# Surface Form Competition: Why the Highest Probability Answer Isn't Always Right

**Ari Holtzman**[1]* **Peter West**[1,2]* **Vered Shwartz** [1,2] **Yejin Choi** [1,2] **Luke Zettlemoyer**[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
{ahai,pawest}@cs.washington.edu

## Abstract

Large language models have shown promising results in *zero-shot* settings (Brown et al., 2020; Radford et al., 2019). For example, they can perform multiple choice tasks simply by conditioning on a question and selecting the answer with the highest probability.

However, ranking by string probability can be problematic due to **surface form competition**—wherein different surface forms compete for probability mass, even if they represent the same underlying concept, e.g. "computer" and "PC." Since probability mass is finite, this lowers the probability of the correct answer, due to competition from other strings that are valid answers (but not one of the multiple choice options).

We introduce Domain Conditional Pointwise Mutual Information, an alternative scoring function that directly compensates for surface form competition by simply reweighing each option according to a term that is proportional to its a priori likelihood within the context of the specific zero-shot task. It achieves consistent gains in zero-shot performance over both calibrated (Zhao et al., 2021) and uncalibrated scoring functions on all GPT-2 and GPT-3 models on a variety of multiple choice datasets.[1]

## 1 Introduction

Despite the impressive results large pretrained language models have achieved in zero-shot settings (Brown et al., 2020; Radford et al., 2019), we argue that current work underestimates the zero-shot capabilities of these models on classification tasks. This is in large part due to **surface form competition**—a property of generative models that causes probability to be rationed between different valid strings, even ones that differ trivially, e.g., by capitalization alone. We show such competition can be largely removed by scoring choices according to Domain Conditional Pointwise Mutual Information ($PMI_{DC}$), which reweights scores by how much *more* likely a hypothesis (answer) becomes given a premise (question) within the specific task domain.

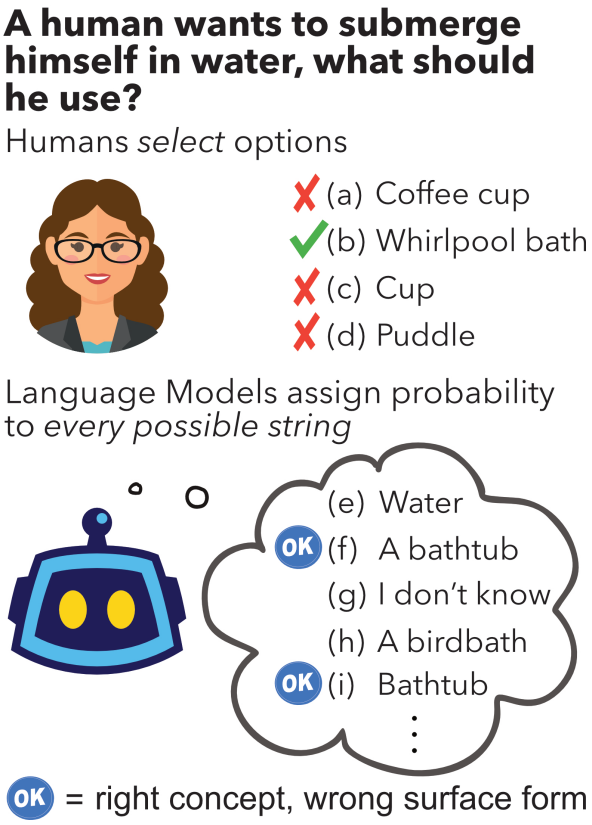


Figure 1: While humans select from given options, language models implicitly assign probability to every possible string. This creates surface form competition between different strings that represent the same concept. Example from CommonsenseQA (Talmor et al., 2019).

More specifically, consider the example question (shown in Figure 1): "A human wants to submerge himself in water, what should he use?" with

*A joint investigation.

[1]Code is available at https://github.com/peterwestuw/surface-form-competition

multiple choice options "Coffee cup", "Whirlpool bath", "Cup", and "Puddle." From the given options, "Whirlpool bath" is the only one that makes sense. Yet, other answers are valid and easier for a language model to generate, e.g., "Bathtub" and "A bathtub." Since all surface forms compete for finite probability mass, allocating significant probability mass to "Bathtub" decreases the amount of probability mass assigned to "Whirlpool bath." While the total probability of generating *some correct answer* may be high (i.e., across all valid surface forms), only one of these is a listed option. This is particularly problematic here, because "Whirlpool bath" will be much lower probability than "Bathtub," due to its rarity. More generally, methods that do not account for surface form competition will favor answers with fewer lexical paraphrases of this type, without any consideration of the task or question under consideration.

$\text{PMI}_{\text{DC}}$ factors out the probability of a specific surface form, instead looking at how much more probable a hypothesis becomes when conditioned on a premise. We use a *domain premise* string to estimate the *unconditional* probability of a hypothesis in a given domain. For CommonsenseQA, for example, we compute the probability of each answer option immediately following the string "? the answer is:", and then divide the *conditional* by this estimate to calculate $\text{PMI}_{\text{DC}}$. This scaling factor renormalizes according to the surface form competition that is inherent to the domain or task, e.g. completions of the domain premise that are just inherently unlikely will be upweighted more. After reweighting we can much more directly measure what the question tells us about the answer and vice versa (the mutual information, see §3 for a full derivation and discussion). Hypotheses no longer need to compete: both "Whirlpool bath" and "Bathtub " will be considered similarly connected to the question after the domain premise reweighting, and so both will attain a high score.

$\text{PMI}_{\text{DC}}$ is the only method that consistently outperforms raw, normalized, and calibrated probability scoring methods on more than a dozen multiple choice datasets and it does so for every model in the GPT-2 and GPT-3 families (§4). To better explain these gains, we show it is possible to use the distinct structure of the COPA dataset (Roemmele et al., 2011) to remove surface form competition entirely, which we call scoring-by-premise, and show that all methods perform well in this idealized setting (§5). Finally, we analyze three exceptions where $\text{PMI}_{\text{DC}}$ does worse than other methods and discuss how inherent differences in dataset construction and task specification cause different methods to work better or worse (§6).

## 2 Background and Related Work

**Zero-shot vs. Few-Shot** Zero-shot inference has long been of interest in NLP, Computer Vision, and ML in general (Socher et al., 2013; Guadarrama et al., 2013; Romera-Paredes and Torr, 2015). However, Radford et al. (2019) popularized the notion that language models have many zero-shot capabilities that simply need to be discovered by prompting the model. For instance placing "TL;DR" (internet slang for Too Long; Didn't Read) at the end of an article causes the model to generate a summary. Efficiently discovering the right prompt is difficult and has become an active area of research (Reynolds and McDonell, 2021; Shin et al., 2020; Jiang et al., 2020).

Brown et al. (2020) demonstrated that few-shot learning without fine-tuning is possible with very large language models. Very recent work has shown it is possible to get smaller models to exhibit few-shot learning behavior, but again using fine-tuning (Schick and Schütze, 2020b,a; Shin et al., 2020; Zhao et al., 2021). Improving zero-shot inference of large models has been a less active area, as most zero-shot behavior is assumed to be directly tied to model quality. It is unclear how to train a better model for a particular task without any task-specific data.

**Surface Form Competition** When applying generative models to multiple choice problems simply choosing the *highest probability* answer becomes problematic due to valid surface forms competing for probability. Indeed, recent work in question answering has demonstrated the importance of considering all multiple choice options together (Khashabi et al., 2020), rather than independently assigning each answer a score and simply choosing the highest. This is a difficult strategy to adapt to left-to-right generative language models, which implicitly choose between *all* possible strings. The use of unsupervised language models pretrained on relatively expansive corpora exacerbates surface form competition because such language models generate a much wider distribution than a given question answering dataset contains.

"What is the most populous nation in North

America?" Posed with this question, a language model such as GPT-3 can generate a correct response such as "USA", "United States", or "United States of America" with high probability. While correct strings like this all contribute to the probability of a correct generation, they may have vastly different probabilities: a common string "United States" will be much more likely than rarer forms like "U.S. of A.". In generative scenarios, as long as most of the probability mass goes to valid strings the generation is likely to be valid. This is not the case for multiple choice problems. Posed with two possible answers, "USA" and "Canada", GPT-3 may still choose the correct answer by probability. However, if we substitute out "USA" for "U.S. of A.", it becomes very likely that GPT-3 will assign higher probability to "Canada", a less likely answer conceptually, but a much more likely surface form. Beyond this, incorrect generic answers such as "I don't know" can take up much of the probability space, relegating the desired answers to the tail of the distribution where calibration with softmax is less trustworthy (Holtzman et al., 2020).

**PMI** Work in dialogue has used PMI to promote diversity (Zhou et al., 2019; Yao et al., 2017; Li et al., 2016; Mou et al., 2016; Tang et al., 2019). Recently, Brown et al. (2020) used a scoring function resembling $\text{PMI}_{\text{DC}}$ for zero-shot question answering, though they only use the string "A:" as a prompt for the unconditional probability estimate, whereas we use a task-specific domain premise (see §3 for details). Furthermore, Brown et al. (2020) only report this scoring method on three datasets (ARC, OpenBookQA, and RACE, included here) out of the more than 20 tested and do not compare scores with their standard method, averaging log-likelihoods (AVG in this work). In contrast, we complete a comprehensive comparison on GPT-3 *and* GPT-2, as well as shedding light on the underlying issue of surface form competition in §5.

**Contextual Calibration** Recently, Zhao et al. (2021) describe a new method for **calibrating** the probabilities of an LM so that straightforward probability-based scoring works better. Though geared towards few-shot learning, the authors devise a clever means of using "dummy" answers for zero-shot learning. Zhao et al. (2021) calibrate for three forms of bias: (1) majority label bias, (2) recency bias, and (3) common token bias. Of these, only (3) applies to the zero-shot case, as "majority label" and "recency" refer to the given few-shot examples' labels and ordering. $\text{PMI}_{\text{DC}}$ directly compensates for common token bias by dividing by the domain conditional probability of each answer without the need for new model parameters, and performs superior to contextual calibration (CC) in the majority of cases.

**Prompt Sensitivity** Recent work highlights the sensitivity of LMs with respect to the *inputs*, and proposes to consider various paraphrases of the prompt to overcome this sensitivity (Davison et al., 2019; Jiang et al., 2020), as well as noting that certain trigger tokens (Shin et al., 2020) can strongly effect the output of such models. In this work we focus on fixing issues regarding the surface form of possible *outputs*.

**Interpreting Language Models** Language models tend to model selectional preferences and thematic fit (Pantel et al., 2007; Erk et al., 2010) which are different from semantic plausibility because the former is more concerned with typicality (Wang et al., 2018). Probability, possibility and plausibility give distinct and relevant notions (Van der Helm, 2006), but reporting bias (Gordon and Van Durme, 2013) means that language models only model what people are likely to write (on websites, that were reachable under the given scrape, etc.). $\text{PMI}_{\text{DC}}$ aims to adjust for these challenges to better measure the underlying agreement between language models and human judgements, but of course is still subject to the limits of the language models it is used with.

## 3 Zero-shot Scoring Strategies

This paper does not define any new modeling or finetuning methods. Rather, we propose the broad use of $\text{PMI}_{\text{DC}}$ scoring for any given model and prompt. $\text{PMI}_{\text{DC}}$ compensates for the fact that different correct answers compete for probability, even though only one will be listed as the correct multiple choice option.

We begin by describing the two most common methods currently in use.

### 3.1 Standard Methods

The first baseline approach is simply selecting the highest-probability option, e.g. baselines in Zhao et al. (2021) and Jiang et al. (2020), which we here refer to as LM. Given a prompt $\mathbf{x}$ (e.g. "The man broke his toe because") and set of possible answers

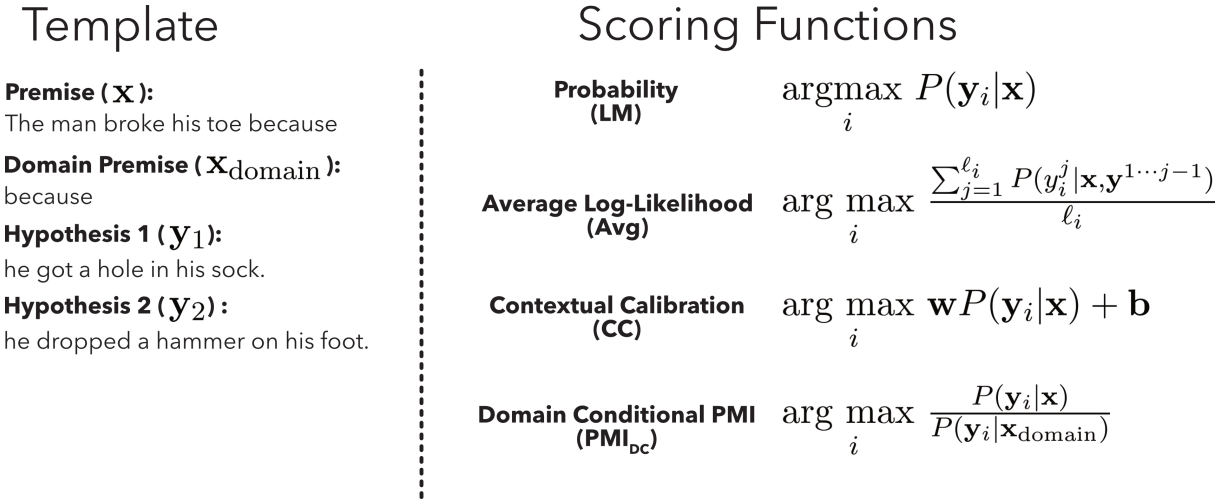


Figure 2: An example from COPA (Roemmele et al., 2011) with the template we use as well as the scoring functions we test. LM returns the highest probability option, while AVG length-normalizes log-likelihoods and chooses the highest option. $\text{PMI}_{\text{DC}}$ is a measurement of the mutual information between hypothesis and premise, intuitively how much $\mathbf{x}$ explains $\mathbf{y}_i$ and vice versa. CC computes an affine transform of LM (adding extra parameters $\mathbf{W}$ and $\mathbf{b}$), adjusting to a given answer set which must be finite and fixed for a given task; $\text{PMI}_{\text{DC}}$ does not require knowledge about possible answers and is readily applied to tasks where answers vary from question to question.

$\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$ (e.g. "he got a hole in his sock.", "he dropped a hammer on his foot"), LM is defined as:

$$\arg\max_i P(\mathbf{y}_i|\mathbf{x}). \quad (1)$$

However, using length normalized log-likelihoods (Brown et al., 2020) has become standard due to its superior performance, and is commonly used for generation tasks (Mao et al., 2019; Oluwatobi and Mueller, 2020). For causal language models, e.g., GPT-2 and GPT-3, Equation 1 can be decomposed as:

$$P(\mathbf{y}_i|\mathbf{x}) = \prod_{j=1}^{\ell} P(y_i^j|\mathbf{x}, y_i^1, \cdots, y_i^{j-1})$$

where $y_i^j$ is the $j$th token of $\mathbf{y}_i$ and $\ell_i$ is the number of tokens in $\mathbf{y}_i$. The AVG strategy is defined as:

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} \log P(y_i^j|\mathbf{x}, \mathbf{y}^{1\cdots j-1})}{\ell_i}.$$

### 3.2 Domain Conditional PMI

Direct probability is not an adequate zero-shot scoring function due to surface form competition. A natural solution is to factor out the probability of specific surface forms, which is what Pointwise Mutual Information (PMI) does:

$$\text{PMI}(\mathbf{x}, \mathbf{y}) = \log \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} = \log \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{x})}. \quad (2)$$

In effect, this is how much more likely the hypothesis ("he dropped a hammer on his foot") becomes given the premise ("The man broke his toe because"). In a multiple-choice setting—where the premise $\mathbf{x}$ does not change across hypotheses—this is proportional to $P(\mathbf{x}|\mathbf{y})$, i.e., the probability of the *premise* given the *hypothesis*. We experiment with this scoring-by-premise in Section 5.

While Equation 2 estimates how related premise $\mathbf{x}$ is to hypothesis $\mathbf{y}$ in general, we found that estimates of $P(\mathbf{y})$ vary wildly. We believe this is because many possible answers are extremely rare in a general setting and therefore their unconditional probability is not well calibrated for the purposes of a given task.

We are specifically trying to measure $P(\mathbf{y})$ in a given domain, e.g., for the "because" relation in our running example, shown in Figures 2 & 3. To quantify this, we propose *Domain Conditional* PMI:

$$\text{PMI}_{\text{DC}}(\mathbf{x}, \mathbf{y}, \text{domain}) = \frac{P(\mathbf{y}|\mathbf{x}, \text{domain})}{P(\mathbf{y}|\text{domain})} \quad (3)$$

$$= \frac{P(\mathbf{y}|\mathbf{x}, \text{domain})}{P(\mathbf{y}|\mathbf{x}_{\text{domain}})} \quad (4)$$

or how much $\mathbf{x}$ tells us about $\mathbf{y}$ within a given domain.

Typically, $P(\mathbf{y}|\mathbf{x}, \text{domain}) = P(\mathbf{y}|\mathbf{x})$ because the premise $\mathbf{x}$ implies the domain in the datasets considered here and for most modern datasets, e.g., "The man broke his toe because" sets the

model up to predict a dependent clause that is the cause of some event, without further representation of the domain. In order to estimate $P(\mathbf{y}|\text{domain})$—the probability of seeing hypothesis $\mathbf{y}$ in this domain—we use a short domain-relevant template $\mathbf{x}_{\text{domain}}$, which we call a "domain premise", often this is just the ending of the conditional premise $\mathbf{x}$. For example, to help predict a causal relation like in Figure 2 we use $\mathbf{x}_{\text{domain}} =$ "because" and thus divide by $P(\mathbf{y}|\text{because})$. For details about templates see Appendix A.

### 3.3 Non-standard Baselines

**Unconditional** We also compare to the unconditional (in-domain) estimate as a scoring function:

$$\arg\max_i P(\mathbf{y}_i|\mathbf{x}_{\text{domain}}). \quad (5)$$

We refer to this as UNC. It ignores the premise completely, only using a domain premise $\mathbf{x}_{\text{domain}}$ (e.g., using $P(\mathbf{y}|\text{because})$ as the score). Yet, it is sometimes competitive, for instance on BoolQ (Clark et al., 2019). UNC is a sanity check on whether zero-shot inference is actually using the information in the question to good effect.

**Contextual Calibration** Finally, we compare to the reported zero-shot numbers of Zhao et al. (2021). *Contextual Calibration* adjusts LM with an affine transform to make a closed set of answers equally likely. Contextual Calibration thus requires adding new parameters to the model and can only be applied to tasks where the set of answers is known in advance, see Zhao et al. (2021) for details. In contrast, $\text{PMI}_{\text{DC}}$ requires nothing but a human-written template (as all zero-shot methods do, including Contextual Calibration), requires no new parameters, and can be used on closed and open answer sets alike.

## 4 Multiple Choice Experiments

### 4.1 Setup

We use GPT-2 via the HuggingFace Transformers library (Wolf et al., 2020) and GPT-3 via OpenAI's beta API.[2] We do not finetune any models, nor do we alter their output. Our code is public, complete with data loading scripts for each dataset to support maximum reproducibility.[3] See the appendix (§A) for example instances from each dataset in our templated format.

### 4.2 Datasets

We report results on 16 splits of 13 datasets, and briefly describe each dataset here.

**Continuation** These datasets require the model to select a continuation to previous text, making them a natural way to test language models. Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011) asks for various "because" and "so" relationships, as shown in Figure 2. StoryCloze (SC) (Mostafazadeh et al., 2017) gives the model a choice between two alternative endings to 5 sentence stories. Finally, HellaSwag (HS) uses GPT-2 to generate, BERT to filter, and crowd workers to verify possible continuations to a passage.

**Question Answering** RACE-M & -H (Lai et al., 2017) (R-M & R-H) are both drawn from English exams given in China, the former being given to Middle Schoolers and the latter to High Schoolers. Similarly, ARC Easy & Challenge (Clark et al., 2018) (ARC-E & ARC-C) are standardized tests described as "natural, grade-school science questions," with the "Easy" split found to be solvable by either a retrieval or word co-occurrence system, and the rest of the questions put in the "Challenge" split. Open Book Question Answering (OBQA) (Mihaylov et al., 2018) is similar to both of these, but was derived using and intended to be tested with a knowledge source (or "book") available; we do not make use of the given knowledge source, following Brown et al. (2020). Finally, CommonsenseQA (CQA) (Talmor et al., 2019) leverages CONCEPTNET (Speer et al., 2017) to encourage crowd workers to write questions with challenging distractors.

**Open Set vs. Closed Set Datasets** The above datasets are all "open set" in that multiple choice answers may be any string. Below we describe "closed set" datasets with a pre-specified set of answers. These are more difficult for the zero-shot case, as this closed set of answers is often somewhat domain specific, but we do not adapt the model to this closed set of answers in any capacity, unlike Contextual Calibration (Zhao et al., 2021).

**Boolean Question Answering** As the bridge between open set and closed set problems, we use BoolQ (Clark et al., 2019) (BQ), which poses ques-

[2] https://beta.openai.com/

[3] https://github.com/peterwestuw/surface-form-competition

Multiple Choice Accuracy on GPT-3

| Params. | 2.7B | | | | | 6.7B | | | | 13B | | | | 175B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Unc | LM | Avg | $PMI_{DC}$ | CC | Unc | LM | Avg | $PMI_{DC}$ | Unc | LM | Avg | $PMI_{DC}$ | Unc | LM | Avg | $PMI_{DC}$ | CC |
| COPA | 0.548 | 0.684 | 0.684 | **0.744** | - | 0.564 | 0.758 | 0.736 | **0.770** | 0.566 | 0.792 | 0.778 | **0.842** | 0.560 | 0.852 | 0.828 | **0.892** | - |
| SC | 0.509 | 0.660 | 0.683 | **0.731** | - | 0.514 | 0.702 | 0.733 | **0.768** | 0.520 | 0.741 | 0.778 | **0.799** | 0.519 | 0.793 | 0.831 | **0.840** | - |
| HS | 0.311 | 0.345 | **0.414** | 0.342 | - | 0.347 | 0.408 | **0.535** | 0.400 | 0.388 | 0.488 | **0.662** | 0.458 | 0.435 | 0.576 | **0.772** | 0.535 | - |
| R-M | 0.224 | 0.378 | 0.424 | **0.426** | - | 0.212 | 0.433 | 0.459 | **0.485** | 0.229 | 0.496 | 0.506 | **0.513** | 0.225 | 0.557 | **0.564** | 0.557 | - |
| R-H | 0.214 | 0.303 | 0.327 | **0.360** | - | 0.220 | 0.348 | 0.368 | **0.398** | 0.229 | 0.382 | 0.392 | **0.421** | 0.222 | 0.424 | 0.433 | **0.437** | - |
| ARC-E | 0.316 | **0.504** | 0.447 | 0.447 | - | 0.335 | **0.582** | 0.523 | 0.515 | 0.338 | **0.662** | 0.597 | 0.577 | 0.362 | **0.735** | 0.670 | 0.633 | - |
| ARC-C | 0.211 | 0.216 | 0.255 | **0.305** | - | 0.218 | 0.268 | 0.298 | **0.330** | 0.223 | 0.321 | 0.343 | **0.385** | 0.226 | 0.402 | 0.432 | **0.455** | - |
| OBQA | 0.100 | 0.172 | 0.272 | **0.428** | - | 0.114 | 0.224 | 0.354 | **0.480** | 0.104 | 0.282 | 0.412 | **0.504** | 0.106 | 0.332 | 0.438 | **0.580** | - |
| CQA | 0.159 | 0.332 | 0.360 | **0.447** | - | 0.174 | 0.400 | 0.429 | **0.503** | 0.164 | 0.488 | 0.479 | **0.585** | 0.163 | 0.610 | 0.574 | **0.667** | - |
| BQ | **0.622** | 0.585 | 0.585 | 0.535 | - | 0.378 | **0.610** | **0.610** | **0.610** | **0.622** | 0.611 | 0.611 | 0.603 | 0.378 | 0.625 | 0.625 | **0.640** | - |
| RTE | 0.473 | 0.487 | 0.487 | **0.516** | 0.495 | 0.527 | **0.552** | **0.552** | 0.487 | 0.527 | 0.527 | 0.527 | **0.549** | 0.473 | 0.560 | 0.560 | **0.643** | 0.578 |
| CB | 0.089 | 0.518 | 0.518 | **0.571** | 0.500 | 0.089 | 0.339 | 0.339 | **0.393** | 0.089 | **0.518** | **0.518** | 0.500 | 0.089 | 0.482 | 0.482 | **0.500** | 0.482 |
| SST-2 | 0.499 | 0.537 | 0.5376 | **0.723** | 0.714 | 0.499 | 0.545 | 0.545 | **0.800** | 0.499 | 0.690 | 0.690 | **0.810** | 0.499 | 0.636 | 0.636 | 0.714 | **0.758** |
| SST-5 | 0.181 | 0.200 | 0.204 | **0.235** | - | 0.181 | 0.278 | 0.227 | **0.320** | 0.181 | 0.186 | **0.296** | **0.191** | 0.176 | 0.270 | 0.273 | **0.296** | - |
| AGN | 0.250 | 0.690 | 0.690 | **0.679** | 0.632 | 0.250 | **0.642** | **0.642** | 0.574 | 0.250 | 0.698 | 0.698 | **0.703** | 0.250 | **0.754** | **0.754** | 0.747 | 0.739 |
| TREC | 0.130 | 0.294 | 0.192 | **0.572** | 0.388 | 0.226 | 0.302 | 0.228 | **0.616** | 0.226 | **0.340** | 0.214 | 0.324 | 0.226 | 0.472 | 0.254 | **0.584** | 0.574 |

Table 1: Comparison of scoring algorithms when using GPT-3 for zero-shot inference on multiple choice questions.

tions based on a multi-sentence passage, with a "yes" or "no" answer.

**Entailment** Entailment datasets focus on the question of whether a hypothesis sentence B is entailed by a premise sentence A. Recognizing Textual Entailment (RTE) (Dagan et al., 2005) requires predicting an "entailment" or "contradiction" label while Commitment Bank (De Marneffe et al., 2019) allows for (a small fraction of) questions to be marked "neutral".

**Text Classification** We consider three more complex classification datasets: SST-2 & -5 (Socher et al., 2013) for various granularities of sentiment classification, AG's News (Zhang et al., 2015) (AGN) for topic classification, and TREC (Li and Roth, 2002) for question classification.

### 4.3 Results

We report results for GPT-3 and GPT-2 in Tables 1 and 2. The trend favoring $PMI_{DC}$ is clear in these tables, but a more digestible summary is shown in Table 3 which aggregates the percentage of splits over which a given method achieves the best score or ties for first-place. In this summarized view it is clear that $PMI_{DC}$ consistently outperforms other scoring methods when aggregated over a variety of datasets. Indeed, the smallest margin (in number of datasets won or tied) between $PMI_{DC}$ and the best competing method is on GPT-3 13B with AVG, but that margin is 50 percentage points. This does not imply that $PMI_{DC}$ is *always* better or that it will be better by a large margin, though it often is. It does suggest that $PMI_{DC}$ is a significantly better bet on a new dataset, as it more consistently matches or outperforms every other method.

## 5 Removing Surface Form Competition

What if we used the probability of the *premise* given the *hypothesis* $P(\mathbf{x}|\mathbf{y}_i)$ instead? While we are still measuring the probability of a surface form (e.g. "the man broke his toe"), it is the *same* surface form across different options ("he had a hole in his sock", "he dropped a hammer on his foot"), eliminating the surface form competition between options. $\mathbf{y}_i$ and $\mathbf{y}'_i$ can now both attain high scores if they are both correct answers and cause $\mathbf{x}$ to be likely. We call this scoring-by-premise.

Due to the nature pretrained causal language models like GPT-2 and GPT-3, it is not usually possible to measure the premise given the hypothesis. This is because the phrasing of the connection between the premise and hypothesis often only works one way, e.g. it is strange to see an answer before a question. We exploit the structure of the COPA dataset (Roemmele et al., 2011) to create a "COPA Flipped" dataset via a simple alternative template, shown in Figure 3. COPA consists of cause and effect pairs (CAUSE *so* EFFECT, and EFFECT *because* CAUSE). In the original dataset, whatever comes second (either CAUSE or EFFECT) has multiple options that a model must choose between. These can be reversed by flipping the order of CAUSE and EFFECT, then substituting

Multiple Choice Accuracy on GPT-2

| Params. | 125M | | | | 350M | | | | 760M | | | | 1.6B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Unc | LM | Avg | $PMI_{DC}$ | Unc | LM | Avg | $PMI_{DC}$ | Unc | LM | Avg | $PMI_{DC}$ | Unc | LM | Avg | $PMI_{DC}$ | CC |
| COPA | 0.564 | 0.610 | **0.632** | 0.628 | 0.558 | 0.670 | 0.660 | **0.700** | 0.556 | 0.698 | 0.676 | **0.694** | 0.560 | 0.690 | 0.684 | **0.716** | - |
| SC | 0.495 | 0.600 | 0.615 | **0.670** | 0.489 | 0.630 | 0.667 | **0.716** | 0.503 | 0.661 | 0.688 | **0.734** | 0.512 | 0.676 | 0.715 | **0.763** | - |
| HS | 0.271 | 0.286 | **0.295** | 0.291 | 0.298 | 0.322 | **0.376** | 0.328 | 0.309 | 0.350 | **0.432** | 0.351 | 0.331 | 0.384 | **0.489** | 0.378 | - |
| R-M | 0.222 | 0.361 | 0.406 | **0.409** | 0.213 | 0.387 | 0.420 | **0.424** | 0.214 | 0.393 | **0.439** | **0.439** | 0.223 | 0.415 | 0.446 | **0.447** | - |
| R-H | 0.209 | 0.275 | 0.310 | **0.344** | 0.215 | 0.304 | 0.326 | **0.363** | 0.215 | 0.318 | 0.345 | **0.383** | 0.219 | 0.330 | 0.357 | **0.391** | - |
| ARC-E | 0.313 | **0.429** | 0.378 | 0.393 | 0.327 | **0.494** | 0.434 | 0.424 | 0.334 | **0.527** | 0.467 | 0.470 | 0.334 | **0.562** | 0.496 | 0.499 | - |
| ARC-C | 0.198 | 0.201 | 0.235 | **0.282** | 0.197 | 0.228 | 0.254 | **0.286** | 0.221 | 0.231 | 0.266 | **0.316** | 0.211 | 0.252 | 0.279 | **0.338** | - |
| OBQA | 0.11 | 0.164 | 0.272 | **0.324** | 0.108 | 0.186 | 0.302 | **0.386** | 0.108 | 0.194 | 0.296 | **0.432** | 0.114 | 0.224 | 0.348 | **0.460** | - |
| CQA | 0.170 | 0.255 | 0.307 | **0.364** | 0.165 | 0.309 | 0.352 | **0.418** | 0.170 | 0.333 | 0.368 | **0.445** | 0.171 | 0.386 | 0.385 | **0.478** | - |
| BQ | **0.622** | 0.588 | 0.588 | 0.511 | **0.622** | 0.608 | 0.608 | 0.497 | **0.622** | 0.580 | 0.580 | 0.467 | **0.622** | 0.563 | 0.563 | 0.495 | - |
| RTE | **0.527** | 0.516 | 0.516 | 0.498 | 0.473 | 0.531 | 0.531 | **0.549** | 0.473 | 0.531 | 0.531 | **0.542** | 0.473 | 0.477 | 0.477 | **0.534** | 0.485 |
| CB | 0.089 | 0.482 | 0.482 | **0.500** | 0.089 | **0.500** | **0.500** | **0.500** | 0.089 | 0.482 | 0.482 | **0.500** | 0.089 | **0.500** | **0.500** | **0.500** | 0.179 |
| SST-2 | 0.499 | 0.636 | 0.636 | **0.671** | 0.499 | 0.802 | 0.802 | **0.862** | 0.499 | 0.770 | 0.770 | **0.856** | 0.499 | 0.840 | 0.840 | **0.875** | 0.820 |
| SST-5 | 0.181 | 0.274 | 0.244 | **0.300** | 0.176 | 0.185 | 0.272 | **0.393** | 0.176 | 0.203 | **0.267** | 0.220 | 0.176 | 0.304 | 0.291 | **0.408** | - |
| AGN | 0.250 | 0.574 | 0.574 | **0.630** | 0.250 | 0.643 | 0.643 | **0.644** | 0.250 | 0.607 | 0.607 | **0.641** | 0.250 | 0.648 | 0.648 | **0.654** | 0.600 |
| TREC | 0.226 | 0.230 | 0.144 | **0.364** | 0.226 | **0.288** | 0.122 | 0.216 | 0.226 | 0.228 | 0.226 | **0.440** | 0.226 | 0.228 | 0.240 | 0.328 | **0.340** |

Table 2: Comparison of scoring algorithms when using GPT-2 for zero-shot inference on multiple choice questions.

Percent of Ties or Wins by Method

| Method | Unc | LM | Avg | $PMI_{DC}$ | CC |
|---|---|---|---|---|---|
| 125M | 12.50 | 6.25 | 12.50 | **68.75** | - |
| 350M | 6.25 | 18.75 | 12.50 | **68.75** | - |
| 760M | 6.25 | 6.25 | 12.50 | **75.00** | - |
| 1.6B | 6.25 | 12.50 | 12.50 | **80.00** | 20.00 |
| 2.7B | 6.25 | 6.25 | 6.25 | **86.66** | 0.00 |
| 6.7B | 6.25 | 25.00 | 25.00 | **75.00** | - |
| 13B | 6.25 | 18.75 | 18.75 | **68.75** | - |
| 175B | 6.25 | 12.50 | 18.75 | **62.50** | 6.25 |

Table 3: Percentage of datasets that a given method produced the best score or was tied for best score with other methods, aggregated over each model size. The first four rows use GPT-2 and summarize data from Table 2, while the final four rows use GPT-3 and summarize data from Table 1. Since ties are included, rows sometimes sum to more than 100. CC is only measured on the 5 datasets we use where Zhao et al. (2021) also report accuracies.

the natural inverse relation ("because"→"so" and "so"→"because" ).

Table 4 shows scores on COPA and COPA Flipped side-by-side. The clearest trend is that in COPA Flipped everything except UNC (Unc in the table) produces the *exact* same result. This is because flipping the hypothesis and premise means that it's the *context* that changes and not the *continuation*. LM, AVG, and $PMI_{DC}$ only differ from each other over different continuations, not over different contexts for the same continuation.

The second important observation is that in COPA Flipped all of these methods generally do about as well as $PMI_{DC}$ on the unflipped version. Indeed, on average they do a little bit better! This is because surface form competition has been eradicated: as the continuation being scored is always the same, it does not matter if other continuations would have been equally good. In COPA Flipped, it only matters how well the context explains the continuation, as measured via ease of prediction. This is not subject to surface form competition because there is only one hypothesis to explain, it is not competing with any other hypotheses for probability mass.

Not all datasets are so easily flippable, so manually flipping individual questions is not a generally applicable strategy. Luckily, $PMI_{DC}$ is symmetric across flipping:

$$\begin{aligned}
&\arg\max_i \frac{P(\mathbf{y}_i|\mathbf{x}, \text{domain})}{P(\mathbf{y}_i|\text{domain})} \\
&= \arg\max_i \frac{P(\mathbf{x}|\mathbf{y}_i, \text{domain})}{P(\mathbf{x}|\text{domain})} \\
&= \arg\max_i P(\mathbf{x}|\mathbf{y}_i, \text{domain})
\end{aligned}$$

In theory, $PMI_{DC}$'s selected answer should be the same between on COPA and COPA Flipped, though we expect small differences due to "so" and "because" not being perfect inverses. Actually, the flipped score meets or exceeds $PMI_{DC}$ on the unflipped data on most models in this setting. One possible reason for this is that on COPA Flipped

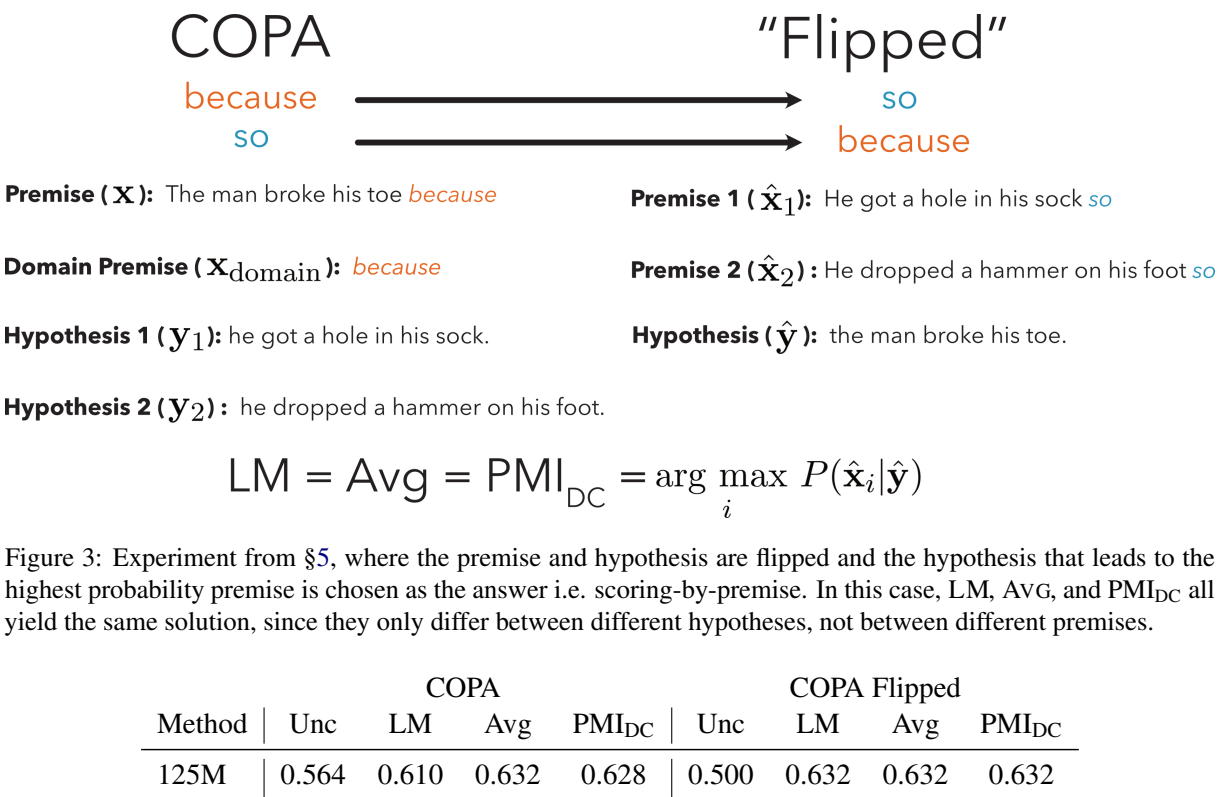


Figure 3: Experiment from §5, where the premise and hypothesis are flipped and the hypothesis that leads to the highest probability premise is chosen as the answer i.e. scoring-by-premise. In this case, LM, AVG, and $\text{PMI}_{\text{DC}}$ all yield the same solution, since they only differ between different hypotheses, not between different premises.

| | COPA | | | | COPA Flipped | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Unc | LM | Avg | $\text{PMI}_{\text{DC}}$ | Unc | LM | Avg | $\text{PMI}_{\text{DC}}$ |
| 125M | 0.564 | 0.610 | 0.632 | 0.628 | 0.500 | 0.632 | 0.632 | 0.632 |
| 350M | 0.558 | 0.670 | 0.660 | 0.700 | 0.500 | 0.664 | 0.664 | 0.664 |
| 760M | 0.556 | 0.698 | 0.676 | 0.694 | 0.500 | 0.708 | 0.708 | 0.708 |
| 1.6B | 0.560 | 0.690 | 0.684 | 0.716 | 0.500 | 0.730 | 0.730 | 0.730 |
| 2.7B | 0.548 | 0.684 | 0.684 | 0.744 | 0.500 | 0.684 | 0.684 | 0.684 |
| 6.7B | 0.564 | 0.758 | 0.736 | 0.770 | 0.500 | 0.768 | 0.768 | 0.768 |
| 13B | 0.566 | 0.792 | 0.778 | 0.842 | 0.500 | 0.790 | 0.790 | 0.790 |
| 175B | 0.560 | 0.852 | 0.828 | 0.892 | 0.500 | 0.836 | 0.836 | 0.836 |

Table 4: To demonstrate the presence of surface form competition, we show that methods that don't directly adjust for competing surface forms (i.e. LM and AVG) have the exact same score as $\text{PMI}_{\text{DC}}$ when scoring the premise. Indeed, using this method LM can do just as well as $\text{PMI}_{\text{DC}}$ applied to the non-flipped version of COPA because surface form competition is not present in this case. For details see §5.

there is no division to calculate $\text{PMI}_{\text{DC}}$; since language model probabilities are approximations, division may induce multiplicative error.

## 6 Analysis

**Failure Cases** There are three datasets on which $\text{PMI}_{\text{DC}}$ does not consistently perform better than other scoring methods: HellaSwag, ARC Easy, and BoolQ. Interestingly (and without any foreknowledge on our part) each one is dominated by a different scoring method.

HellaSwag is most amenable to AVG. After examining a number of data points and the scores of each algorithm, we observe that HellaSwag is more focused on the *internal coherence* of the hypotheses themselves (given basic topic information from the premise), rather than *external coherence*, e.g., a hypothesis being a valid or specific answer to a posed question. This is appears to be largely because it was generated by GPT-2 (Radford et al., 2019) and filtered with BERT, which results in relatively on-topic but somewhat odd hypotheses that humans can distinguish from natural data.

ARC Easy yields highest scores to LM, i.e. just selecting the highest probability option. Upon manual observation, the strongest pattern we were able to find was the presence of stock answers, e.g., the fact that clouds are generated when "ocean water evaporates and then condenses in the air." Indeed, Clark et al. (2018) note that ARC Easy are the questions that were solved either by a retrieval or word co-occurrence baseline, while examples that were answered incorrectly by both were put into

the ARC Hard split. This appears to adequately explain our observations.

Finally, BoolQ, a binary reading comprehension dataset, in which all answers are either "yes" or "no" is best solved by an unconditional baseline. This is largely because the dataset presents truly complex questions, that require more reasoning than GPT-2 or 3 are capable of doing out-of-the-box. The unconditional baseline simply manages to infer the majority label without data, but none of the methods reported actually do better than the majority baseline. The one exception is $\text{PMI}_{\text{DC}}$ with the largest GPT-3 model, but that is a less than 2 percentage point improvement.

**Why does length normalization work?** Past work offers little explanation for why AVG should be a successful strategy for choosing the correct answer, other than the vague intuitive notion that estimates are strongly length biased and require some kind of length compensation. Length bias may be caused by the final softmax layer of current language models assigning too much probability mass to irrelevant options at each time-step, a property noted in open-ended generation, character-level language modeling, and machine translation (Holtzman et al., 2020; Al-Rfou et al., 2019; Peters et al., 2019). If this is the issue, it would mean that longer sequences are even less probable than they should be, biasing the model towards shorter answers.

Another possible argument is that length normalization may account for *unconditional probability* in a similar way to $\text{PMI}_{\text{DC}}$. Length normalization is measured over Byte Pair Encoding (BPE) tokens (Sennrich et al., 2016) and BPE tends to produce a vocabulary such that most tokens are equally frequent (Wang et al., 2020). Furthermore, recent evidence suggests that language is approximately uniformly information dense (Levy, 2018; Levy and Jaeger, 2007; Jaeger, 2006). As such, length in BPE tokens may correspond roughly to a *unigram* estimate of log-probability, supposing that BPE tokens have approximately uniform unigram log-probability. The adjustment made by AVG is still somewhat different than $\text{PMI}_{\text{DC}}$, (division of log terms rather than subtraction) but could have a similar effect. On inspected examples AVG and $\text{PMI}_{\text{DC}}$ often agreed on answer ranking more than other pairs of methods. We leave exploration of this phenomenon to future work.

## 7 Discussion

§5 shows that surface form competition is the cause of certain inadequacies with the standard probability-based scoring methods. $\text{PMI}_{\text{DC}}$ alleviates this, but so does scoring-by-premise in which the probability of the premise given the hypothesis is used as the score. The takeaway here is that generative models assigning probability to a given option isn't the same as selection. $\text{PMI}_{\text{DC}}$ is not a panacea for zero-shot inference. Rather, it is a technique that aligns the prediction being made by the model with the actual task humans would like to get done, which is closer to "choose the hypothesis that explains the premise" than "generate the exact surface form of the hypothesis".

Generative models are density estimation functions that assign probability to every possible string, which is a very different task than the human act of selecting an answer. We expect surface form competition anywhere that generative models are used for selection where there is sufficient uncertainty. There are also many more complex tasks we would like to apply generative models to, such as giving a model a prompt and having it write an essay, but they largely still elude us. Perhaps some of the difficulty comes from using these models in a way that doesn't align with the specific nature of the task.

Consider long-form generation, where currently the best we can do is sample from a large language model. This strategy is clearly problematic because language models marginalize over the contextual probability of every author they've encountered (weighted by the frequency with which they've encountered them) and then draw from this wide and extremely noisy distribution, a kind of "authorship competition".

No explicit modeling of concepts was needed to counteract surface form competition, instead using a variation of PMI for a scoring function was enough. We tentatively posit that similarly general methods (in contrast to explicit structure) may help in many other areas, especially as training becomes infeasible for most individual labs due to the cost and engineering effort required to build extremely large language models.

## 8 Conclusion

We conduct a large-scale comparison of standard and recent scoring functions for zero-shot inference across all GPT-2 and GPT-3 models. We show that

Domain Conditional PMI consistently outperforms previous scoring functions on a wide variety of multiple choice datasets. We also argue that compensating for *surface form competition* is the cause of this boost, by demonstrating that other methods work just as well as Domain Conditional PMI when surface form competition is eliminated. We analyze failure cases of our proposed method and show that they are tied to dataset construction and task definition. Finally, we report a set of qualitative analyses to explain why Domain Conditional PMI outperforms existing zero-shot scoring methods.

## Acknowledgments

This work was supported in part by the ARO (AROW911NF-16-1-0121), the NSF (IIS-1562364), DARPA under the MCSprogram through NIWC Pacific (N66001-19-2-4031) and the Allen Institute for AI (AI2). We thank Mitchell Wortsman, Gabriel Ilharco, and Tim Dettmers for giving thorough and insightful feedback on preliminary drafts.

## References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2019. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3159–3166.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *International Conference on Learning Representations*.

Tim Florian Jaeger. 2006. *Redundancy and syntactic reduction in spontaneous speech*. Ph.D. thesis, Stanford University Stanford, CA.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Roger Levy. 2018. Communicative efficiency, uniform information density, and the rational speech act theory. In *CogSci*.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. Improving neural story generation by targeted common sense grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5990–5995.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358.

Olabiyi Oluwatobi and Erik Mueller. 2020. DLGNet: A transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 54–62, Online. Association for Computational Linguistics.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 564–571, Rochester, New York. Association for Computational Linguistics.

Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.

Ruud Van der Helm. 2006. Towards a clarification of probability, possibility and plausibility: how semantics could help futures practice to improve. *Foresight*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9154–9160.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

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

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2190–2199.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28:649–657.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. 2019. Unsupervised context rewriting for open domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1834–1844.

## A Templates

Table 5 shows the template used for each model with an example instance.

| Type | Dataset | Template |
|---|---|---|
| Continuation | COPA | [The man broke his toe]$_{\text{P}}$ [because]$_{\text{DP}}$ [he got a hole in his sock.]$_{\text{UH}}$ |
| | | [I tipped the bottle]$_{\text{P}}$ [so]$_{\text{DP}}$ [the liquid in the bottle froze.]$_{\text{UH}}$ |
| | StoryCloze | [Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week.]$_{\text{P}}$ [The story continues:]$_{\text{DP}}$ [Jennifer felt bittersweet about it.]$_{\text{UH}}$ |
| | HellaSwag | [A female chef in white uniform shows a stack of baking pans in a large kitchen presenting them. the pans]$_{\text{P}}$ [contain egg yolks and baking soda.]$_{\text{UH}}$ |
| QA | RACE | [There is not enough oil in the world now. As time goes by, it becomes less and less, so what are we going to do when it runs out [...].]$_{\text{P}}$ question: [According to the passage, which of the following statements is true]$_{\text{P}}$[?]$_{\text{DP}}$ answer: [There is more petroleum than we can use now.]$_{\text{UH}}$ |
| | ARC | [What carries oxygen throughout the body?]$_{\text{P}}$ [the answer is:]$_{\text{DP}}$ [red blood cells.]$_{\text{UH}}$ |
| | OBQA | [Which of these would let the most heat travel through?]$_{\text{P}}$ [the answer is:]$_{\text{DP}}$ [a steel spoon in a cafeteria.]$_{\text{UH}}$ |
| | CQA | [Where can I stand on a river to see water falling without getting wet?]$_{\text{P}}$ [the answer is:]$_{\text{DP}}$ [bridge.]$_{\text{UH}}$ |
| Boolean QA | BoolQ | title: [The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 [...]]$_{\text{P}}$ question: [Have the San Jose Sharks won a Stanley Cup?]$_{\text{P}}$ [answer:]$_{\text{DP}}$ [No.]$_{\text{UH}}$ |
| Entailment | RTE | [Time Warner is the world's largest media and Internet company.]$_{\text{P}}$ question: [Time Warner is the world's largest company.]$_{\text{P}}$ [true or false? answer:]$_{\text{DP}}$ [true.]$_{\text{UH}}$ |
| | CB | question: Given that [What fun to hear Artemis laugh. She's such a serious child.]$_{\text{P}}$ Is [I didn't know she had a sense of humor. ]$_{\text{P}}$ true, false, or neither? [the answer is:]$_{\text{DP}}$ [true.]$_{\text{UH}}$ |
| Text Classification | SST-2 | "[Illuminating if overly talky documentary]$_{\text{P}}$" [[The quote] has a tone that is]$_{\text{DP}}$ [positive.]$_{\text{UH}}$ |
| | SST-5 | "[Illuminating if overly talky documentary]$_{\text{P}}$" [[The quote] has a tone that is]$_{\text{DP}}$ [neutral.]$_{\text{UH}}$ |
| | AG's News | title: [Economic growth in Japan slows down as the country experiences a drop in domestic and corporate [...]]$_{\text{P}}$ summary: [Expansion slows in Japan]$_{\text{P}}$ [topic:]$_{\text{DP}}$ [Sports.]$_{\text{UH}}$ |
| | TREC | [Who developed the vaccination against polio?]$_{\text{P}}$ [The answer to this question will be]$_{\text{DP}}$ [a person.]$_{\text{UH}}$ |

Table 5: The templates used for each task, along with an example instance (with a single random candidate answer). Original questions (premises) are colored blue, and original answers (hypotheses) are colored red. Long premises are abbreviated with "[...]". The full premises, conditional hypotheses and domain premises are marked in $[\cdot]_{\text{P}}$, $[\cdot]_{\text{UH}}$, and $[\cdot]_{\text{DP}}$ respectively. For a complete description of our templating methodology, please see our code at https://github.com/peterwestuw/surface-form-competition