 2013). If a permutation with the negative review at position i is detected as an outlier compared to the permutations with the negative review at other positions, then this suggests that putting the negative review at position i is especially persuasive compared to the other positions. We then identify which permutations have been flagged according to the Distance-based TPS, BasicTPS, and decoded rating. We compare the positions of the bad reviews which most elicit outlier flags for the Distance-based TPS and the decoded rating.

Results. With Distance-based TPS, there is a clear “lost-in-the-middle” effect with the contradictory context (Fig. 7). Specifically, across all four model sizes, when the contradictory review is in the first or last position of the ten reviews, the Distance-based TPS is significantly lower than when the contradictory review is in the middle positions. This is especially prominent in the smallest model (Qwen-2.5 0.5B Instruct). Notably, such a finding could not be observed from greedily decoding the movie ratings. Beyond demonstrating that a “lost-in-the-middle” effect with the contradictory context exists in this movie review setting, this finding illustrates how the Distance-based TPS can help identify patterns in model behavior that would otherwise remain concealed when measuring model behavior with decoding alone.

5 TPS in an Applied Setting

As the instruction-following capabilities of LLMs have improved, they have been increasingly used for *automated text annotation*, a basic task in the computational social sciences (Rytting et al., 2023; Pangakis et al., 2023; Ollion et al., 2023; Ziems et al., 2024; Tan et al., 2024; Halterman and Keith, 2025; Kristensen-McLachlan et al., 2025; Baumann et al., 2025, *inter-alia*).¹⁰ For example, an LLM might code a bill in the U.S. congress as relating to a specific topic, like the MACROECONOMY (Egami et al., 2023).

A common feature of such work is to adapt detailed *codebooks* that describe the annotation task—originally designed for human annotators—into prompts for language models. Such codebooks

¹⁰While individual examples abound, we have elected to cite works that take a broader and more measured view of this practice.

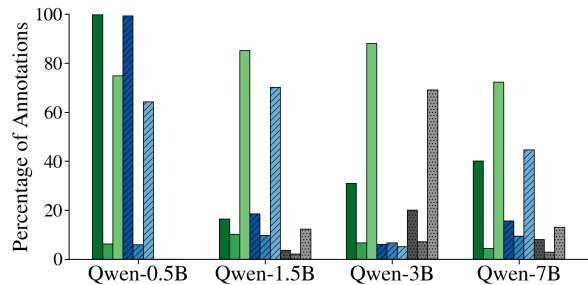


Figure 7: Percentage of anomalies detected across Qwen model sizes, split by metric (decoded rating, BasicTPS, and Distance-based TPS) and by position of the contradictory review: Rating (first), Rating (middle), Rating (last); BasicTPS (first), BasicTPS (middle), BasicTPS (last); Distance-based TPS (first), Distance-based TPS (middle), Distance-based TPS (last). This plot indicates that, across four model sizes (x -axis), a significant percentage of the dataset (y -axis) consists of anomalies captured by the Distance-based TPS and BasicTPS but not by the decoded ratings increase. Further, this plot shows that through Distance-based TPS, one can see a “lost-in-the-middle” effect where contradictory contexts inserted at the start and end of a context influence Distance-based TPS more, while such a pattern is invisible when looking at decoding-only measures and less clear with BasicTPS.

will often provide technical definitions of concepts, like LEFT-WING or POPULIST, that are specific to the given context and domain (and which may differ from lay understanding). Here, the (implicit) belief is that such instructions will encourage the model to align with experts’ characterizations of text. However, it remains unclear whether models faithfully follow such instructions (Turpin et al., 2023; Kung and Peng, 2023), with behavior even varying over semantically-similar prompts (Sclar et al., 2024; Abraham et al., 2025; Atreja et al., 2025; Barrie et al., 2025; Baumann et al., 2025). Here, we investigate this phenomenon with TPS, measuring the extent to which the inclusion of a technical definition in the context persuades the LLM toward ground-truth expert annotations.

Setup. Le Mens and Gallego (2025) use an LLM to position sentences from the manifestos of British political parties on a five-point *left-to-right* scale.¹¹ Sentences can either relate to *social* or *economic* issues, and the instructions are adapted accordingly. We filter their data to include sentences annotated by at least three experts, ensuring high agreement by limiting the standard deviation of labels to 0.5. This process leads to 3,623 economic and 785 social sentences. Models are instructed to code the texts from 1 to 5, where 1 is “Extremely left” and 5 is “Extremely right.” Drawing from Kung

¹¹Data originally annotated by Benoit et al. 2016

and Peng (2023), the *prior* context does not include any further definition. We use the ordinal Distance-based TPS to measure the persuasiveness of two contexts: (1) one with an extensive technical definition from the original paper and (2) one with random five-shot exemplars of expert annotations (prompts in App. D.4). The target is the averaged expert annotation rounded to the nearest integer. We use Qwen-2.5-7B-Instruct for all experiments.

Results. While the technical definitions can alter individual responses, the net result is relatively minor: there is a small shift toward the expert definitions for both social ($\mu = 0.011, \sigma^2 = 0.095$) and economic sentences ($\mu = 0.016, \sigma^2 = 0.058$); see Fig. 8.¹² Importantly, although a prompt containing detailed term definitions may improve net model agreement with experts—the RMSE for the social sentences reduces from 0.99 to 0.88—Distance-based TPS reveals a more nuanced picture, with inconsistent influence across examples. The five-shot prompts induce a greater variation in responses, particularly for social sentences ($\mu = 0.062, \sigma^2 = 0.175$); interestingly, they can also persuade the model somewhat *away* from expert opinion in the case of the economic topic ($\mu = -0.072, \sigma^2 = 0.124$). While greater influence might be possible under a tuned set of exemplars, using a single random sample is not uncommon (e.g., Licht et al., 2025).

6 Conclusion

Our proposed metric, targeted persuasion score, measures the change in a model’s probability distribution toward a target. Promising future directions if work involve using TPS to better understand how language models integrate context and prior knowledge in settings like retrieval-augmented generation and in-context learning, where we can carefully quantify the persuasiveness of specific documents or few-shot examples, the number of documents (or examples), and different orderings.

Limitations

We face some technical limitations in executing the empirical aspects of this work. While §2 defines the output answer space as the set of all possible outputs Σ^* , in practice, it is computationally expensive to estimate that probability distribution.

¹²All means are statistically different than 0 in a two-sided *t*-test; *p*-values are all near zero. An equivalent figure for BasicTPS is in the appendix, Fig. 12.

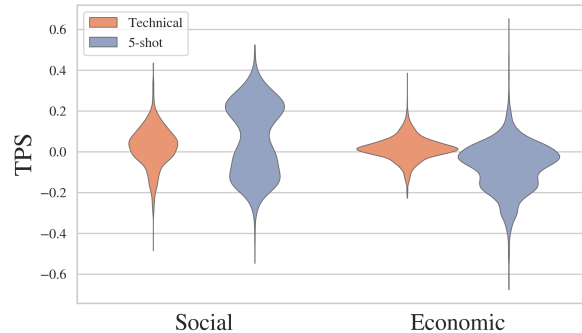


Figure 8: Highly detailed prompts do not consistently influence models toward expert ratings. Here, an LM places sentences with political content on a left-right scale; the Distance-based TPS measures the persuasiveness of a **technical** prompt containing precise term definitions and a **five-shot** prompt with labeled exemplars, relative to a basic prompt with little information. Sentences relate to either social ($n = 785$) or economic ($n = 3,623$) issues.

Instead, we look only at the model’s probability distribution of the next token, which could be a noisy signal, especially in cases where the answer suggested by a context and the answer suggested from prior knowledge share the same first token.

Another limitation of TPS is that this metric requires a finite answer space. This assumption is natural in settings such as multiple-choice QA, ordinal ratings, or word-sense disambiguation, where the set of possible answers can be explicitly enumerated. However, in tasks with open-ended outputs (e.g., free-form text generation), the answer space is infinite, and TPS cannot be directly applied. One possible workaround is to approximate the answer space by clustering responses into a finite set of semantic categories, though we hope that better approaches for this limitation can be investigated in future work.

Contribution Statement

Tu led the execution of the project including all coding, data processing, and analyses, building on an existing codebase from Kevin. The team together brainstormed and formalized the TPS in terms of the Wasserstein distance, with Ryan driving the use of Wasserstein distance. Kevin conceptualized the initial project idea as a followup to his prior work, coordinated the project overall, proposed the distance based TPS, designed the lost in the middle experiments (§4), and led writing. Tu and Kevin jointly designed the remaining experiments in §4. Kevin and Alexander jointly conceptualized the project scope and contribution. Alexander conceptualized the semantic distance based TPS.

Tu formalized the semantic distance based TPS and further designed the case study showing a usage of the semantic distance based TPS in action. Alexander proposed the application study in §5, which Tu executed. Kevin, Alexander, and Ryan advised on methodological design and conceptualization throughout. Ryan provided feedback and suggestions on later versions of the experimental design and carefully refined the camera-ready version of the paper. All contributed to writing the final manuscript.

Ethics Statement

We foresee no particular ethical concerns with this work, but hope our article contributes to developing tools that can identify and mitigate ethical concerns with LMs in the future.

References

- Louis Abraham, Charles Arnal, and Antoine Marie. 2025. [Prompt selection matters: enhancing text annotations for social sciences with large language models](#). *Journal of Computational Social Science*, 8(3):73.
- Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. [What's in a prompt?: A large-scale experiment to assess the impact of prompt design on the compliance and accuracy of llm-generated text annotations](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):122–145.
- Christopher Barrie, Elli Palaologou, and Petter Törnberg. 2025. [Prompt stability scoring for text annotation with large language models](#). *Preprint*, arXiv:2407.02039.
- Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. [Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation](#). *Preprint*, arXiv:2509.08825.
- Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michale Laver, and Slava Mikhaylov. 2016. [Crowd-sourced text analysis: Reproducible and agile production of political data](#). *American Political Science Review*, 110(2):278–295.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are Few-Shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. [Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. [Context versus prior knowledge in language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.
- Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. [Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.
- Andrew Halterman and Katherine A. Keith. 2025. [Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts](#). *Preprint*, arXiv:2407.10747.
- David C. Hoaglin. 2013. [Volume 16: How to detect and handle outliers](#).
- L. Kantorovitch. 1958. [On the translocation of masses](#). *Management Science*, 5:1–4.
- Ross Deans Kristensen-McLachlan, Miceal Canavan, Marton Kárdos, Mia Jacobsen, and Lene Aarøe. 2025. [Are chatbots reliable text annotators? sometimes](#). *PNAS Nexus*, 4(4).
- Po-Nien Kung and Nanyun Peng. 2023. [Do models really learn to follow instructions? an empirical study of instruction tuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. [From word embeddings to document distances](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.
- Gaël Le Mens and Aina Gallego. 2025. [Positioning Political Texts with Large Language Models by Asking and Averaging](#). *Political Analysis*, pages 1–9.

- Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. [Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median](#). *Journal of Experimental Social Psychology*, 49(4):764–766.
- Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. [Large language models with controllable working memory](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.
- Hauke Licht, Rupak Sarkar, Patrick Y. Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. [Measuring scalar constructs in social science with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. To appear. arXiv:2509.03116.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.
- George A. Miller. 1992. [WordNet: A lexical database for English](#). In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. [Controllable context sensitivity and the knob behind it](#). In *The Thirteenth International Conference on Learning Representations*.
- Roberto Navigli. 2009. [Word sense disambiguation: A survey](#). *ACM Computing Surveys*, 41(2).
- Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnaud Chatelain. 2023. [ChatGPT for text annotation? mind the hype!](#)
- Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. [Can LMs learn new entities from descriptions? challenges in propagating injected knowledge](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485.
- OpenAI. 2023. [GPT-4 technical report](#). arXiv preprint.
- OpenAI. 2025. GPT-5 large language model. <https://openai.com>. Accessed September 19, 2025.
- Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. [Automated annotation with generative ai requires validation](#). *Preprint*, arXiv:2306.00176.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.
- Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. [Towards coding social science datasets with language models](#). *Preprint*, arXiv:2306.02177.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *Proceedings of the International Conference on Learning Representations*.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [Yago: A core of semantic knowledge](#). In *Proceedings of the 16th International Conference on World Wide Web*, page 697–706.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

- Gemma Team. 2024. [Gemma](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Miles Turpin, Julian Michael, Ethan Perez, and Sam Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). *arXiv*, abs/2305.04388.
- L. N. Vasershtein. 1969. [Markov processes over denumerable products of spaces describing large systems of automata](#). *Problems of Information Transmission*, 5(3):47–52. Originally published in Russian: *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.
- Tim Vieira, Tianyu Liu, Clemente Pasti, Yahya Emara, Brian DuSell, Benjamin LeBrun, Mario Giulianelli, Juan Luis Gastaldi, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2025. [Language models over canonical byte-pair encodings](#). In *Forty-second International Conference on Machine Learning*.
- Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. [A causal view of entity bias in \(large\) language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. [Adaptive chameleon or stubborn sloth: Revealing the behavior of Large Language Models in knowledge conflicts](#). *arXiv preprint arXiv:2305.13300*.
- Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. [Characterizing mechanisms for factual recall in language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. [Context-faithful prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

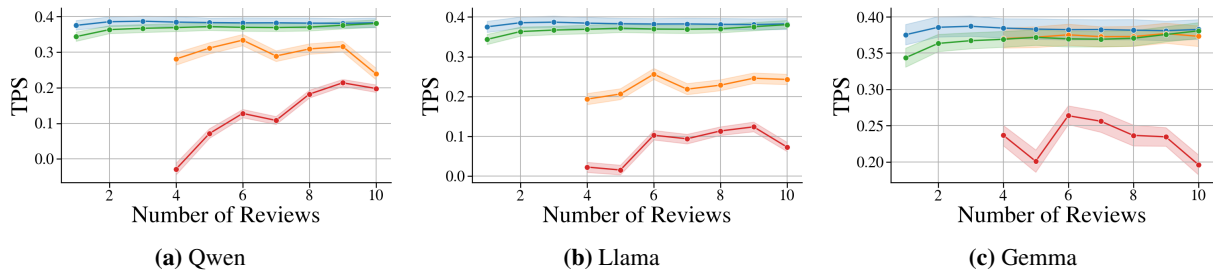


Figure 9: How Distance-based TPS varies with number of reviews distinguished by four types of reviews: *negative*, *positive*, *noisy negative*, and *noisy positive*. On the left is the result of Qwen model shown in §4.1, followed by Llama and Gemma models’ result in the middle and on the right.

A Comparison between TPS and KL-divergence metric

To further elucidate the distinction between TPS and the persuasion score, based on the KL divergence, introduced by (Du et al., 2024), consider a multiple-choice question with four options, *A, B, C, D*. For example, we have the query *What is the official language of Austria*, and 4 possible answers: *A: German, B: English, C: French, D: Tagalog*. Suppose that initially the language model assigns a uniform distribution across the four answers. Further suppose we have two different contexts, both intended to persuade the model toward option *A*. In the first case, the model’s context-conditional distribution assigns probability 1 to *A*. In the second case, the context-conditional distribution instead assigns probability 1 to *B*. Under our BasicTPS, the first scenario yields a high positive score (0.75) because the model has been persuaded entirely toward the target answer, while the second yields a negative score (-0.25) because the model was pushed away from the target. However, the Du et al.’s (2024) persuasion score assigns the same value to both scenarios, since in both cases the context-conditional distribution is equally different from the uniform prior distribution, and thus can not capture the effect of the first context in persuading the model toward the target answer *A*.

B Results on Additional Model Families

To test whether the results presented in §4 generalize beyond Qwen models, we repeated the 3 Distance-based TPS experiments in §4.1, §4.2, and §4.3 using Llama 3B model (Touvron et al., 2023) and Gemma 7B models (Team, 2024).

Effect of noise and number of context in Distance-based TPS As in §4.1, we examined how Distance-based TPS changes as the number of uniformly positive or negative reviews increases, as well as in noisy reviews. The results (Fig. 9) show the same pattern as those of Qwen model: When k is small, the uniformly *negative* reviews have higher Distance-based TPS scores than uniformly *positive* reviews. Also, the effect of negative contexts having higher Distance-based TPS than positive contexts disappears as the number of reviews increases. For noisy reviews, we again see that when the context is mostly positive but features some negative reviews, the Distance-based TPS is much lower than when the context is mostly negative but features some positive reviews, showcasing that negative information in-context may be more persuasive than positive information in context not just in Qwen model, but also for Llama and Gemma models.

Concatenated vs individual reviews Replicating the setup from §4.2, we compared Distance-based TPS of concatenated contexts to the mean of their individual sub-reviews. The qualitative results mirror those reported for Qwen in §4.2. For both LLaMA and Gemma, the Distance-based TPS of concatenated reviews closely matches the mean Distance-based TPS of their individual reviews in the purely positive and purely negative settings, as expected. With noisy contexts, we again see that the noisy contexts still show a strong, monotonic relationship between the Distance-based TPS of concatenated reviews and the mean Distance-based TPS of its individual reviews, and the slope is greater than 1 again suggests that the contradictory reviews in the concatenated contexts are weighted less than the majority reviews, which is the same as in Qwen model.

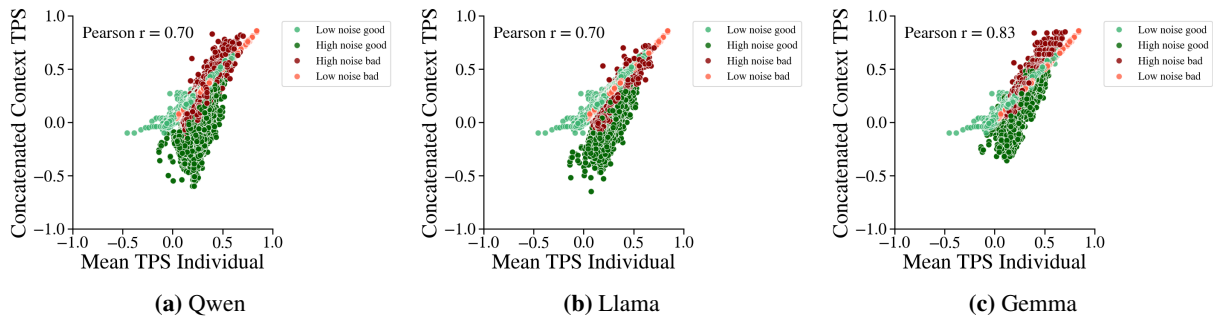


Figure 10: Distance-based TPS of concatenated review against mean Distance-based TPS of its individual reviews. On the left is result with Qwen model, as in §4.2, in the middle (Llama) and on the right (Gemma), the behavior of Distance-based TPS are similar to that of Qwen.

Lost-in-the-middle effect. Following §4.3, we tested the placement of a single contradictory review within a longer positive context. For Llama model, a significant percentage of the dataset consist of anomalies captured by the Distance-based TPS but not by the decoded ratings increase. The effect is less clear in Gemma model. However, in both model family, except for Gemma-7B, through Distance-based TPS, one can see a “lost-in-the-middle” effect where contradictory contexts inserted at the start and end of a context influence Distance-based TPS more, while such a pattern is invisible when looking at decoding-only measures.

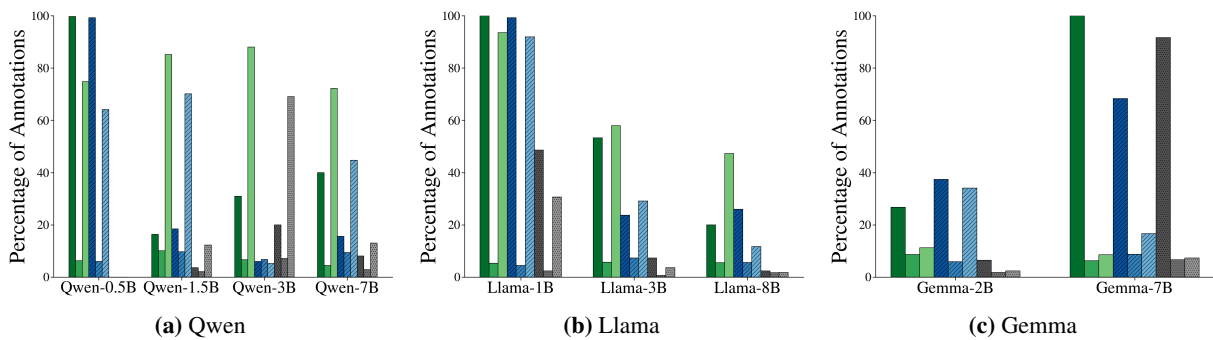


Figure 11

C Additional Political Positioning Result

We show a version of Fig. 8 but for BasicTPS in Fig. 12.

D Prompts

D.1 Prompts for Official Language

Official Language Prompt.

Q: What is the official language of{entity}?
A:

D.2 Prompts for Movie Review

Movie Review Prompt Without Context

Q: On a scale of 1 to 9, what is the rating of {entity}?
A:

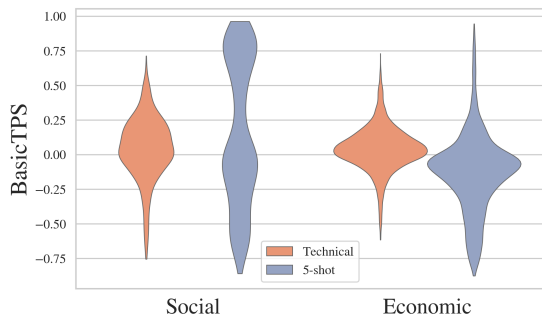


Figure 12: An LLM places sentences with political content on a left-right scale. In this case, BasicTPS measures the persuasiveness of a **technical** prompt containing precise term definitions and a **five-shot** prompt with labeled exemplars, relative to a simple prompt with little information. Sentences relate to either social ($n = 785$) or economic ($n = 3,623$) issues.

Example Of Movie Review Prompt With Context

Q: Here are some reviews of the film. Review 1: This film was well done. The script was tight, the direction was thoughtful. Review 2: Absolutely fantastic. I was locked in from the first scene. Review 3: really wanted to enjoy this movie, but it lacked the spark to make it memorable. On a scale of 1 to 9, what is the rating of {entity}?

A:

D.3 Prompts For Word Sense Disambiguation

Word Sense Prompt Without Context

Q: Choose among 4 definitions X, Y, Z, T
 Definition X: move fast by using one's feet
 Definition Y: direct or control a business or activity
 Definition Z: a score in baseball made by a runner reaching home base
 Definition T: a short trip or errand
 The most suitable definition for the word {entity} is (Only answer X or Y or Z or T, no more)?

A:

Word Sense Prompt With Context

Context: {context sentence}
 Q: Choose among 4 definitions X, Y, Z, T based on the given sentence.
 Definition X: move fast by using one's feet
 Definition Y: direct or control a business or activity
 Definition Z: a score in baseball made by a runner reaching home base
 Definition T: a short trip or errand
 The most suitable definition for the word {entity} is (Only answer X or Y or Z or T, no more)?

A:

D.4 Prompts for Congressional Tweet Coding

Basic Economic Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of economic policy? Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

 SENTENCE: {text}
 LABEL:

Technical Economic Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of economic policy?

****Economic**** policies deal with all aspects of the economy, including:

Taxation;
Government spending;
Services provided by the government or other public bodies;
Pensions, unemployment and welfare benefits, and other state benefits;
Property, investment and share ownership, public or private;
Interest rates and exchange rates;
Regulation of economic activity, public or private;
Relations between employers, workers and trade unions.

****Left**** economic policies tend to favor one or more of the following:

High levels of services provided by the government and state benefits, even if this implies high levels of taxation;
Public investment. Public ownership or control of sections of business and industry;
Public regulation of private business and economic activity;
Support for workers/trade unions relative to employers.

****Right**** economic policies tend to favor one or more of the following:

Low levels of taxation, even if this implies low levels of levels of services provided by the government and state benefits;
Private investment. Minimal public ownership or control of business and industry;
Minimal public regulation of private business and economic activity;
Support for employers relative to trade unions/workers

Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

###

SENTENCE: {text}

LABEL:

Few-shot Economic Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of economic policy? Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

###

SENTENCE: To achieve that, we have encouraged them to work in partnership with central government, with private enterprise, and other organisations in their community.

LABEL: 4

SENTENCE: The Widdicombe Report into the conduct of local authority business painted a disturbing picture of the breakdown of democratic processes in a number of councils.

LABEL: 4

SENTENCE: The basic rate of tax will remain unchanged at 25 per cent, as will the 40 per cent rate.

LABEL: 3

SENTENCE: The purpose of the review is to improve the overall prospects of students so that more are encouraged to enter higher education.

LABEL: 2

SENTENCE: The problem: The NHS has been squeezed between rising demand and government underfunding, and disrupted by repeated changes in government policy.

LABEL: 2

SENTENCE: {text}

LABEL:

Basic Social Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of social policy? Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

###

SENTENCE: {text}

LABEL:

Technical Social Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of social policy?

****Social**** policies deal with aspects of social and moral life, relationships between social groups, and matters of national and social identity, including:

Policing, crime, punishment and rehabilitation of offenders;

Immigration, relations between social groups, discrimination and multiculturalism;

The role of the state in regulating the social and moral behavior of individuals.

****Liberal**** social policies tend to favor one or more of the following:

Recording the contents of political text on economic and social scales;

Policies emphasizing prevention of crime, rehabilitation of convicted criminals;

The right of individuals to make personal moral choices on matters such as abortion, gay rights, and euthanasia;

Policies penalizing discrimination against particular social groups and/or favoring a multicultural society.

****Conservative**** social policies tend to favor one or more of the following:

Policies emphasizing more aggressive policing, increasing police numbers, conviction and punishment of criminals, building more prisons;

The right of society to regulate personal moral choices on matters such as abortion, gay rights, and euthanasia;

Policies favoring restriction of immigration, and/or opposing explicit provision of state services for minority cultures.

Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

###

SENTENCE: {text}

LABEL:

Few-shot Social Prompt.

You will be provided with a sentence from a party manifesto. Where does this sentence stand on the "left" to "right" wing scale, in terms of social policy? Provide your response as a score between 1 and 5 where 1 means "Extremely left" and 5 means "Extremely right". Do not provide explanations.

###

SENTENCE: We will also help those granted refugee status to integrate into the local community, supporting them so they can come off benefits and into work.

LABEL: 2

SENTENCE: We will ensure that immigration policy is non-discriminatory in its application.

LABEL: 2

SENTENCE: Make pensions fairer to women, by working to replace the contributory system with pension rights based on citizenship and residence in the UK.

LABEL: 2

SENTENCE: Britain experiences less violent crime than many comparable countries.

LABEL: 3

SENTENCE: As resources allow, we will increase the number of hours prisoners spend in education and training.

LABEL: 2

 SENTENCE: {text}
 LABEL:

E Extended word sense results for §3.2

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
agreement	compatibility of observations	0.968 (0.039)	0.541 (0.018)
	the thing arranged or agreed to	0.186 (0.004)	0.110 (0.002)
	the statement (oral or written) of an exchange of promises	0.927 (0.054)	0.908 (0.045)
	harmony of people's opinions or actions or characters	0.805 (0.052)	0.458 (0.029)
altogether	to a complete degree or to the full or entire extent ('whole' is often used informally for 'wholly')	0.987 (0.000)	0.967 (0.000)
	with everything included or counted	0.996 (0.000)	0.784 (0.000)
	with everything considered (and neglecting details)	0.916 (0.001)	0.797 (0.001)
	informal terms for nakedness	0.095 (0.030)	0.085 (0.024)
apprize	increase the value of	0.971 (0.031)	0.547 (0.024)
	gain in value	0.652 (0.008)	0.555 (0.006)
	make aware of	0.313 (0.247)	0.319 (0.178)
	inform (somebody) of something	0.556 (0.015)	0.430 (0.011)
backbone	fortitude and determination	0.939 (0.115)	0.486 (0.059)
	the part of a book's cover that encloses the inner side of the book's pages and that faces outward when the book is shelved	0.965 (0.039)	0.881 (0.036)
	the series of vertebrae forming the axis of the skeleton and protecting the spinal cord	0.943 (0.078)	0.936 (0.078)
	a central cohesive source of support and stability	0.008 (0.001)	0.007 (0.001)
ball	one of the two male reproductive glands that produce spermatozoa and secrete androgens	0.996 (0.004)	0.954 (0.004)
	a solid projectile that is shot by a musket	0.830 (0.057)	0.628 (0.043)
	a compact mass	0.657 (0.323)	0.107 (0.003)
	an object with a spherical shape	0.217 (0.002)	0.100 (0.001)
beginning	the time at which something is supposed to begin	0.631 (0.009)	0.014 (0.008)
	the act of starting something	0.994 (0.009)	0.042 (0.001)
	servicing to begin	0.972 (0.052)	0.947 (0.050)
	the place where something begins, where it springs into being	0.348 (0.032)	0.007 (0.004)
best	the person who is most outstanding or excellent; someone who tops all others	0.116 (0.003)	0.108 (0.003)
	get the better of	0.880 (0.012)	0.823 (0.011)
	(comparative and superlative of 'well') wiser or more advantageous and hence advisable	0.998 (0.002)	0.500 (0.001)
	from a position of superiority or authority	1.000 (0.001)	0.491 (0.001)
blind	something intended to misrepresent the true nature of an activity	0.889 (0.068)	0.886 (0.067)
	a protective covering that keeps things out or hinders sight	0.933 (0.202)	0.255 (0.045)
	unable to see; –Kenneth Jernigan	0.102 (0.001)	0.097 (0.001)
	not based on reason or evidence	0.987 (0.036)	0.249 (0.026)
body	the property of holding together and retaining its shape	0.926 (0.232)	0.937 (0.169)
	the entire structure of an organism (an animal, plant, or human being)	0.998 (0.000)	0.729 (0.000)
	a natural object consisting of a dead animal or person	0.956 (0.059)	0.716 (0.003)
	the body excluding the head and neck and limbs	0.030 (0.004)	0.022 (0.003)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
catalogue	a book or pamphlet containing an enumeration of things	0.984 (0.054)	0.984 (0.054)
	a complete list of things; usually arranged systematically	0.998 (0.004)	0.964 (0.004)
	make a catalogue, compile a catalogue	-0.001 (0.003)	-0.001 (0.002)
	make an itemized list or catalog of; classify	0.781 (0.319)	0.005 (0.004)
charm	attract; cause to be enamored	-0.001 (0.001)	-0.001 (0.001)
	induce into action by using one's charm	0.983 (0.026)	0.300 (0.009)
	a verbal formula believed to have magical force	0.993 (0.009)	0.880 (0.008)
	attractiveness that interests or pleases or stimulates	0.978 (0.048)	0.978 (0.048)
club	a team of professional baseball players who play and travel together	0.877 (0.123)	0.770 (0.077)
	a building that is occupied by a social club	0.016 (0.011)	0.009 (0.006)
	a formal association of people with similar interests	0.939 (0.170)	0.524 (0.095)
	a spot that is open late at night and that provides entertainment (as singers or dancers) as well as dancing and food and drink	0.988 (0.020)	0.879 (0.015)
compass	an area in which something acts or operates or has power or control:	0.953 (0.013)	0.952 (0.013)
	the limit of capability	0.995 (0.019)	0.673 (0.013)
	get the meaning of something	0.969 (0.067)	0.947 (0.065)
	travel around, either by plane or ship	0.049 (0.000)	0.048 (0.000)
consecutive	in regular succession without gaps	0.888 (0.141)	0.171 (0.058)
	in a consecutive manner	-0.003 (0.009)	-0.003 (0.008)
	one after the other	0.643 (0.310)	0.541 (0.257)
	successive (without a break)	0.986 (0.034)	0.986 (0.034)
constitute	to compose or represent:	0.067 (0.000)	0.046 (0.000)
	form or compose	0.870 (0.156)	0.759 (0.135)
	create and charge with a task or function	1.000 (0.001)	0.458 (0.000)
	set up or lay the groundwork for	0.966 (0.001)	0.503 (0.001)
contend	come to terms with	0.932 (0.090)	0.284 (0.036)
	maintain or assert	0.965 (0.049)	0.797 (0.046)
	to make the subject of dispute, contention, or litigation	0.039 (0.009)	0.012 (0.003)
	be engaged in a fight; carry on a fight	0.903 (0.170)	0.685 (0.136)
dampen	smother or suppress	0.590 (0.185)	0.520 (0.164)
	make moist	0.943 (0.217)	0.703 (0.137)
	make vague or obscure or make (an image) less visible	0.932 (0.196)	0.284 (0.005)
	lessen in force or effect	0.323 (0.002)	0.282 (0.002)
dependable	worthy of reliance or trust	0.097 (0.137)	0.096 (0.136)
	worthy of being depended on	0.000 (0.002)	0.001 (0.002)
	financially sound	0.972 (0.035)	0.011 (0.013)
	consistent in performance or behavior	0.864 (0.231)	-0.017 (0.106)
die	stop operating or functioning	0.995 (0.010)	0.995 (0.010)
	pass from physical life and lose all bodily attributes and functions necessary to sustain life	0.997 (0.005)	0.027 (0.000)
	lose sparkle or bouquet	0.932 (0.078)	0.023 (0.002)
	cut or shape with a die	0.000 (0.002)	0.000 (0.001)
disconnected	lacking orderly continuity	0.998 (0.002)	0.763 (0.001)
	(music) marked by or composed of disconnected parts or sounds; cut short crisply	0.896 (0.002)	0.872 (0.002)
	marked by sudden changes in subject and sharp transitions	0.316 (0.003)	0.293 (0.003)
	having been divided; having the unity destroyed; - Samuel Lubell; - E.B.White	0.780 (0.001)	0.716 (0.001)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
doctor	restore by replacing a part or putting together what is torn or broken	0.928 (0.214)	0.930 (0.213)
	alter and make impure, as with the intention to deceive a person who holds Ph.D. degree (or the equivalent) from an academic institution	0.995 (0.013)	0.923 (0.012)
	a licensed medical practitioner	1.000 (0.001)	0.012 (0.000)
		0.002 (0.003)	0.002 (0.003)
draw out	remove, usually with some force or effort; also used in an abstract sense	0.987 (0.000)	0.853 (0.000)
	lengthen in time; cause to be or last longer	0.115 (0.001)	0.105 (0.001)
	remove as if by suction	0.983 (0.058)	0.949 (0.049)
	deduce (a principle) or construe (a meaning)	0.886 (0.054)	0.807 (0.026)
edge	provide with a border or edge	-0.022 (0.075)	-0.000 (0.001)
	advance slowly, as if by inches	0.995 (0.006)	0.370 (0.002)
	lie adjacent to another or share a boundary	0.995 (0.017)	0.008 (0.000)
	the attribute of urgency in tone of voice	0.998 (0.004)	0.968 (0.004)
effect	(of a law) having legal validity	0.997 (0.005)	0.677 (0.002)
	a phenomenon that follows and is caused by some previous phenomenon	0.346 (0.086)	0.346 (0.086)
	an outward appearance	0.970 (0.085)	0.803 (0.059)
	produce	0.620 (0.009)	0.619 (0.009)
evidence	give evidence	0.155 (0.243)	0.088 (0.138)
	your basis for belief or disbelief; knowledge on which to base belief	0.965 (0.126)	0.762 (0.099)
	provide evidence for	0.007 (0.005)	0.004 (0.003)
	provide evidence for; stand as proof of; show by one's behavior, attitude, or external attributes	0.875 (0.132)	0.462 (0.069)
father	the founder of a family	0.997 (0.002)	0.337 (0.001)
	a male parent (also used as a term of address to your father)	0.002 (0.002)	0.000 (0.002)
	a person who founds or establishes some institution	0.809 (0.281)	0.335 (0.162)
	make children	0.843 (0.285)	0.852 (0.267)
fault	a wrong action attributable to bad judgment or ignorance or inattention	0.020 (0.003)	0.019 (0.003)
	(geology) a crack in the earth's crust resulting from the displacement of one side with respect to the other	0.993 (0.010)	0.986 (0.005)
	the quality of being inadequate or falling short of perfection	0.998 (0.002)	0.792 (0.001)
	an imperfection in an object or machine	0.975 (0.010)	0.914 (0.007)
favor	consider as the favorite	0.971 (0.078)	0.435 (0.035)
	an advantage to the benefit of someone or something	0.971 (0.007)	0.964 (0.007)
	an inclination to approve	0.998 (0.001)	0.971 (0.001)
	promote over another	0.027 (0.004)	0.027 (0.002)
gap	an open or empty space in or between things	0.168 (0.002)	0.163 (0.002)
	a conspicuous disparity or difference as between two figures	0.821 (0.010)	0.800 (0.010)
	a narrow opening	0.996 (0.011)	0.804 (0.011)
	an act of delaying or interrupting the continuity	0.998 (0.002)	0.994 (0.002)
get together	get people together	-0.002 (0.003)	-0.002 (0.003)
	a small informal social gathering	0.885 (0.263)	0.846 (0.256)
	work together on a common enterprise or project	0.943 (0.172)	0.606 (0.110)
	become part of; become a member of a group or organization	0.681 (0.345)	0.233 (0.119)
go off	happen in a particular manner	0.815 (0.366)	0.780 (0.207)
	burst inward	0.684 (0.402)	0.651 (0.279)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	run away; usually includes taking something or somebody along	0.219 (0.014)	0.216 (0.014)
	go off or discharge	0.744 (0.047)	0.734 (0.046)
guarantee	make certain of	0.834 (0.251)	0.001 (0.007)
	promise to do or accomplish	0.706 (0.339)	0.700 (0.336)
	stand behind and guarantee the quality, accuracy, or condition of	-0.005 (0.008)	-0.002 (0.003)
	give surety or assume responsibility	0.943 (0.194)	0.008 (0.002)
gush	issue in a jet; come out in a jet; stream or spring forth	0.009 (0.016)	0.007 (0.013)
	gush forth in a sudden stream or jet	-0.001 (0.001)	-0.001 (0.001)
	a sudden rapid flow (as of water)	0.002 (0.001)	0.002 (0.001)
	praise enthusiastically	0.925 (0.227)	0.800 (0.196)
hell	noisy and unrestrained mischief	0.971 (0.034)	0.026 (0.001)
	violent and excited activity	0.989 (0.010)	0.989 (0.010)
	any place of pain and turmoil	0.001 (0.002)	-0.000 (0.000)
	a cause of difficulty and suffering	0.717 (0.255)	0.014 (0.005)
hide	make undecipherable or imperceptible by obscuring or concealing	0.992 (0.020)	0.853 (0.017)
	cover as if with a shroud	0.996 (0.006)	0.926 (0.006)
	prevent from being seen or discovered	0.961 (0.006)	0.011 (0.002)
	be or go into hiding; keep out of sight, as for protection and safety	0.031 (0.001)	-0.000 (0.001)
hike	increase	0.795 (0.022)	0.118 (0.020)
	an increase in cost	0.868 (0.161)	0.807 (0.160)
	the amount a salary is increased	0.990 (0.028)	0.813 (0.025)
	a long walk usually for exercise or pleasure	0.191 (0.010)	0.030 (0.002)
honorable	worthy of being honored; entitled to honor and respect	0.406 (0.009)	0.290 (0.007)
	adhering to ethical and moral principles	0.511 (0.204)	0.365 (0.146)
	not disposed to cheat or defraud; not deceptive or fraudulent	0.941 (0.216)	0.398 (0.088)
	deserving of esteem and respect	0.081 (0.092)	-0.108 (0.095)
ice	decorate with frosting	0.828 (0.028)	0.768 (0.028)
	water frozen in the solid state	0.633 (0.048)	0.455 (0.048)
	a rink with a floor of ice for ice hockey or ice skating	0.510 (0.019)	0.361 (0.016)
	diamonds	0.948 (0.072)	0.417 (0.063)
identical	being the exact same one; not any other:	0.066 (0.000)	0.054 (0.000)
	exactly alike; incapable of being perceived as different	0.985 (0.005)	0.086 (0.000)
	(of twins) derived from a single egg or ovum	0.943 (0.002)	0.919 (0.002)
	coinciding exactly when superimposed	0.978 (0.089)	0.113 (0.005)
incidental	(frequently plural) an expense not budgeted or not specified	0.429 (0.000)	0.222 (0.000)
	(sometimes followed by 'to') minor or casual or subordinate in significance or nature or occurring as a chance concomitant or consequence	0.554 (0.021)	0.287 (0.011)
	not of prime or central importance; - Pubs.MLA	0.996 (0.005)	0.382 (0.001)
	following or accompanying as a consequence	0.632 (0.344)	0.594 (0.330)
incompetent	not doing a good job	0.821 (0.121)	0.694 (0.101)
	legally not qualified or sufficient	0.987 (0.012)	0.983 (0.012)
	not meeting requirements	0.913 (0.132)	0.792 (0.112)
	showing lack of skill or aptitude	-0.095 (0.078)	-0.080 (0.066)
indicate	indicate a place, direction, person, or thing; either spatially or figuratively	0.217 (0.007)	0.002 (0.002)
	suggest the necessity of an intervention; in medicine	0.777 (0.002)	0.007 (0.002)
	be a signal for or a symptom of	0.996 (0.004)	0.986 (0.004)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	give evidence of	0.961 (0.077)	0.875 (0.074)
insert	put or introduce into something	0.633 (0.001)	0.466 (0.001)
	fit snugly into	0.680 (0.294)	0.673 (0.252)
	insert casually	0.349 (0.016)	0.256 (0.012)
	introduce	0.581 (0.364)	0.385 (0.147)
inspire	heighten or intensify	0.034 (0.000)	0.027 (0.000)
	serve as the inciting cause of	0.990 (0.005)	0.649 (0.003)
	draw in (air)	0.919 (0.189)	0.793 (0.162)
	spur on or encourage especially by cheers and shouts	0.965 (0.015)	0.819 (0.013)
intend	denote or connote	0.336 (0.342)	-0.032 (0.153)
	design or destine	0.998 (0.006)	0.996 (0.006)
	have in mind as a purpose	0.956 (0.194)	0.943 (0.191)
	mean or intend to express or convey	0.002 (0.000)	0.002 (0.000)
interrupt	interfere in someone else's activity	0.761 (0.268)	0.648 (0.229)
	destroy the peace or tranquility of	0.987 (0.013)	0.960 (0.012)
	terminate	0.475 (0.341)	0.474 (0.341)
	make a break in	0.001 (0.003)	0.000 (0.003)
invigorate	impart vigor, strength, or vitality to	0.235 (0.008)	0.173 (0.005)
	make lively	0.990 (0.002)	0.705 (0.002)
	give life or energy to	0.803 (0.170)	0.776 (0.168)
	heighten or intensify	0.822 (0.001)	0.545 (0.000)
judge	judge tentatively or form an estimate of (quantities or time)	0.994 (0.013)	0.928 (0.007)
	form a critical opinion of	0.896 (0.205)	0.795 (0.120)
	pronounce judgment on	0.065 (0.132)	0.050 (0.106)
	put on trial or hear a case and sit as the judge at the trial of	0.824 (0.075)	0.647 (0.059)
law	a generalization that describes recurring facts or events in nature	0.983 (0.063)	0.030 (0.002)
	the force of policemen and officers	0.824 (0.225)	0.825 (0.224)
	the collection of rules imposed by authority	0.992 (0.000)	0.026 (0.000)
	the learned profession that is mastered by graduate study in a law school and that is responsible for the judicial system	0.005 (0.003)	0.000 (0.000)
lecture	deliver a lecture or talk	-0.000 (0.000)	-0.000 (0.000)
	censure severely or angrily	0.980 (0.040)	0.896 (0.037)
	a speech that is open to the public	0.938 (0.198)	0.305 (0.064)
	a lengthy rebuke	0.803 (0.282)	0.184 (0.222)
lonely	marked by dejection from being alone	0.982 (0.038)	0.879 (0.034)
	lacking companions or companionship	0.001 (0.002)	0.001 (0.002)
	characterized by or preferring solitude	0.873 (0.147)	0.873 (0.147)
	devoid of creatures	0.864 (0.226)	0.784 (0.206)
lose	be set at a disadvantage	0.993 (0.021)	0.825 (0.018)
	fail to make money in a business; make a loss or fail to profit	0.001 (0.000)	0.001 (0.000)
	fail to perceive or to catch with the senses or the mind	0.987 (0.036)	0.471 (0.017)
	place (something) where one cannot find it again	0.998 (0.004)	0.614 (0.003)
modest	limited in size or scope	0.997 (0.005)	0.825 (0.005)
	not large but sufficient in size or amount	0.998 (0.003)	0.655 (0.002)
	low or inferior in station or quality	0.382 (0.001)	0.382 (0.001)
	humble in spirit or manner; suggesting retiring mildness or even cowed submissiveness	0.603 (0.019)	0.603 (0.019)
nice	exhibiting courtesy and politeness	0.955 (0.075)	0.780 (0.062)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	excessively fastidious and easily disgusted	0.991 (0.014)	0.804 (0.011)
	done with delicacy and skill	0.796 (0.230)	0.498 (0.177)
	socially or conventionally correct; refined or virtuous	0.001 (0.004)	0.001 (0.003)
nigh	(of actions or states) slightly short of or not quite accomplished; all but	0.991 (0.009)	0.591 (0.005)
	being on the left side	0.999 (0.001)	0.512 (0.001)
	not far distant in time or space or degree or circumstances	0.927 (0.166)	0.510 (0.091)
	near in time or place or relationship	-0.000 (0.001)	-0.000 (0.001)
offend	act in disregard of laws, rules, contracts, or promises	0.347 (0.018)	0.291 (0.016)
	strike with disgust or revulsion	0.867 (0.124)	0.612 (0.101)
	hurt the feelings of	0.656 (0.021)	0.543 (0.017)
	cause to feel resentment or indignation	0.987 (0.021)	0.902 (0.021)
orbit	move in an orbit	-0.076 (0.267)	-0.065 (0.230)
	an area in which something acts or operates or has power or control:	0.945 (0.217)	0.843 (0.099)
	the (usually elliptical) path described by one celestial body in its revolution about another	0.896 (0.000)	0.770 (0.000)
	a particular environment or walk of life	0.946 (0.200)	0.914 (0.088)
passage	a journey usually by ship	0.996 (0.005)	0.662 (0.005)
	a bodily reaction of changing from one place or stage to another	0.986 (0.002)	0.686 (0.001)
	the motion of one object relative to another	0.313 (0.003)	0.312 (0.003)
	a path or channel or duct through or along which something may pass	0.689 (0.022)	0.676 (0.022)
pitter-patter	describing a rhythmic beating	0.877 (0.198)	0.872 (0.198)
	rain gently	-0.007 (0.028)	-0.004 (0.023)
	make light, rapid and repeated sounds	0.954 (0.058)	0.761 (0.046)
	as of footsteps	0.980 (0.032)	0.969 (0.031)
plastic	forming or capable of forming or molding or fashioning	0.936 (0.185)	0.937 (0.184)
	capable of being influenced or formed	0.957 (0.066)	0.789 (0.044)
	a card (usually plastic) that assures a seller that the person using it has a satisfactory credit rating and that the issuer will see to it that the seller receives payment for the merchandise delivered	-0.001 (0.001)	-0.001 (0.001)
	capable of being molded or modeled (especially of earth or clay or other soft material)	0.996 (0.002)	0.754 (0.002)
pliable	susceptible to being led or directed	-0.013 (0.204)	-0.013 (0.191)
	capable of being shaped or bent or drawn out	0.968 (0.022)	0.853 (0.019)
	able to adjust readily to different conditions	0.999 (0.001)	0.941 (0.001)
	capable of being bent or flexed or twisted without breaking	0.872 (0.133)	0.809 (0.131)
pluck	pull or pull out sharply	0.088 (0.002)	0.002 (0.000)
	pull lightly but sharply with a plucking motion	0.741 (0.205)	0.021 (0.006)
	strip of feathers	0.975 (0.051)	0.022 (0.001)
	look for and gather	0.963 (0.082)	0.956 (0.082)
plunder	destroy and strip of its possession	0.691 (0.368)	0.582 (0.332)
	steal goods; take as spoils	0.994 (0.014)	0.864 (0.009)
	plunder (a town) after capture	0.346 (0.007)	0.322 (0.007)
	take illegally; of intellectual property	0.644 (0.001)	0.599 (0.001)
plunk	set (something or oneself) down with or as if with a noise	0.777 (0.004)	0.560 (0.003)
	drop steeply	0.216 (0.032)	0.158 (0.023)
	with a short hollow thud	0.987 (0.030)	0.897 (0.030)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	pull lightly but sharply with a plucking motion	0.998 (0.002)	0.805 (0.002)
prevail	be valid, applicable, or true	-0.000 (0.003)	-0.000 (0.001)
	continue to exist	0.947 (0.099)	0.003 (0.000)
	be larger in number, quantity, power, status or importance	0.999 (0.002)	0.987 (0.002)
	prove superior	0.997 (0.006)	0.004 (0.000)
rag	censure severely or angrily	0.962 (0.092)	0.902 (0.092)
	treat cruelly	0.608 (0.029)	0.474 (0.023)
	cause annoyance in; disturb, especially by minor irritations	0.869 (0.207)	0.755 (0.178)
	harass with persistent criticism or carping	0.383 (0.015)	0.300 (0.012)
rattling	quick and energetic	0.535 (0.408)	0.573 (0.310)
	extraordinarily good or great ; used especially as intensifiers	0.999 (0.003)	0.873 (0.002)
	a rapid series of short loud sounds (as might be heard with a stethoscope in some types of respiratory disorders)	0.253 (0.095)	0.187 (0.070)
	used as intensifiers; 'real' is sometimes used informally for 'really'; 'rattling' is informal	0.696 (0.019)	0.515 (0.014)
refer	be relevant to	0.068 (0.031)	0.025 (0.015)
	make reference to	0.769 (0.166)	0.298 (0.066)
	have as a meaning	0.987 (0.025)	0.681 (0.017)
	seek information from	0.924 (0.195)	0.417 (0.096)
relegate	refer to another person for decision or judgment	0.847 (0.259)	0.673 (0.206)
	expel, as if by official decree	0.901 (0.247)	0.900 (0.247)
	assign to a lower position; reduce in rank	0.032 (0.007)	0.026 (0.005)
	assign to a class or kind	0.551 (0.416)	0.431 (0.331)
scourge	a person who inspires fear or dread	0.993 (0.020)	0.968 (0.020)
	whip	0.068 (0.197)	0.025 (0.194)
	cause extensive destruction or ruin utterly	0.809 (0.027)	0.582 (0.019)
	something causing misery or death	0.997 (0.006)	0.968 (0.006)
silly	inspiring scornful pity; - Dashiell Hammett	0.957 (0.103)	0.941 (0.103)
	lacking seriousness; given to frivolity	0.720 (0.298)	0.634 (0.262)
	ludicrous, foolish	0.092 (0.001)	0.081 (0.001)
	dazed from or as if from repeated blows	0.525 (0.381)	0.487 (0.357)
slant	degree of deviation from a horizontal plane	0.628 (0.349)	0.476 (0.267)
	to incline or bend from a vertical position	-0.000 (0.000)	-0.000 (0.000)
	heel over	0.844 (0.226)	0.380 (0.102)
	present with a bias	0.997 (0.005)	0.551 (0.003)
slide	the act of moving smoothly along a surface while remaining in contact with it	0.986 (0.058)	0.734 (0.043)
	to pass or move unobtrusively or smoothly	0.009 (0.001)	0.006 (0.000)
	(music) rapid sliding up or down the musical scale	0.997 (0.006)	0.710 (0.005)
	move obliquely or sideways, usually in an uncontrolled manner	0.986 (0.004)	0.630 (0.003)
slight	being of delicate or slender build; - Frank Norris	0.995 (0.007)	0.868 (0.004)
	(quantifier used with mass nouns) small in quantity or degree; not much or almost none or (with 'a') at least some	0.007 (0.001)	0.006 (0.001)
	lacking substance or significance; ; ; a fragile claim to fame"	0.998 (0.001)	0.952 (0.001)
	pay no attention to, disrespect	0.996 (0.000)	0.995 (0.000)
smack	directly	0.495 (0.394)	0.478 (0.393)
	deliver a hard blow to	0.992 (0.008)	0.007 (0.000)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	have a distinctive or characteristic taste	0.796 (0.218)	0.004 (0.001)
	have an element suggestive (of something)	-0.005 (0.049)	0.000 (0.000)
smother	deprive of the oxygen necessary for combustion	0.925 (0.070)	0.690 (0.052)
	conceal or hide	0.829 (0.240)	0.345 (0.146)
	envelop completely	0.988 (0.017)	0.843 (0.015)
	deprive of oxygen and prevent from breathing	0.002 (0.001)	0.001 (0.001)
so	subsequently or soon afterward (often used as sentence connectors)	0.993 (0.005)	0.308 (0.002)
	in the way indicated; ; ; ('thusly' is a nonstandard variant)	-0.184 (0.350)	-0.101 (0.195)
	in truth (often tends to intensify)	0.995 (0.002)	0.992 (0.002)
	(used to introduce a logical conclusion) from that fact or reason or as a result	0.981 (0.004)	0.549 (0.002)
socialize	prepare for social life	0.995 (0.011)	0.684 (0.007)
	make conform to socialist ideas and philosophies	0.984 (0.050)	0.984 (0.050)
	take part in social activities; interact with others	-0.001 (0.003)	-0.001 (0.002)
	train for a social environment	0.734 (0.301)	0.632 (0.101)
specialise	become more focus on an area of activity or field of study	0.908 (0.081)	0.028 (0.002)
	devote oneself to a special area of work	0.005 (0.002)	0.001 (0.000)
	be specific about	0.838 (0.196)	0.814 (0.192)
	suit to a special purpose	0.994 (0.010)	0.022 (0.000)
specialize	become more focus on an area of activity or field of study	0.913 (0.091)	0.028 (0.003)
	suit to a special purpose	0.994 (0.013)	0.021 (0.000)
	be specific about	0.797 (0.233)	0.773 (0.230)
	devote oneself to a special area of work	0.004 (0.002)	0.001 (0.000)
standard	the ideal in terms of which something can be judged	0.686 (0.223)	0.596 (0.181)
	a basis for comparison; a reference point against which other things can be evaluated	0.251 (0.002)	0.128 (0.001)
	regularly and widely used or sold	0.930 (0.067)	0.399 (0.027)
	conforming to the established language usage of educated native speakers; (American); (British)	0.791 (0.003)	0.389 (0.001)
stark	devoid of any qualifications or disguise or adornment	0.961 (0.106)	0.668 (0.074)
	severely simple	0.001 (0.003)	0.001 (0.003)
	providing no shelter or sustenance	0.997 (0.005)	0.562 (0.003)
	without qualification; used informally as (often pejorative) intensifiers	0.936 (0.091)	0.779 (0.076)
steamroller	make level or flat with a steamroller	0.734 (0.217)	0.734 (0.218)
	overwhelm by using great force	0.949 (0.080)	0.155 (0.003)
	proceed with great force	0.971 (0.034)	0.151 (0.003)
	bring to a specified state by overwhelming force or pressure	0.132 (0.003)	0.132 (0.002)
strangle	kill by squeezing the throat of so as to cut off the air	-0.001 (0.001)	-0.001 (0.001)
	struggle for breath; have insufficient oxygen intake	0.393 (0.348)	0.272 (0.240)
	prevent the progress or free movement of	0.932 (0.116)	0.769 (0.096)
	conceal or hide	0.582 (0.297)	0.456 (0.235)
suggest	suggest the necessity of an intervention; in medicine	0.984 (0.002)	0.777 (0.002)
	make a proposal, declare a plan for something	0.965 (0.037)	0.668 (0.026)
	imply as a possibility	0.025 (0.001)	0.018 (0.001)
	call to mind	0.643 (0.303)	0.024 (0.011)
sway	win approval or support for	0.814 (0.000)	0.802 (0.000)
	move or walk in a swinging or swaying manner	0.179 (0.011)	0.177 (0.011)
	cause to move back and forth	0.995 (0.008)	0.907 (0.008)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	move back and forth or sideways	0.947 (0.210)	0.883 (0.195)
tape	register electronically	0.990 (0.054)	0.813 (0.044)
	memory device consisting of a long thin plastic strip coated with iron oxide; used to record audio or video signals or to store computer information	0.996 (0.008)	0.599 (0.008)
	measuring instrument consisting of a narrow strip (cloth or metal) marked in inches or centimeters and used for measuring lengths	0.987 (0.035)	-0.004 (0.028)
	a recording made on magnetic tape	0.003 (0.001)	0.002 (0.001)
terrible	extreme in degree or extent or amount or impact	0.082 (0.012)	0.080 (0.012)
	causing fear or dread or terror	0.974 (0.023)	0.718 (0.016)
	exceptionally bad or displeasing	0.989 (0.015)	0.639 (0.010)
	intensely or extremely bad or unpleasant in degree or quality	0.864 (0.031)	0.864 (0.029)
total	complete in extent or degree and in every particular	0.011 (0.001)	0.009 (0.001)
	determine the sum of	0.348 (0.216)	0.920 (0.088)
	add up in number or quantity	0.988 (0.022)	0.989 (0.017)
	constituting the full quantity or extent; complete	0.949 (0.090)	0.427 (0.064)
transplant	lift and reset in another soil or situation	0.018 (0.001)	0.013 (0.000)
	transfer from one place or period to another	0.929 (0.135)	0.636 (0.119)
	an operation moving an organ from one organism (the donor) to another (the recipient)	0.999 (0.001)	0.145 (0.000)
	the act of removing something from one location and introducing it in another location	0.926 (0.218)	0.143 (0.006)
unknown	not known before	0.976 (0.002)	0.790 (0.002)
	an unknown and unexplored region	0.034 (0.001)	0.026 (0.001)
	not famous or acclaimed	0.935 (0.093)	0.927 (0.088)
	being or having an unknown or unnamed source	0.985 (0.014)	0.419 (0.006)
vaporise	cause to change into a vapor	0.054 (0.001)	0.000 (0.000)
	change into a vapor	0.263 (0.257)	-0.033 (0.058)
	turn into gas	0.953 (0.077)	0.945 (0.076)
	lose or cause to lose liquid by vaporization leaving a more concentrated residue	0.935 (0.010)	0.008 (0.000)
vaporize	decrease rapidly and disappear	0.940 (0.215)	0.659 (0.166)
	lose or cause to lose liquid by vaporization leaving a more concentrated residue	0.994 (0.014)	0.907 (0.014)
	turn into gas	0.417 (0.100)	0.283 (0.068)
	kill with or as if with a burst of gunfire or electric current or as if by shooting	0.549 (0.000)	0.373 (0.000)
verbalise	be verbose	0.109 (0.008)	0.098 (0.002)
	express in speech	0.833 (0.077)	0.822 (0.076)
	convert into a verb	0.953 (0.037)	0.102 (0.001)
	articulate; either verbally or with a cry, shout, or noise	0.975 (0.004)	0.877 (0.003)
visualise	view the outline of by means of an X-ray	0.386 (0.038)	0.225 (0.022)
	make visible	0.976 (0.080)	0.734 (0.079)
	imagine; conceive of; see in one's mind	0.976 (0.003)	0.700 (0.002)
	form a mental picture of something that is invisible or abstract	0.366 (0.309)	0.162 (0.241)
voice	the sound made by the vibration of vocal folds modified by the resonance of the vocal tract	0.474 (0.014)	0.282 (0.008)
	an advocate who represents someone else's policy or purpose	0.516 (0.002)	0.308 (0.002)
	expressing in coherent verbal form	0.998 (0.003)	0.889 (0.003)

Word	Sense	Mean BasicTPS (SD)	Mean Distance-basedTPS (SD)
	the melody carried by a particular voice or instrument in polyphonic music	0.992 (0.006)	0.532 (0.004)
water	supply with water, as with channels or ditches or streams	0.009 (0.000)	0.006 (0.000)
	liquid excretory product	0.980 (0.051)	0.979 (0.051)
	a facility that provides a source of water	0.985 (0.048)	0.784 (0.038)
	the part of the earth's surface covered with water (such as a river or lake or ocean)	0.990 (0.003)	0.641 (0.002)
waxy	having the paleness of wax; - Bram Stoker	0.986 (0.001)	0.083 (0.000)
	easily impressed or influenced	0.979 (0.046)	0.896 (0.042)
	made of or covered with wax	0.012 (0.001)	0.001 (0.000)
	capable of being bent or flexed or twisted without breaking	0.998 (0.003)	0.803 (0.002)
weigh	show consideration for; take into account	0.936 (0.218)	0.689 (0.087)
	have weight; have import, carry weight	0.646 (0.007)	0.451 (0.005)
	to be oppressive or burdensome; ,	0.992 (0.003)	0.264 (0.000)
	determine the weight of	0.355 (0.000)	0.245 (0.000)
work up	come up with	0.957 (0.017)	0.938 (0.017)
	form or accumulate steadily	0.927 (0.041)	0.043 (0.001)
	bolster or strengthen	0.091 (0.001)	0.035 (0.000)
	develop	0.866 (0.246)	0.605 (0.146)