Less Likely Brainstorming: Using Language Models to Generate Alternative Hypotheses

Liyan Tang[◇] **Yifan Peng**[♠] **Yanshan Wang**[♣] **Ying Ding**[◇]
Greg Durrett[◇] **Justin F. Rousseau**[◇]
[◇]The University of Texas at Austin
[♠]Weill Cornell Medicine [♣]University of Pittsburgh
lytang@utexas.edu

Abstract

A human decision-maker benefits the most from an AI assistant that corrects for their biases. For problems such as generating interpretation of a radiology report given findings, a system predicting only highly likely outcomes may be less useful, where such outcomes are already obvious to the user. To alleviate biases in human decision-making, it is worth considering a broad differential diagnosis, going beyond the most likely options. We introduce a new task, "less likely brainstorming," that asks a model to generate outputs that humans think are relevant but less likely to happen. We explore the task in two settings: a brain MRI interpretation generation setting and an everyday commonsense reasoning setting. We found that a baseline approach of training with less likely hypotheses as targets generates outputs that humans evaluate as either likely or irrelevant nearly half of the time; standard MLE training is not effective. To tackle this problem, we propose a controlled text generation method that uses a novel contrastive learning strategy to encourage models to differentiate between generating likely and less likely outputs according to humans. We compare our method with several state-of-the-art controlled text generation models via automatic and human evaluations and show that our models' capability of generating less likely outputs is improved.[1]

1 Introduction

Cognitive errors occur when an abnormality is identified, but its importance is incorrectly understood, resulting in an incorrect final diagnosis (Onder et al., 2021; Bruno et al., 2015). For example, radiologists may look for confirmatory evidence to support a diagnostic hypothesis and ignore or discount evidence that refutes the hypothesis (confirmation bias; Busby et al. (2018); Onder et al. (2021)). One way to reduce the likelihood of such cognitive errors is to provide cognitive "help" by having a devil's advocate (Seah et al., 2021; Waite et al., 2017). For this purpose, we propose a new text generation task called "**less likely brainstorming**" to produce less likely but relevant consultations to bring fresh eyes to examine a case—a powerful way to correct diagnostic errors.

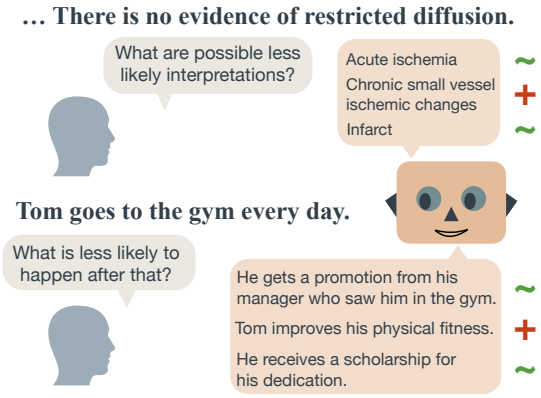

Figure 1: Examples from MRIINTERPRET and E-CARE datasets. The task is to generate interpretations or hypotheses that humans would consider to be "less likely" to happen but still relevant to the context. "+" and "∼" represent likely and less likely outputs, respectively.

Here, we consider less likely hypotheses in two scenarios. First, they can be hypotheses that humans think are likely but not among the most likely to happen. These hypotheses are critical to providing second opinion of a prior clinical study but are often difficult to generate by traditional decoding techniques. Second, they can be hypotheses that *are* indeed impossible according to humans, but are close to being true if certain counterfactual assumptions about the input hold. These hypotheses are also helpful as they are often ignored by clinicians. There is a tendency for clinicians to look for a confirmatory diagnostic hypothesis but ignore a refutable one. Note that a less likely hypothesis

[1]Code is available at https://github.com/Liyan06/Brainstorm.

reflects the likelihood of a potential diagnosis *from the human perspective*, not from the probability of model output.

We propose BRAINSTORM, a novel contrastive learning strategy to generate "less likely" hypotheses. We treat this problem as a text generation task as text generation models are the most flexible for providing predictions and explanations for complex tasks; they can generalize to new examples and produce complex, structured diagnoses in many formats. Generation of the "less likely hypotheses" is conditioned on an indicator variable set to trigger the model to prefer outputs are less likely according to humans. For this purpose, we propose two additional loss objectives to effectively learn the relationship between input context, the indicator, and outputs. Without our training strategy, using naive controlled generation training, we find that conditioning on the indicator often leads to generating "highly likely" or irrelevant outputs.

We explore this task in two settings: everyday commonsense reasoning and brain magnetic resonance imaging (MRI) interpretation generation (more details in Section 5). In the everyday commonsense reasoning setting, we adapt ART (Bhagavatula et al., 2020) and E-CARE (Du et al., 2022), which both contain "less plausible" or "implausible" hypotheses that fit our definition of less likely. An illustrative example asking for less likely hypotheses can be found in Figure 1. We show that our approach can generate more "less likely" hypotheses than baselines, including models directly fine-tuned on this set, past controllable generation approaches (Lu et al., 2022), or models with alternate decoding (Li et al., 2022; Liu et al., 2021). In the brain MRI interpretation setting, we experiment with predicting diagnoses from brain MRI reports (see Figure 1). Assessment by a neurologist reveals that our model successfully shifts the distribution of generated diagnoses further toward the tail while still generating relevant diagnoses.

2 Related Work

Uncertainty in Radiology Interpretation Uncertainty plays a significant role in the process of clinical decision making (Croskerry, 2013). When facing uncertainty, physicians may resort to various erroneous strategies, such as denying the presence of uncertainty resulting in various interpretation biases. These biases could lead to unexpected consequences (Kim and Lee, 2018; Eddy, 1984), including missed diagnoses, misdiagnoses, unnecessary diagnostic examinations and even life-threatening situations (Farnan et al., 2008). Recent work (Seah et al., 2021; Waite et al., 2017) have provided deep-learning based methods and suggestions in reducing errors from interpretation bias on medical imaging. To the best of our knowledge, we are the first to explore reducing bias from interpreting radiology reports via our less likely text generation framework.

Controllable text generation and decoding methods Controllable text generation is the task of generating text that adheres certain attributes, such as language detoxification (Zhang and Song, 2022; Liu et al., 2021; Dathathri et al., 2020), formality modification (Mireshghallah et al., 2022; Yang and Klein, 2021) and open-ended story generation (Mori et al., 2022; Lin and Riedl, 2021; Fan et al., 2018). The task of controllable text generation encompasses both training-time and decoding-time methods. Training-time approaches include CTRL (Keskar et al., 2019), which learns to utilize control codes to govern attributes in order to generate the desired text, and QUARK (Lu et al., 2022), which leverages a strong attribute classifier as a reward function to unlearn unwanted attributes. These methods typically rely on training data that contains both the desired and undesired attributes to be effective in the supervised setting. Our method falls into this category.

On the other hand, decoding-time methods utilize off-the-shelf pre-trained LMs (PLMs) and aim to re-rank the probability of generated text based on specific constraints. PPLM (Dathathri et al., 2020) and FUDGE (Yang and Klein, 2021) are typical methods in this category that train an attribute classifier to guide PLMs to generating desired text. DEXPERTS (Liu et al., 2021) and Contrastive Decoding (Li et al., 2022) are more recent methods that re-weight generation probabilities by contrasting the output distributions between different LMs. We select those two as strong baselines for comparison against our proposed model.

Contrastive Learning in NLP Contrastive learning (CL) has been applied to a wide range of representation learning tasks in NLP, such as learning task-agnostic sentence representation (Gao et al., 2021) and improving natural language understanding (Jaiswal et al., 2021; Qu et al., 2021). It has recently been applied to text generation tasks as

well (An et al., 2022; Cao and Wang, 2021; Lee et al., 2021) where additional hard positive or negative examples are created through techniques such as back-translation or perturbation.

3 Problem Setting

The problem we tackle in this work can be viewed as a controllable text generation task. Let x be a premise or a brain MRI report findings, we want a model to generate a likely/less likely hypothesis or interpretation y given an indicator i by drawing from the distribution $P(y \mid x, i)$. The indicator i can take two values: $+$ to indicate generating likely outputs and $\sim$ to generate less likely outputs.

For example, given a premise x ="*Tom goes to the gym every day.*" in Figure 1 from the E-CARE dataset (more details in Section 5), we want a model to generate a hypothesis $y^{\sim}$ that is less likely to happen ($i = \sim$) after x, such as *"He gets a promotion from his manager who saw him in the gym."*. Although this hypothesis fits into the same scenario as the premise as it directly connects to the premise involving Tom's daily gym attendance, it is less likely to happen since the causal relationship between going to the gym and receiving a promotion is not common. The understanding of what is "less likely" can be based on the concept of bounded rationality (Simon, 1955), where likely hypotheses are those that are likely given known premises, but less likely hypotheses may stem from additional unknown premises.

It is important to note that when we refer to an output as "less likely/likely", we mean that it is less likely/likely based on human understanding of x. All models we experiment with in this work generate outputs that have high probability according to the model, regardless of whether they are likely or less likely to happen according to humans.

4 Methodology

In this section, we present our method as well as baseline models we compare against. Requirements for these models can be found in Table 1. We use BART (Lewis et al., 2020) as the backbone LM for all experimental settings.

4.1 BRAINSTORM

Our encoder-decoder system takes the concatenation of a pair (x, i) as input and returns one or multiple generated output sequences y. At decoding time t, our model iteratively decodes the next token conditioned on the left-hand context, i.e., $y_{<t}$:

$$P_{\text{LM}}(y) = \prod_{t}^{T} P_{\text{LM}}(y_t \mid x, i, y_{<t}) \quad (1)$$

where $P_{\text{LM}}(y_t \mid x, i, y_{<t})$ is the next token distribution given the context. The task inputs are described in Section 5.

Besides the standard maximum likelihood training with human reference, we incorporate two additional loss objectives to guide models to associate the context, indicators, and target sequences. The training approach is illustrated in Figure 2.

Margin Loss First, given the indicator i, we want the model to assign a higher estimated probability to human reference y than its opposite indicator $\neg i$. Therefore, we apply a margin-based loss:

$$\mathcal{L}_{\text{margin}} = \max(0, P(y \mid x, \neg i) - P(y \mid x, i) + m) \quad (2)$$

where m is the margin value. This loss objective tells models that if the indicator is modified, then the target sequence should have lower probability. Margin loss does not require both likely and less likely outputs y^{+} and $y^{\sim}$.

Similarity Loss We propose two versions of a contrastive similarity loss based on the availability of examples that can be used in CL. When both positive and negative examples are available in the same batch, we define the similarity loss as

$$\mathcal{L}_{\text{sim}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_{x,i}, \mathbf{z}_y)/\tau)}{\sum_{\hat{y} \in \text{batch}} \exp(\text{sim}(\mathbf{z}_{x,i}, \mathbf{z}_{\hat{y}})/\tau)} \quad (3)$$

Here, $\mathbf{z}_{x,i}$, $\mathbf{z}_y$, and $\mathbf{z}_{\hat{y}}$ represent the hidden representations of input (x, i), human reference y, and an output $\hat{y}$ in the same batch. $\mathcal{L}_{\text{sim}}$ encourages the model to maximize the agreement between $\mathbf{z}_{x,i}$ and its corresponding output $\mathbf{z}_y$. This loss objective encourages a model to learn the relation between certain indicators and the target sequence by contrasting the target sequence with all negative outputs in the batch.

This objective term resembles that in CoNT (An et al., 2022) which takes self-generated outputs as negative samples; here, we conditioned the input on special indicators. Note that at the training time, the indicator i could be either $+$ or $\sim$. When the indicator $i = +$, the hard negative is the human reference of $y^{\sim}$, and vice versa. We set the weight of the term in Equation (3) associated with the

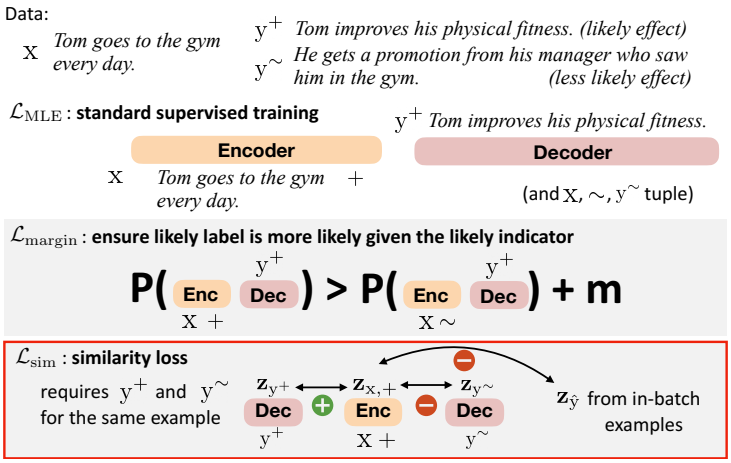

Figure 2: An overview of BRAINSTORM using an example from E-CARE, which consists of three objectives. $\mathbf{z}_{x,i}$ is the encoder representation of the input x conditioned on an indicator i. $\mathbf{z}_{y^+}, \mathbf{z}_{y^\sim}$ and $\mathbf{z}_{\hat{y}}$ are the decoder representations of positive, hard negative, and other negative target sequences within the same batch, respectively. The $\mathcal{L}_{\text{sim}}$ objective is highlighted in red where it requires both likely and less likely data.

hard negative to 10 throughout the experiment to increase its importance relative to in-batch negatives.

When positive and negative examples are not available at the same time (denoted by a lack of a "pair" check in Table 1), we propose an alternative similarity loss objective $\mathcal{L}'_{\text{sim}}$ that minimizes the similarity of encoder representation $\mathbf{z}_{x,i}$ and $\mathbf{z}_{x,\neg i}$, without comparing to outputs in the batch:

$$\mathcal{L}'_{\text{sim}} = \text{sim}(\mathbf{z}_{x,i}, \mathbf{z}_{x,\neg i}). \quad (4)$$

We use cosine similarity for both versions.

Final Loss The overall training objective of BRAINSTORM is the combination of the standard maximum likelihood estimation (MLE) $\mathcal{L}_{\text{MLE}}$, margin loss, and similarity loss:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{CE}} + w_s\mathcal{L}_{\text{sim}} + w_m\mathcal{L}_{\text{margin}} \quad (5)$$

where w_s and w_m are hyperparameters. BRAINSTORM′ replaces $\mathcal{L}_{\text{sim}}$ by $\mathcal{L}'_{\text{sim}}$.

4.2 Baselines

4.2.1 Training-Time Baselines

MLE and MLE-LL MLE is trained on all data. It is a conditional model $p(y \mid x, i)$ that learns to generate both y^+ and $y^\sim$ depending on i. MLE-LL learns to generate less likely outputs $y^\sim$ by only training on $(x, y^\sim)$. Both models are trained with standard MLE.

QUARK (Lu et al., 2022) is a state-of-the-art controllable text generation method that outperforms methods such as unlikelihood training (Welleck et al., 2020). QUARK trains an LM to generate text with fewer undesirable properties by maximizing rewards assigned by a reward function. In this study, we use the DeBERTa model (He et al., 2020) as the reward function to help generate more $y^\sim$ (more details in Section 6).

4.2.2 Decoding-Time Baselines

Modified DEXPERTS DEXPERTS (Liu et al., 2021) combines a base LM M along with two language models called "expert" (M_{exp}) and "anti-expert" (M_{anti}) that model text with desired and undesired properties, respectively. The next token distribution is determined by $P_{\text{DExperts}}(y_t) = \sigma(z'_t + \alpha(z_t^{\text{exp}} - z_t^{\text{anti}}))$ where z is the logits for the next token y_t and z'_t is the truncated logits from M under any truncation sampling methods such as top-k sampling. For simplicity, we omit the preceding context in the notation. The hyperparameter α controls how far the final token distribution deviates from model M.

In our setting, we modify this definition to be

$$P_{\text{DExperts}'}(y_t) = \sigma(z_t^\sim + \alpha(z_t^{\text{neu}} - z_t^+)) \quad (6)$$

Here, z_t^+ is from the model that learns to generate $\hat{y}^+$ by only training on (x, y^+) pairs. z_t^{neu} is from the model that learns to generate both y^+ and $y^\sim$ conditioned on the indicator. Unlike MLE, this model does not condition on indicators to generate hypotheses. Instead, it leverages text with both desired (generating $y^\sim$) and undesired properties (generating y^+). It is shown to effectively maintain

Methods	Data			Need Clf.
	+	∼	pair	
Training-time Method				
MLE-LL		✓		
MLE			✓	
QUARK	✓	✓	✓	✓
BRAINSTORM			✓	
BRAINSTORM′	✓	✓		
Decoding-time Method				
DEXPERTS			✓	
CD			✓	

Table 1: Requirements for various methods. +/∼/pair means a method requires y^+/$y^\sim$/both for x. QUARK can take any type of data as inputs but requires a trained classifier. We use BRAINSTORM′ as an alternative of BRAINSTORM if y^+ and $y^\sim$ are not both available for x. DEXPERTS and CD require that both y^+ and $y^\sim$ *could be* available for x (which is not the case for MRIINTERPRET, Section 7).

the fluency of the generated text (Liu et al., 2021). $z_t^\sim$ is from a base LM that generates $y^\sim$ only. It can be MLE-LL or BRAINSTORM.

Modified Contrastive Decoding Contrastive Decoding (CD) combines a larger M_{exp} and a smaller "amateur" model (M_{ama}) and searches for text under a constrained search space (Li et al., 2022). The resulting outputs are intended to amplify the strengths of M_{exp} and remove undesired properties that appear in M_{ama}. A scaling factor τ_{CD} controls the penalties of the amateur model in CD.

In our setting, two models have the same size. M_{ama} learns to generate y^+; M_{exp} can be MLE-LL or BRAINSTORM. Intuitively, the ability to generate $y^\sim$ is preserved, while the tendency to generate y^+ is factored out.

Hyperparameters We experiment with a wide range of values for α in DEXPERTS and τ_{CD} in CD and show how the fraction changes across these values in Figure 3. We keep the recommended value for the remaining hyperparameters. Unless specified otherwise, we generate outputs using diverse beam search (Vijayakumar et al., 2016).

5 Experimental Settings

We investigate our methods in both brain MRI settings and everyday commonsense reasoning settings (Table 5).

5.1 Everyday Commonsense Reasoning

Two datasets from the commonsense reasoning domain were adapted. See examples in Figure 4 from Appendix.

ART (*A*bductive *R*easoning in narrative *T*ext; Bhagavatula et al. (2020)) is a large-scale benchmark dataset that tests models' language-based abductive reasoning skills over narrative contexts. Each instance in the dataset consists of two observations O_1 and O_2 (O_1 happened before O_2), as well as a likely and a less likely hypothesis event (happening in between O_1 and O_2) collected from crowd workers. Each "likely" hypothesis is causally related to two observations and each "less likely" hypothesis is created by editing each "likely" hypothesis. The original task is to generate a likely hypothesis given the observation pair (O_1, O_2).

E-CARE (Explainable CAusal REasoning; Du et al. (2022)) tests models' causal reasoning skills. Each instance in the dataset consists of a premise, a "likely" and a "less likely" hypothesis, and a conceptual explanation of the causality. The likely hypothesis can form a valid causal fact with the premise. Two tasks are introduced: (1) causal reasoning: choosing the "likely" hypothesis given a premise and (2) explanation generation: generating an explanation for the causal fact.

Adapted Setting In our adapted setting, we want a model F to generate $y^\sim$ given either an observation pair (ART) or a premise (E-CARE) x. Formally, let E be a binary evaluator $E(x, y) \in \{1, 0\}$ that classifies an output y into either y^+ or $y^\sim$ based on x. We want a model F that generates $\hat{y} = F(x, i = \sim)$, where $E(x, \hat{y}) = 0$.

Evaluation For ART, we use the default training, validation and test sets to evaluate our models. For E-CARE, we randomly construct training and validation sets from the original training set and use the default validation set as the test set since the original test set is not available. All hyperparameters are determined on the validation set.

For each instance x in the test set, we ask a model F to generate $\hat{y} = F(x, i = \sim)$, then measure the fraction of less likely hypotheses according to an evaluator E.

To reduce ambiguity and encourage more consistent human evaluations, we formally define all relevancy categories from rounds of pilot studies. More detailed definitions and annotation instructions can be found in Appendix B and C. We measure both

the (1) *relevancy* and (2) *fluency* of generated hypothesis in human evaluation.

5.2 MRIINTERPRET

We present a new dataset MRIINTERPRET based on the findings and impression sections of a set of de-identified radiology reports we collected from brain MRIs. Each instance consists of a findings x, an indicator i, and a likely/less likely interpretation y of the findings x depending on i.

Dataset Construction We first find phrases such as "likely represents", "consistent with", and "may be unrelated to" that represent uncertainty from each sentence of reports. We view these phrases as indicators of the presence of interpretations; denote them by s^+ or $s^\sim$. A likely or less likely indicator (Appendix F) suggests a likely or less likely interpretation of a finding. For each likely indicator s^+, we treat the sub-sentence preceding s^+ concatenated with prior 6 sentences as the findings x, and the completion of the sentence following s^+ as the likely interpretation y^+ of the findings x. We include prior sentences to provide more context for reaching interpretations. For less likely indicators $s^\sim$, we treat the sub-sentence either following or preceding $s^\sim$ as the less likely interpretation of the findings depending on how $s^\sim$ is stated. An example can be found in Figure 4.

Indicator Unification We have collected a variety of indicators and decided to unify them into a minimum set for both likely and less likely indicators. More details of indicator unification can be found in Appendix F.

Evaluation To ensure the human evaluation for MRIINTERPRET to be as reliable as possible, we carefully curate a thorough annotation instruction guideline with precise definitions for all relevancy labels in Section 7 and Appendix E.

6 Evaluation on Commonsense Reasoning

6.1 Automatic Evaluation

Our first evaluation relies on automatically assessing whether system outputs are likely or less likely according to humans. We fine-tune DeBERTa models (He et al., 2020) for our automatic evaluation on two everyday commonsense datasets. They take the pair of (x, y) as input and predict whether y is a likely or less likely hypothesis. In our settings, the fine-tuned DeBERTa model achieves 85% accuracy on the test set of ART and achieves 80% on the original validation set of E-CARE.

	ART		E-CARE	
Model	**Frac** (↑)	**PPL** (↓)	**Frac** (↑)	**PPL** (↓)
MLE	54.1	42.6	54.5	80.4
MLE-LL	56.6	42.5	52.6	84.8
+ CD	59.9	49.8	**63.4**	107.3
+ DEXPERTS	56.2	51.7	57.2	108.3
BRAINSTORM	**79.4**	40.7	58.1	**69.2**
+ CD	**79.7**	50.2	**67.2**	88.1
+ DEXPERTS	**79.0**	51.5	58.1	89.3
QUARK	**85.9**	**27.5**	**68.2**	80.8
BRAINSTORM				
$-\mathcal{L}_{\text{margin}}$	**69.3**	44.9	54.6	73.2
$-\mathcal{L}_{\text{sim}}$	58.2	52.6	53.2	83.7
BRAINSTORM′	58.3	52.0	55.1	71.2

Table 2: Performance of generating less likely hypothesis on ART test set and E-CARE validation set. For DEXPERTS and CD, we list the fractions where models reach minimum PPL. The ablation study of our proposed method is shown at the bottom.

Table 2 compares a number of methods on our commonsense reasoning datasets. We answer several questions based on these results. We perform a paired bootstrap test for each result by comparing to MLE-LL. We highlight results that are better at 0.05 level of significance.

Can we just train on $(x, y^\sim)$? Interestingly, the baseline model MLE-LL that only trained on $(x, y^\sim)$ pairs generates "likely" hypotheses approximately half of the time. This is possibly an effect of the pre-training regimen; furthermore, generating likely hypotheses may be easier and past work has shown that seq2seq models can amplify behaviors like copying that are easy to learn (Goyal et al., 2022).

Are the proposed two loss objectives effective? We see that compared to MLE-LL, our proposed BRAINSTORM method achieves substantially higher fractions of less likely hypotheses with no cost to quality in terms of perplexity. At the bottom of Table 2, we show that ablating either of the proposed loss objectives worsens performance (and note that ablating both yields MLE). BRAINSTORM′ is not as effective since it does not compare with outputs in the batch, but we can see its merits in MRIINTERPRET (Section 7).

Can decoding-time methods alleviate the problem of generating likely outputs? We explore

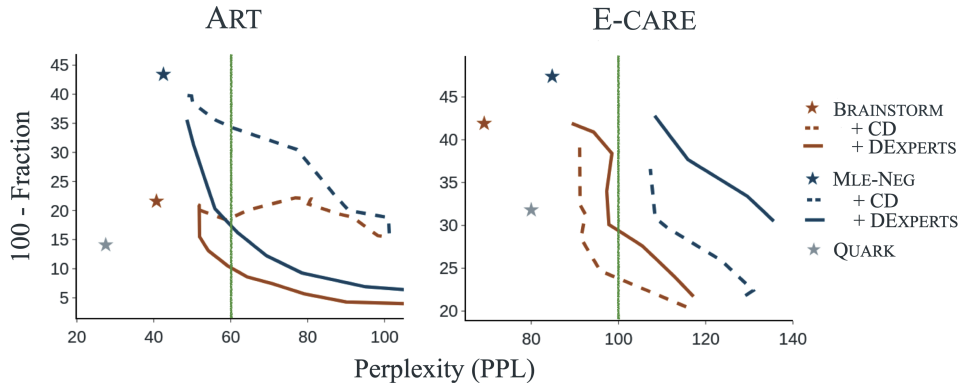

Figure 3: Fraction-perplexity trade-off of decoding-time methods CD and DEXPERTS on ART test set and original E-CARE validation set (our test set). We show the trade-off across various values for τ_{CD} in CD and α in DExperts. Both CD and DExperts can improve the fraction of less likely hypotheses, but at a very high cost to perplexity.

Model	ART					E-CARE				
	Likely (↓)	**L-Likely** (↑)	**Contra.** (?)	**Rep.** (↓)	**Irrel.** (↓)	**Likely** (↓)	**L-Likely** (↑)	**Contra.** (?)	**Rep.** (↓)	**Irrel.** (↓)
MLE-LL	42.3	15.2	22.7	9.5	10.3	35.4	15.6	5.7	18.6	24.7
Quark	**14.7**	**20.8**	51.0	**4.3**	9.2	35.2	15.1	5.7	**3.3**	40.7
BRAINSTORM	**20.9**	**20.2**	41.3	**4.8**	12.8	37.1	**20.1**	4.7	**12.7**	25.4

Table 3: Human evaluations on ART and E-CARE. We see that our method is able to produce more "less likely" (L-Likely) outputs on both datasets. We calculated the mean of the ratings from multiple annotators for each sample.

whether DEXPERTS and CD can further raise the fraction of less likely generations when combined with either MLE-LL or BRAINSTORM. These methods have hyperparameters that trade off how much of the "undesired" behavior each can remove from the system. We compute several fraction-perplexity trade-off curves in Figure 3. Notably, although the fraction of less likely outputs can improve, **both of these methods significantly increase the perplexity of generations**, which corresponds with notably worse fluency of the text. Although these points apparently have high less likely fractions, we caution that the distribution of the text may deviate from the text that DeBERTa was fine-tuned on, meaning that our classifiers may not work well in these ranges. The green lines reflect thresholds where we observe serious degradation in output quality starting to occur. Below this perplexity threshold, the automatic evaluation suggests that both methods demonstrate some capability in alleviating the models' tendency in generating "likely" hypotheses without too great a cost to perplexity. Note that DEXPERTS is more effective than CD in ART and vice versa in E-CARE.

Table 2 reports the settings where models achieve the minimum perplexities; at these points, perplexity is substantially increased but the fraction of less likely hypotheses is not substantially changed for the majority of results.

Can QUARK yield improvement? In Table 2, the automatic evaluation results show that QUARK exceeds BRAINSTORM by generating 6% more "less likely" hypothesis in ART and 10% more in E-CARE. It also has lower perplexity in ART. To further compare the two models, we conducted a human evaluation on the outputs from two models, and the result shows that QUARK generates lower-quality "less likely" hypotheses (Section 6.2).

6.2 Human Evaluation

To further validate the results, we conduct a finer-grained human evaluation on a sample of 100 examples from the test sets of both datasets along two axes – relevancy and fluency. We refined our relevancy evaluation by dividing the "relevancy" category into four subcategories, resulting in a total of five categories for evaluation.: (1) *Likely*; (2) *Less likely*; (3) *Contradictory* - the output is impossible if we assume the input is true; (4) *Repetition* - the output is describing the same meaning as the input; and (5) *Irrelevant* - the output has little connection with input. More thorough category definitions with examples, annotation instruc-

tion and quality checks for AMT annotators can be found in Appendix C. We compare the performance of three models: MLE-LL, BRAINSTORM, and QUARK (Table 3). As QUARK demonstrates better performance in automatic evaluation, we include its generated text in our human evaluation.

Our results show a high level of agreement between the automatic evaluation (Table 2) and human evaluation (Table 3) regarding the fraction of "likely" hypotheses on both datasets. On ART, QUARK and BRAINSTORM decrease the fraction of "likely" hypotheses by 60% and 50%, respectively, compared to MLE-LL. However, on E-CARE, the human evaluation indicates that all three models generate an equivalent number of "likely" hypotheses. By further breaking down the "relevancy" category used in the automatic evaluation, we then have a clearer understanding of the distribution of categories among the models' outputs.

Low-Quality Hypotheses It is not desirable for models to generate outputs that are repetitions of the input (Repetition) or have little connection to the input (Irrelevant). On the ART dataset, all models generate a small proportion of irrelevant outputs, with QUARK and BRAINSTORM reducing the fraction of "Repetition" hypotheses by half, compared to MLE-LL. However, we get more low-quality outputs on E-CARE. While BRAINSTORM is able to reduce the fraction of Repetition hypotheses by a large margin, it is not as effective as QUARK. One possible reason for this is that QUARK is trained to generate outputs that the DeBERTa classifier (the reward model) predicts as less likely; Repetition cases are rarely classified as less likely due to their similarity with the input, but Irrelevant outputs are more likely to be classified this way.

Less Likely versus Contradictory While less likely hypotheses are desirable, contradictory hypotheses are less so. A typical way of generating a contradictory hypothesis is by simply adding negation: *Lisa went laptop shopping yesterday* → *Lisa* ***didn't*** *go laptop shopping yesterday*. However, such examples have little value as the negation brings no new information to the input and is not a useful counterfactual for a user to see.

We evaluate the models' outputs on the ART dataset, where a significant number of contradictory hypotheses are generated, and find that 43 out of 100 hypotheses generated by QUARK include the words "didn't" or "not," while only 10 hypotheses generated by BRAINSTORM and MLE-LL did so. We posit that this is likely due to the DeBERTa classifier assigning high rewards for hypotheses that include negation words, and QUARK effectively learning this shortcut.

7 Human Evaluation on MRIINTERPRET

To evaluate the models' performance on the radiological interpretation generation setting, we select 30 findings from our validation set that ask for less likely interpretation. For each finding, we select the human reference and generate the top 5 less likely interpretations from 2 baselines (MLE-LL and MLE) and BRAINSTORM′, resulting in $30 \times (5 \times 3 + 1) = 480$ interpretations. We randomized the order of these interpretations before evaluation.

Due to the structure of the indicators in this dataset, methods that require examples to have both y^+ and $y^\sim$ for the same data (see "pair" in Table 1) are not able to be used. Since QUARK relies on a trained classifier, we choose not to use QUARK as well. A trained classifier on MRIINTERPRET is not reliable since the training set only consists of naturally occurring data, which is highly imbalanced (see Table 5 in Appendix). This leads the classifier to perform poorly on the "less likely" class, which is the minority class but is also the class of greatest interest in this study. We find that augmenting the training data with counterfactual cases is not easy. For example, "*the lack of evidence of restricted diffusion makes it less likely to be*" is a naturally occurring prompt from a less likely example, and attempting to change it to a sentence such as "*the lack of evidence of restricted diffusion could represent*" yields a statement that turns out to be out of distribution from the training data and models do not behave reliably in these counterfactual cases.

For each generated interpretation, we evaluate its (1) **relevancy** to the findings and (2) whether it contains any **hallucinations** about findings (Appendix E.2). For relevancy, we asked a neurologist to classify each interpretation into: (1) *Relevant and likely*; (2) *Relevant and less likely*; and (3) *Irrelevant*. Further, for those classified as "Relevant and less likely", we further evaluate how well the interpretation fits into the context of the findings by grading them on three levels: high, medium and low, ranging from high matches that represent the most obvious less likely interpretations to low matches that represent relevant but exceedingly rare diagnosis. We provide detailed definitions for these

Model	Likely	Less likely			Irrel.
		High	Med.	Low	
MLE-LL	6.7	40.7	21.2	14.7	16.7
MLE	7.3	50.0	22.1	13.3	7.3
BRAINSTORM′	6.7	42.0	**32.6**	8.7	10.0
Reference	3.3	76.7	13.4	3.3	3.3

Table 4: Human Evaluation on MRIINTERPRET. Results are shown as percentages. We evaluated 30 × 5 = 150 less likely interpretations generated from each model and 30 less likely interpretations from human reference. Results show that our proposed model successfully shifts the distribution of generated interpretations further toward the tail of the "relevant but less likely" category but still generates relevant diagnoses.

categories and include comprehensive annotation guidelines in Appendix E to facilitate consistency in future studies.

Results are shown in Table 4. Most human references (which the neurologist was blinded to) are annotated as either a high or medium match under the relevant but less likely category, suggesting the reliability of the neurologist's annotation. We find that training on all data (MLE) instead of exclusively on less likely data (MLE-LL) would effectively help generate more relevant but less likely interpretations and reduce the amount of irrelevant ones. One possible reason is that MRIINTERPRET is a highly imbalanced dataset (Table 5).

By comparing the outcomes between human reference and BRAINSTORM, we find that BRAINSTORM tends to shift the distribution of generated interpretations towards generating lower matched interpretations, which effectively extends the beam of potential diagnoses that meet the criteria of "relevant but less likely" based on refuting findings. Anecdotally, interpretations in this medium category reflect the sort of alternative hypotheses and "outside-the-box" suggestions that represent the original goal of our approach.

8 Conclusion

In this work, we propose a new text generation task "less likely brainstorming" for reducing cognitive errors in interpreting findings of MRI reports. We found that simply training on less likely data does not help with generating less likely interpretations and hence propose a novel CL method to tackle the problem. In two settings, we show that our proposed training technique can effectively generate more "less likely" hypotheses, producing interpretations that radiologists may not think of, outperforming past training- and decode-time modifications to generation models.

Limitations

Our brain MRI interpretations were evaluated by a single neurologist. Such annotations require deep expertise and are not easily carried out with high quality by trainees, which limited the amount of data we were able to collect. To ensure that the annotation would be as reliable as possible, we carefully thought of the dimensions in evaluating the generated interpretations and proposed a thorough annotation instruction guideline. We believe that future work can conduct more extensive studies using our annotation guidelines as a starting point. Further, the radiology reports we experiment with are from a single academic medical center, which makes the generalizability unclear. Future work is needed to evaluate the performance of our models on data from different medical centers. Finally, future work is needed to evaluate relevant and likely outputs from MRI interpretations to address different forms of interpretation bias and to expand the beam of potential likely diagnoses based on the findings.

Beyond the brain MRI interpretation experiments, our generation experiments are limited to a set of pre-trained models optimized for carrying out generation tasks in English. It is possible that multilingual models generating in languages other than English will show different properties. We are limited by the availability of resources for automatic evaluation in these settings, but a more extensive multilingual evaluation with human users could be conducted in the future.

Ethical Risks

We are proposing better ways for incorporating systems into the radiological diagnostic process. This is aimed at helping improve human decision-making and mitigating the limitations of traditional fully-automatic approaches. However, we believe that it is imperative to rigorously test and evaluate these methods before they can be put into practical clinical settings. We are not claiming that these methods are ready for real-world adoption at this stage.

Acknowledgments

We would like to thank Darcey Riley and TAUR lab at UT for discussion about DExperts and for providing feedback on this work. We acknowledge the funding support from National Science Foundation AI Center Institute for Foundations of Machine Learning (IFML) at University of Texas at Austin (NSF 2019844), as well as NSF CAREER Award IIS-2145280 and IIS-2145640, National Library of Medicine under Award No. 4R00LM013001, and a gift from Salesforce, Inc.

References

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. Cont: Contrastive neural text generation. abs/2205.14690.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. 2015. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35(6):1668–1676.

Lindsay P. Busby, Jesse L. Courtier, and Christine M. Glastonbury. 2018. Bias in radiology: The how and why of misses and misinterpretations. *RadioGraphics*, 38(1):236–247.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pat Croskerry. 2013. From mindless to mindful practice — cognitive bias and clinical decision making. *New England Journal of Medicine*, 368(26):2445–2448.

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

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

David M. Eddy. 1984. Variations in physician practice: The role of uncertainty. *Health Affairs*, 3(2):74–89.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

J M Farnan, J K Johnson, D O Meltzer, H J Humphrey, and V M Arora. 2008. Resident uncertainty in clinical decision making and impact on patient care: a qualitative study. *Quality and Safety in Health Care*, 17(2):122–126.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Ajay Jaiswal, Liyan Tang, Meheli Ghosh, Justin F. Rousseau, Yifan Peng, and Ying Ding. 2021. Radbert-cl: Factually-aware contrastive learning for radiology report classification. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 196–208. PMLR.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Kangmoon Kim and Young-Mee Lee. 2018. Understanding uncertainty in medicine: concepts and implications in medical education. *Korean Journal of Medical Education*, 30(3):181–188.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization.

Zhiyu Lin and Mark Riedl. 2021. Plug-and-blend: A framework for controllable story generation with blended control codes. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 62–71, Virtual. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Ximing Lu, Sean Welleck, Liwei Jiang, Jack Hessel, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. *ArXiv*, abs/2205.13636.

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics.

Yusuke Mori, Hiroaki Yamane, Ryohei Shimizu, and Tatsuya Harada. 2022. Plug-and-play controller for story completion: A pilot study toward emotion-aware story writing assistance. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 46–57, Dublin, Ireland. Association for Computational Linguistics.

Omer Onder, Yasin Yarasir, Aynur Azizova, Gamze Durhan, Mehmet Ruhi Onur, and Orhan Macit Ariyurek. 2021. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights into Imaging*, 12(1).

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jarrel C Y Seah, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M Jones. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8):e496–e506.

Herbert A. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. 2017. Interpretive error in radiology. *American Journal of Roentgenology*, 208(4):739–749.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

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.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model.

Hanqing Zhang and Dawei Song. 2022. Discup: Discriminator cooperative unlikelihood prompt tuning for controllable text generation. In *The 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi.

A Dataset statistics

Dataset statistics can be found in Table 5.

B Definition of Relevancy Categories on Everyday Commonsense

To encourage more consistent human evaluations, we formally define all relevancy categories as the following. These definitions are refined from rounds of pilot studies to reduce ambiguity for human annotations. Example outputs and explanations for each relevancy category can be found in the annotation interface (Figure 5 and 7).

B.1 E-CARE

Relevant A hypothesis is relevant if it fits with the same scenario as the premise. It should not introduce new people, places, or things that are not at least plausibly in the same source scenario.

Likely For the hypothesis to be likely, it must also be causally related to the premise – either the premise causes the hypothesis or the hypothesis causes the premise (you will see both versions of the task below). There should not be clearly more likely hypotheses than it.

Relevant and Less likely The hypothesis is still the same scenario as the premise (relevant). However, it is less likely to be causally related to the premise. There could be other hypotheses that are superior to the given hypothesis.

Irrelevant The generated hypothesis does not describe the same scenario as the premise or is not causally related to the premise.

Contradictory The hypothesis contradicts the premise – it says something that is impossible if we assume the premise to be true (e.g., the premise states that something happened and the hypothesis states that that thing did not happen).

Repetition The hypothesis is very similar to the premise – it either contains a text span that is a repetition of the premise, or it is expressing nearly the same meaning as the premise.

B.2 ART

Relevant A hypothesis is relevant if it fits with the same scenario as the observation pair. It should not introduce new people, places, or things that are not at least plausibly in the same source scenario.

Likely For the hypothesis to be likely, it must also be strongly related to O_1 and O_2 in a causal fashion – to the extent possible, the first observation O_1 should cause the hypothesis and the hypothesis causes the second observation O_2. There should not be clearly more likely hypotheses than it.

Relevant and Less likely The hypothesis is still the same scenario as the observation pair (relevant). However, it is less likely to be causally related to the observation pair – maybe it could happen following O_1, but not necessarily. There could be other hypotheses that are superior to the given hypothesis.

Irrelevant The hypothesis does not describe the same scenario as the observation pair: it either involves different people, places, or things, or the events it describes have very little connection to O_1 and O_2.

Contradictory The hypothesis contradicts either observation O_1 or observation O_2 – it says something that is impossible if we assume O_1 and O_2 to be true (e.g., O_2 states that something happened and the hypothesis states that that thing did not happen).

Repetition The hypothesis is very similar to either O_1 or O_2 – it either contains a text span that is a repetition of O_1 or O_2, or it is expressing nearly the same meaning as O_1 or O_2.

C Annotation on Everyday Commonsense

The human evaluation by crowdworkers has been judged to be IRB exempt. We hired crowd annotators from US through Amazon Mechanical Turk. These annotators have lifetime approval rates over 99% and more than 1000 approved HITs. We first conducted a quality check on ART and E-CARE. For each dataset, we randomly selected 100 examples from the test set and each example is evaluated by 7 annotators, resulting in $100 \times 7 = 700$ annotations for each dataset. We finally selected 7 qualified crowdworkers from each of the datasets. The procedure of filtering out non-qualified workers is shown below. For qualified crowdworkers, we randomly select another 100 examples from each dataset and conduct a final annotation round, resulting in $100 \times 7 \times 2 = 1400$ annotations in total. We set maximum time on completing each HIT to 1 hour and each HIT takes approximately 1.5 minutes. We paid annotators $0.3/HIT, which

Dataset	Train				Val		Test	
	Likely		Less Likely		Less Likely		Less Likely	
MRIINTERPRET	10097		1005		121		—	
ART	50509		50509		1781		3562	
E-CARE	cause	effect	cause	effect	cause	effect	cause	effect
	6855	6580	6855	6580	762	731	1088	1044

Table 5: A summary of dataset statistics. All datasets are in English. For ART and E-CARE, we show the stats of our adapted versions. Since E-CARE has a hidden test set, we randomly split the original training set into a training and a validation set, and we use the original validation set as our test set. Note that each example in E-CARE asks for either the cause or the effect of the premise.

is equivalent to $12/hr and is higher than the minimum USA wage.

Category definitions and annotation instructions with examples are shown in Figure 5, 6, 7 and 8.

Selecting Qualified Workers After we collected all annotations from the pilot study. We filter out workers by following these steps:

1. We first filter out workers that annotated less than 4 HITs. With limited amount of annotated HITs, it is hard to evaluate the consistency of their annotations.

2. For any HIT, if two output sequences are exactly the same but the annotator assigned them different categories, then we remove the worker. For example, in E-CARE, if the premise is "*Tom goes to the gym every day.*" and we have the hypotheses "*He gets a promotion from his manager who saw him in the gym.*" that appears twice, then if one hypothesis is classified as "Relevant and Likely" and another one is classified as "Relevant but Less Likely", we will filter out this annotator.

3. We use the "Repetition" category to further filter out annotators. We believe "Repetition" is the least subjective category in our annotation instruction, and using this category to filter out annotations would lead to minimum bias we can project to the selected annotators. This consists of two steps: (1) A model many generate an output that is exactly the input. For example, a model takes as input "*Tom goes to the gym every day.*" and generate "*Tom goes to the gym every day.*" as well. This happens sometimes across all models. For those cases, we will filter out annotators that assigned categories other than "Repetition"; (2) Besides the exact match, there are cases where a model's output is a paraphrase of the input. For these, to minimize our bias, we choose to use models' outputs that only differs from the input by at most two words to filter out annotators. For example, in ART, if one observation is "*Lisa went laptop shopping yesterday*", and the model's output is "*She went laptop shopping yesterday*", then we filter out annotators that do not assign "Repetition" to it.

After we collected all the annotations from qualified workers, we use the above steps to further filter out works that do not meet our standard. Finally, we got valid annotations by three annotators from each datasets. We use Fleiss kappa to calculate the agreement between annotators. The annotators achieve moderate agreement ($\kappa = 0.447$) on ART and fair agreement ($\kappa = 0.354$) on E-CARE for relevancy evaluation. This is within our expectation since evaluating whether a hypothesis is likely or less likely is subjective.

D Fluency Evaluation on Everyday Commonsense Reasoning

Fluency evaluation can be found in Table 6. Most of generations from models are fluent and grammatically correct.

E Annotation on Brain MRI Interpretation

The use of the brain MRI data is covered by an IRB. A neurologist reviewed each finding sample and evaluated the interpretation on multiple metrics.

E.1 Relevancy

The overall objective of the interpretation generation was to produce less likely diagnoses, or interpretations, based on the absence of specific findings. The findings followed a common pattern of "Absence of [finding x] makes it unlikely to

Model	ART		E-CARE	
	Gram. Correct Fluent	**Contain Flu. Errors**	**Gram. Correct Fluent**	**Contain Flu. Errors**
MLE-LL	93.9	6.1	99.0	1.0
QUARK	94.6	5.4	98.0	2.0
BRAINSTORM	93.5	6.6	95.9	4.1

Table 6: Human evaluation of fluency on everyday commonsense reasoning datasets. Annotators reached substantial agreement on both datasets.

Task	Examples	Output	Explanation
Brain MRI	…… Absence of evidence of restricted diffusion makes it unlikely to be 	~ acute ischemia ~ infarct	In diffusion weighted imaging sequences on MRI of the brain, one of the most common causes of diffusion restriction finding is due to acute ischemic stroke, also known as an infarct. Thus, the absence of restricted diffusion within brain tissue makes an interpretation unlikely to be acute ischemia/infarct.
ANLG	O1: Lisa went laptop shopping yesterday. O2: She was thankful she bought it. 	+ The price raised on the next day. ~ Lisa decided to buy a car.	This scenario makes sense —Lisa was grateful that she had made her laptop purchase the day before, as the price un-expectedly increased the following day. The event focus on Lisa's purchase of a laptop. It is not likely that she would suddenly decide to buy a car without any prior indication of her interest in doing so.
E-CARE	Tom goes to the gym every day. 	+ Tom improves his physical fitness. ~ He gets a promotion from his manager who saw him in the gym.	It directly relates to Tom's daily gym attendance. Regular exercise is a common and effective method for improving physical fitness. It directly connects to the premise involving Tom's daily gym attendance. However, the connection between gym attendance and job promotion is indirect.

Figure 4: Examples from MRIINTERPRET, ART and E-CARE. The example shown in the table for E-CARE asks for a likely/less likely effect of the premise. "+"/"∼" indicates whether humans would consider the output to be likely/less likely according to the context under the Examples column. We explain why humans would consider these outputs as likely/less likely in the Explanation column (this is not in the training data).

Model	Hallucination (%)
MLE-LL	23.3
MLE	30.0
BRAINSTORM	33.3
Reference	6.6

Table 7: Human evaluation on hallucinations. The result shows the percentage of hallucinations found in 150 generated interpretations from each model.

be [interpretation y]." The finding of interest was modified to be standardized across all findings if it used varying terminologies in a similar pattern (see Appendix F for more details). Because the interpretations are oriented in this negated valence, the objective of the output is to produce "relevant but unlikely" interpretations. The annotator rated the interpretation on 3 metrics: (1) relevant and likely, (2) relevant but less likely, and (3) irrelevant.

Relevant and Likely Output was judged as "relevant and likely" if the interpretation erroneously suggested a diagnosis that would be likely, not unlikely, despite the absence of [finding x]. For instance, "Absence of restricted diffusion within the previously described fluid collections along the right convexity makes it unlikely to be". An interpretation of "the presence of a small subdural hematoma" is actually a likely diagnosis given the lack of restricted diffusion in the fluid collection since subdural hematomas do not normally demonstrate restricted diffusion.

Relevant but Less Likely Output was judged as "relevant but less likely" if the interpretation correctly provides a less likely diagnosis due to the absence of [finding x]. For example, "absence of restricted diffusion makes it unlikely to be". An interpretation of "acute ischemia" is unlikely since diffusion restriction is often associated with acute ischemia.

If the interpretation was judged as "relevant but unlikely", the degree to which the interpretation fits with the findings was graded on three levels: (1) high, (2) medium, and (3) low.

- Less likely interpretations were **high matches** if they were within the top 5 diagnoses to fit the statement. These were the most obvious interpretations.
- Less likely interpretations were **medium matches** if they were further down the bar of potential interpretations. They still were relevant to the findings and made sense as being less likely given the absence of the finding of interest, but are less obvious and fall outside of the top 5 diagnoses.
- Less likely interpretations were **low matches** if the interpretation was relevant to the findings, but was an exceedingly rare diagnosis to make it of low value to mention as an interpretation.

Irrelevant Output was judged as "irrelevant" if it was not related to the finding of interest or the structure that the finding of interest is referring to.

E.2 Presence of Hallucination

Lastly, no matter the rating of relevance, presence or absence of hallucination was noted. It was possible to have a relevant but unlikely interpretation with high degree of fit with the finding, but a hallucination that does not appear in the original findings was added. We therefore evaluate whether each interpretation contains hallucinations.

The results are shown in Table 7. The models listed contain a large proportion of hallucinated content especially for MLE and BRAINSTORM. We examined what these hallucinations look like. We found that in the most cases, models hallucinate about the findings (generating some findings that do not actually written in the report) and concatenate those hallucinated findings after their interpretations. For examples, a generated interpretation would be "an acute infarction *although this is limited by the presence of contrast enhancement*", "intracranial abscess *although this is limited by the presence of significant soft tissue swelling*", or "blood products in the ventricular system *as seen on prior CT*."

However, unlike other text generation tasks such as text summarization where hallucinations are hard to identify, hallucinations in MRIINTERPRET follow a pattern of interpretation followed by the non-existent findings. Although future work could work on how to directly generate interpretations without hallucination, a rule-based heuristics can remove the majority of hallucinations in the current version of our system.

O1: Riley went to the store with her mother.
O2: Riley wore her cowboy boots to school the next day.

MLE-LL	Riley's mother bought her cowboy boots.
+ CD (τ_{CD} = 0.5)	Riley had bought cowboy shoes that she had not worn before.
+ CD (τ_{CD} = 1.0)	Her mother bought a new cowboy shirt for Riley.
+ CD (τ_{CD} = 1.5)	Riiley got her new cowboy boots torn.
+ DEXPERTS (α = 0.1)	Riley's mother bought her cowboy boots.
+ DEXPERTS (α = 0.5)	Riley's mother bought her cowboy boots.
+ DEXPERTS (α = 1.0)	Riley didn't have cowboy boots.
BRAINSTORM	Riley bought a pair of cowboy boots for her mother.
+ CD (τ_{CD} = 0.5)	Rice bought cowboy hats with her money and jeans.
+ CD (τ_{CD} = 1.0)	Rice's mother bought her a pair of cowboy boots to wear to school next to her cowboy outfit.
+ CD (τ_{CD} = 1.5)	Her mom bought her a new pair of cowboy boots.
+ DEXPERTS (α = 0.1)	Riley's mother bought her a pair of cowboy boots. Riley hated them.
+ DEXPERTS (α = 0.5)	Riley bought a pair of cowboy pants.
+ DEXPERTS (α = 1.0)	Riley bought a pair of cowboy pants.
QUARK	Riley's mother bought her cowboy boots at the store and Riley hated them.

Table 8: Examples of generating less likely hypotheses across different methods on ART.

Premise: The whole kitchen was burnt down more quickly.
Ask for: less likely cause

MLE-LL	The whole kitchen was burnt down more quickly.
+ CD (τ_{CD} = 0.5)	Tom poured water in the pan of acid.
+ CD (τ_{CD} = 1.0)	There was a big thunderstorm yesterday.
+ CD (τ_{CD} = 1.5)	There is a big storm outside.
+ DEXPERTS (α = 0.1)	The whole kitchen was burnt down more quickly.
+ DEXPERTS (α = 0.5)	The fire in the kitchen spread outwards.
+ DEXPERTS (α = 1.0)	There are more and more fires in this place.
BRAINSTORM	Tom put a lot of fuel on the fire.
+ CD (τ_{CD} = 0.5)	Tom poured a bucket of water to a sink which has a high temperature.
+ CD (τ_{CD} = 1.0)	There was an accident at night.
+ CD (τ_{CD} = 1.5)	Tom poured gasoline to the stove.
+ DEXPERTS (α = 0.1)	There is a fire in the kitchen.
+ DEXPERTS (α = 0.5)	The whole kitchen was filled with smoke.
+ DEXPERTS (α = 1.0)	Tom's kitchen is leaking water.
QUARK	The fire in the kitchen was very hot.

Table 9: Examples of generating less likely hypotheses across different methods on E-CARE.

F Indicator Unification for MRIINTERPRET

We narrowed down the indicators to a smaller set to ensure that our model sees sufficient data for each indicator during training. The indicator mappings are shown in Figure 9 and 10. We also include the way we flip these indicators for the margin loss objective.

G Example of generated outputs

We show examples of generated outputs for both everyday commonsense reasoning datasets in Table 8 and 9.

H Implementation Details

H.1 Significance Test

We perform a paired bootstrap test for each result by comparing to MLE-LL. We highlight results that are better at 0.05 level of significance.

H.2 Computing Infrastructure

We use BART from HuggingFace Transformers (Wolf et al., 2020), which is implemented in the PyTorch framework.

H.3 Training Details

We fine-tune BART-Large (400M parameters) with 1 NVIDIA RTX A6000 GPU on all experiments and it converges in 2 epochs. We use AdamW as our optimizer with adam epsilon set to 1e-8. Learning rate is set to 5e-5 with linear schedule warmup. There is no warm-up step.

H.3.1 Everyday Commomsense Reasoning

We initialize the model from `facebook/bart-large`. The batch size is set to 64 if only using MLE objective and 42 otherwise. We set maximum input length to 100 and maximum output length to 64. Most text should fit into these lengths. The average training time for each model is around 0.8 GPU hours if only using MLE objective and 1.5 GPU hours otherwise.

H.3.2 MRIINTERPRET

We initialize the model from `GanjinZero/biobart-large` (Yuan et al., 2022). The batch size is set to 32. We set maximum input length to 256 and maximum output length to 60. Most text should fit into these lengths. The average training time for each model is around 0.8 GPU hours if only using MLE objective and 1.2 GPU hours otherwise.

H.4 Hyperparameter Setups

BRAINSTORM For the margin loss $\mathcal{L}_{\text{margin}}$ (Equation (2)), we chose m within in the range of 1×10^{-3} and 1×10^{-2} and set it to 0.005 in the log space as it works well throughout our experiments. w_s and w_m are set to 1.0 and 10.0, respectively, as they achieve the best result on the validation set.

QUARK We follows the default parameter setups in the original work with 6000 training steps for both commonsense reasoning datasets.

Decoding We use diverse beam search for all experiments with diversity penalty set to 1.0. We set τ_{CD} in CD from 2×10^{-1} to 1×10^{3}, and α in DEXPERTS from 1×10^{-3} to 1. We keep the recommended values for the remaining hyperparameters.

Instructions

In this HIT, you will be presented with an observation pair (Observation 1, Observation 2) on the left. These are two events that we assume have happened. Then, you are presented with multiple hypotheses on the right that could have happened **between** O1 and O2. Your job is to evaluate the quality of the hypothesis along two axes – **Relevancy** and **Fluency**.

Relevancy
Classify the hypothesis into one of the following categories:

I. Relevant and likely
II. Relevant but less likely
III. Irrelevant
IV. Contradictory
V. Repetition

Relevant
A hypothesis is *relevant* if it fits with the same scenario as the observation pair. It should not introduce new people, places, or things that are not at least plausibly in the same source scenario.

Likely
For the hypothesis to be likely, it must also be strongly related to O1 and O2 in a **causal** fashion – to the extent possible, the first observation O1 should cause the hypothesis and the hypothesis causes the second observation O2. There should not be clearly more likely hypotheses than it.

Relevant and Less likely
The hypothesis is still the same scenario as the observation pair (*relevant*). However, it is **less likely** to be causally related to the observation pair – maybe it could happen following O1, but not necessarily. There could be other hypotheses that are superior to the given hypothesis.

Irrelevant
The hypothesis does not describe the same scenario as the observation pair: it either involves different people, places, or things, or the events it describes **have very little connection** to O1 and O2.

Contradictory
The hypothesis contradicts either observation O1 or observation O2 – it **says something that is impossible if we assume O1 and O2 to be true** (e.g., O2 states that something happened and the hypothesis states that that thing did not happen).

Repetition
The hypothesis is very similar to either O1 or O2 – it either contains a text span that is a **repetition** of O1 or O2, or it is **expressing nearly the same meaning** as O1 or O2.

Fluency
Classify the **hypothesis** into one of the following categories:

I. Contains fluency errors
II. Grammatically correct and fluent

Note: Please **only** evaluate the fluency of the hypothesis as a **standalone** piece of text. That is, evaluate if that one sentence looks okay to you, as opposed to whether or not it makes sense in context.

Examples

Example 1
O1: Baby Jake needed a bath because he had not bathed in two days.
O2: Jake's mom finished by taking him out and drying him with the towel.

Hypothesis 1: Jake's mom gave him a bath and then he fell asleep.
Hypothesis 2: Jake's mom gave him a bath and he loved it.
Hypothesis 3: Jake's mom didn't bathe him and didn't give him a bath.
Hypothesis 4: Jake's mom didn't bathe him.
Hypothesis 5: Jake's mom cooked him a delicious meal.
Hypothesis 6: Jake's dad put him in the bath tub.

Relevancy

- Relevant and likely
  Hypothesis 1, 2. These all involve Jake and his mom (relevant) and make sense in the scenario.
- Relevant but less likely
  Hypothesis 6. This hypothesis could happen. However, “Jake's mom put him in the bath tub” would be a more likely option given the observation O2.
- Contradiction
  Hypothesis 3, 4. These hypotheses are contradictory to the Observation O2. Observation O2 implies that Jake took a bath, but Hypothesis 3, 4 directly say that Jake did not take a bath, which is a contradiction.
- Irrelevant
  Hypothesis 5. There is no connection between the observation pair and the hypothesis, although they involve the same people.

Fluency

- Contains fluency errors
  Hypothesis 3. The phrase "didn't bathe him" and "didn't give him a bath" both refer to the same action (not bathing Jake), so using both phrases in the same sentence is redundant and can be confusing.
- Grammatically correct and fluent
  Hypothesis 1, 2, 4, 5, 6. These hypotheses alone are grammatically correct and convey a clear and logical message.

Example 2
O1: Janice looked at her bags of trash with pride.
O2: Instead, she took a friend and they both got small yogurts together.

Hypothesis 1: Janice was going to buy herself yogurt.
Hypothesis 2: Janice decided to throw away all the trash in the trash.
Hypothesis 3: Janice wanted to go to the store and buy a large bag of trash.
Hypothesis 4: Janice looked at her bags of trash with pride.
Hypothesis 5: Janice was proud of trash that she had collected.

Relevancy

- Relevant and likely
  Hypothesis 1, 2.
- Relevant and less likely
  Hypothesis 3. It is unusual to buy a bag of trash from a store.
- Repetition
  Hypothesis 4, 5. Hypothesis 4 is a repetition of Observation O1. Hypothesis 5 is expressing the same meaning as Observation O1.

Fluency

- Contains fluency errors
  Hypothesis 2. "All the trash in the trash" could be interpreted as redundant because "trash" is already included in "all the trash." It might sound clearer to say "all the trash" or "all the trash in the bin." Both of these phrases would eliminate the potential for redundancy and make the meaning of the sentence more clear.
- Grammatically correct and fluent
  Hypothesis 1, 3, 4, 5. These hypotheses alone are grammatically correct and convey a clear and logical message.

Figure 5: Annotation Interface (I) for ART.

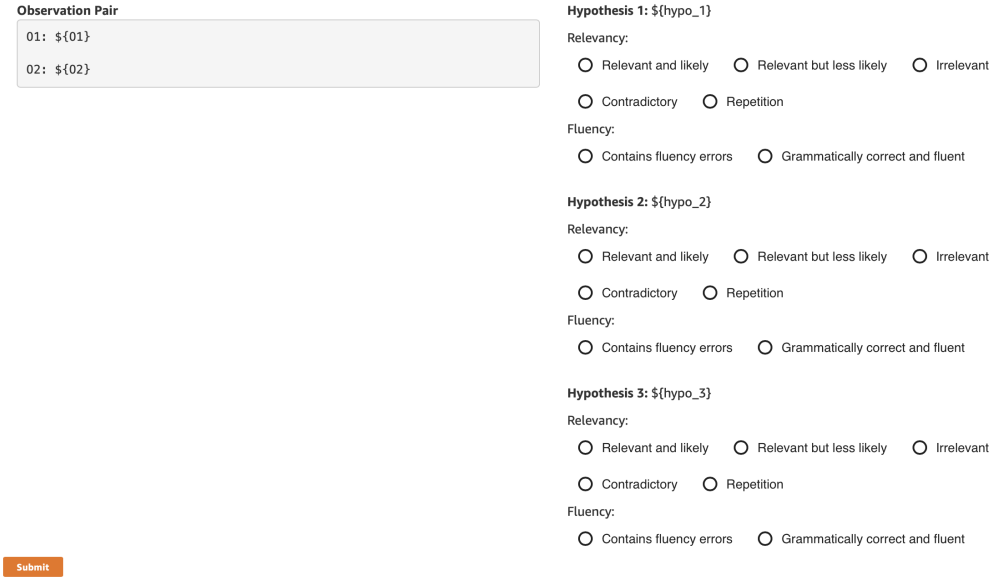

Figure 6: Annotation Interface (II) for ART.

Instructions

In this HIT, you will be presented with a premise statement introducing a scenario, followed by multiple hypotheses statements. These hypotheses statements are supposed to be either causes or effects of the premise. Your job is to evaluate the quality of each of the hypothesis statements along two axes – **Relevancy** and **Fluency**.

Note: You may search for information to verify certain hypotheses.

Relevancy
Classify the hypothesis into one of the following categories:

I. Relevant and likely
II. Relevant but less likely
III. Irrelevant
IV. Contradictory
V. Repetition

Relevant
A hypothesis is *relevant* if it fits with the same scenario as the premise. It should not introduce new people, places, or things that are not at least plausibly in the same source scenario.

Likely
For the hypothesis to be likely, it must also be **causally** related to the premise – either the premise causes the hypothesis or the hypothesis causes the premise (you will see both versions of the task below). There should not be clearly more likely hypotheses than it.

Relevant and Less likely
The hypothesis is still the same scenario as the premise (*relevant*). However, it is less likely to be **causally related** to the premise. There could be other hypotheses that are superior to the given hypothesis.

Irrelevant
The generated hypothesis does not describe the same scenario as the premise or is not causally related to the premise.

Contradictory
The hypothesis contradicts the premise – it **says something that is impossible if we assume the premise to be true** (e.g., the premise states that something happened and the hypothesis states that that thing did not happen).

Repetition
The hypothesis is very similar to the premise – it either contains a text span that is a **repetition** of the premise, or it is **expressing nearly the same meaning** as the premise.

Fluency
Classify the **hypothesis** into one of the following categories:

I. Contains fluency errors
II. Grammatically correct and fluent

Note: Please **only** evaluate the fluency of the hypothesis as a **standalone** piece of text. That is, evaluate if that one sentence looks okay to you, as opposed to whether or not it makes sense in context.

Examples

Example 1
Premise: My mom keeps cleaning my room.
What would be the possible **effect** of the premise?

Hypothesis 1: My mom cleans my room every day.
Hypothesis 2: The dust in my room is getting worse.
Hypothesis 3: My mom never cleans my room.
Hypothesis 4: It's time for her to leave for work.
Hypothesis 5: My mom keeps cleaning my room.
Hypothesis 6: My mom is constantly cleaning my room.
Hypothesis 7: My room keeps clean.

Relevancy

- Relevant and likely
  Hypothesis 7. This scenario makes sense.
- Relevant and less likely
  Hypothesis 2. It is less likely that there are more dust in the room if the room keeps being cleaned.
- Contradiction
  Hypothesis 3. The premise and the hypothesis 3 cannot both be true at the same time, so they contradict each other.
- Irrelevant
  Hypothesis 4. Hypothesis 4 is not related to the premise, although they may involve the same person.
- Repetition
  Hypothesis 1, 5, 6. Hypothesis 5 repeats the premise. Hypothesis 1, 6 are phraphrases of the premise. They express the same meaning as the premise.

Fluency

- Grammatically correct and fluent
  Hypothesis 1, 2, 3, 4, 5, 6, 7. These hypotheses alone are grammatically correct and convey a clear and logical message.

Example 2
Premise: We can see many stripes on their backs.
What would be the possible **cause** of the premise?

Hypothesis 1: There are many zebras in the zoo.
Hypothesis 2: Kudus are African animals.
Hypothesis 3: We can see many stripes on their backs.
Hypothesis 4: There are many Periwinkles in the zoo.
Hypothesis 5: There are many zebras in the zoo the zoo.

Relevancy

- Relevant and likely
  Hypothesis 1, 2, 5. For hypothesis 2, Kudus has stripes on their backs. This can be verified from online search.
- Relevant and less likely
  Hypothesis 4. Periwinkles does not commonly has stripes on their backs (from online search).
- Repetition
  Hypothesis 3.

Fluency

- Contains fluency errors
  Hypothesis 5. It is repetitive and does not make grammatical sense. A more fluent version of this sentence would be "There are many zebras in the zoo."
- Grammatically correct and fluent
  Hypothesis 1, 2, 3, 4. These hypotheses alone are grammatically correct and convey a clear and logical message.

Figure 7: Annotation Interface (I) for E-CARE.

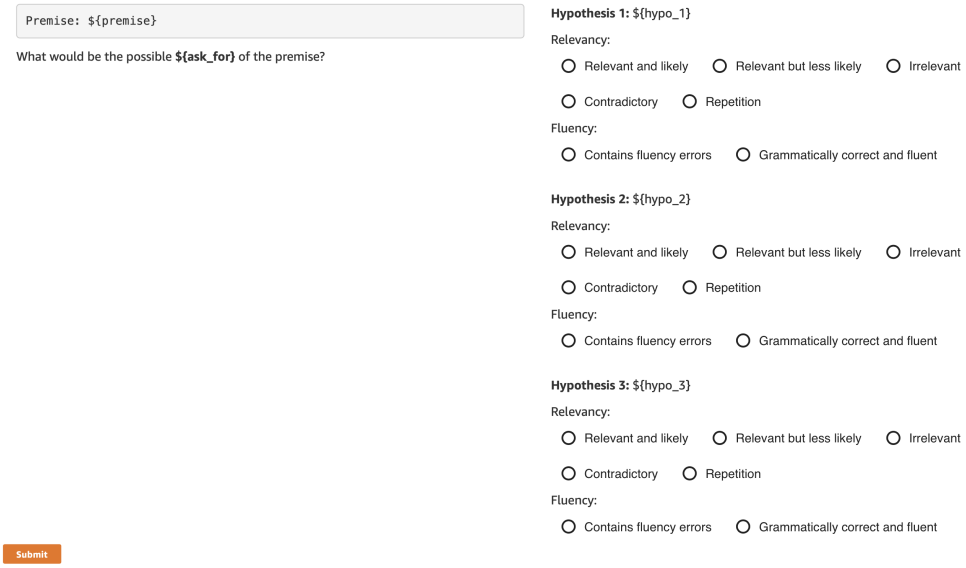

Figure 8: Annotation Interface (II) for E-CARE.

Mappings of likely indicators

likely suggestive of:
{with suggestion of, a reflection of, likely representing, likely reflective of, likely relating, suggesting, in favor of, most likely consistent with, likely consistent with, perhaps related to, possibly related to, most likely related to, raising the possibility of, most likely reflecting, likely relating to, potentially related to, likely the result of, likely reflecting, concerning for, favor of, favored to represent, most likely representing, in keeping with, to be related to, to represent, probably representing, likely due to, probably related to, likely related to, compatible with, more likely to be related to, most likely, possibly representing, most consistent with, suggestive of, potentially reflecting, consistent with, most likely to be related to, representing, potentially representing}

could represent:
{most likely to represent, most likely represent, likely represents, suggests the possibility of, is favored to represent, potentially reflect, could be an indication of, are diagnostic of, may also reflect, could indicate, likely reflects, may be seen with, potentially represent, may be seen in, can represent, likely represent, could possibly be related to, may represent, likely suggest, most likely represents, likely indicate, suggest the possibility of, may be due to, likely reflect, represents, may be a reflection of, could be related to, could reflect, most likely diagnosis is, could potentially be related to, raises possibility of, probably represent, can be seen in the setting of, most likely reflect, raise the possibility of, may reflect, can be seen in, may well represent, would have to represent, may also represent, probably also represent, may be in part related to, could be due to, may indicate, could be consistent with, could represent, likely indicates, could be a reflection of, likely suggests, could also represent, may be related to}

findings could represent:
{considerations would include, differential diagnosis would include, differential considerations include, differential includes, differential would include, diagnostic possibilities include, differential diagnosis also includes}

Figure 9: Unifying "likely" indicators in MRIINTERPRET.

Mappings of less likely indicators

findings are less likely to be:
{another less likely possibility is, less likely differential considerations include, less likely considerations would be, less likely considerations include, less likely considerations would include, less likely possibilities include, less likely possibilities would include}

less likely to be:
{less likely related to, likely not related to, likely unrelated to, not particularly characteristic of, versus less likely, not characteristic of, probably not related to, unlikely to represent}

cannot exclude:
{may be unrelated to, less likely would be, may not be related to, is not related to}

makes it unlikely to be: {makes it unlikely to be}

Flipping Unified Indicators

Likely to Less Likely
likely suggestive of -> less likely to be
could represent -> cannot exclude
findings could represent -> findings are less likely to be"

Less Likely to Likely
findings are less likely to be -> findings could represent
less likely to be -> likely suggestive of
cannot exclude -> could represent
makes it unlikely to be -> could represent

Figure 10: Unifying "less likely" indicators in MRIINTERPRET and how we map flipped indicators.