Explanation Regeneration via Information Bottleneck

Qintong Li[♠*] **Zhiyong Wu**[◇] **Lingpeng Kong**[♠] **Wei Bi**[♡]
[♠]The University of Hong Kong
[◇]Shanghai AI Laboratory [♡]Tencent AI Lab
qtli@connect.hku.hk, wuzhiyong@pjlab.org.cn, lpk@cs.hku.hk, victoriabi@tencent.com

Abstract

Explaining the black-box predictions of NLP models naturally and accurately is an important open problem in natural language generation. These free-text explanations are expected to contain sufficient and carefully-selected evidence to form supportive arguments for predictions. Thanks to the superior generative capacity of large pretrained language models (PLM), recent work built on prompt engineering enables explanations generated without specific training. However, explanations generated through single-pass prompting often lack sufficiency and conciseness, due to the prompt complexity and hallucination issues. To discard the dross and take the essence of current PLM's results, we propose to produce sufficient and concise **e**xplanations via the **i**nformation **b**ottleneck (EIB) theory. EIB regenerates explanations by polishing the single-pass output of PLM but retaining the information that supports the contents being explained by balancing two information bottleneck objectives. Experiments on two different tasks verify the effectiveness of EIB through automatic evaluation and thoroughly-conducted human evaluation.

1 Introduction

Natural language explanations have attracted a lot of attention as a way to uncover the rationales behind black-box predictions. Thanks to the power of large pretrained language models (PLM) (Brown et al., 2020; Zhang et al., 2022), prompting methods proposed in recent studies achieve impressive results in generating free-text explanations (Wei et al.; Lampinen et al., 2022). A clear advantage of such methods is that they involve no additional training from task-specific datasets.

In this paper, we regard a free-text explanation as a description of the relationship between an input context and a hypothesis, e.g., a question and an answer. Although it is difficult to state that one explanation is superior to all others due to the different desiderata of the tasks to be explained, this does not prevent us from answering the question "*what makes a good explanation*" from a practical view. Previous research (Yu et al., 2019; Miller, 2019) points out several semantic constraints should be satisfied in constructed explanations: (i) avoid undesirable content, like repeating context's statement, (ii) ensure adequate background supports, and (iii) emphasize selective evidence. Current machine-generated explanations still exhibit defects on these constraints (Kassner and Schütze, 2020; Welleck et al., 2022). For single-pass prompting methods, they cast the burden of ensuring explanation constraints all on a PLM which "starts from scratch". This inspires us to investigate how to discard the dross and take the essence of current PLM's results.

We propose our **e**xplanation generation approach via the **i**nformation **b**ottleneck theory (Tishby

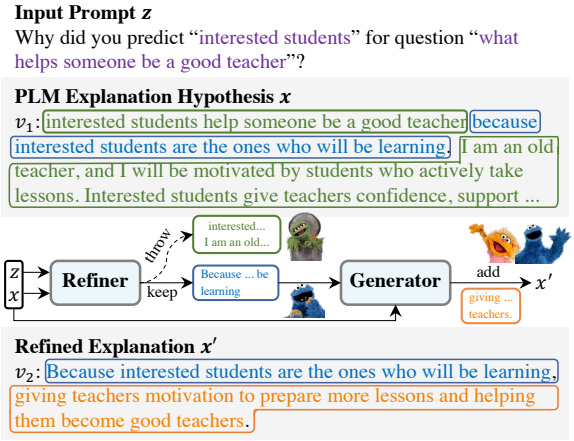


Figure 1: Although PLM generates an informative explanation hypothesis (v_1), this explanation contains redundant or inessential information which may interfere with the holistic understanding of the relationship between question and answer. In comparison, the polished explanation (v_2), improved upon the initial hypothesis, is more concise and reasonable.

[*]Work done while interning at Tencent AI Lab.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12081–12102
July 9-14, 2023

et al., 2000) (EIB), which can refine explanations prompted from PLM into more *meaningful*, *sufficient*, and *concise* ones. It works in two phases, as illustrated in Figure 1. First, given an NLP task sample (e.g., a QA pair), EIB uses a large PLM to produce an initial explanation hypothesis (v_1) by framing the task sample into a prompt input. Second, a *refiner* improves the quality of an explanation hypothesis along the axis of the aforementioned characteristics (i.e., meaningful, sufficient, and concise). The *refiner* is trained following the information bottleneck principle. Concretely, it learns a minimal sufficient bottleneck representation of the explanation v_1, while being maximally explainable about the sample (i.e., the QA pair) by introducing an information loss (Ethayarajh et al., 2022). With the learned bottleneck representation on hand, a generator learns to produce a new explanation. We propose a simple and general procedure for training the refiner by pairing synthetic explanation hypotheses with gold references from existing datasets. EIB is a general explanation generation framework and can be applied to different NLP tasks with no specific task supervision.

We demonstrate the effectiveness of EIB in generating explanations on two popular NLP tasks: commonsense question answering and natural language inference. Experiments show that EIB significantly improves the explanation candidates prompted from PLM, by making them more concise while retaining useful information for explaining task samples. Automatic evaluation and carefully designed human evaluation demonstrate the performance of EIB. Furthermore, an analysis of evaluations shows an imperious demand for better metrics to judge explanations more credibly. We publicly release our code and data[1].

2 Method

Prompting Recently, writing explanations through prompting large PLMs has become a competitive approach. Given an NLP task sample z including input z_c and output z_o, we could infer its explanation $\boldsymbol{x}$ via prompting a PLM: $\boldsymbol{x} = PLM(S(z_c, z_o))$, where function $S(\cdot,\cdot)$ transforms z to prompt formats through predefined templates. For example, if we have a QA sample, question z_c: *Can elephants be put in the fridge?* and answer z_o: *no*, the prompt will be "*The question is can elephants be put in the fridge? The answer is no* because.".

Although prompting has achieved remarkable success, machine-generated explanations still have room for improvement as discussed in the introduction. Therefore, we seek to step further under the current achievement, exploring an effective way to improve explanation quality in terms of meaningfulness, sufficiency, and conciseness.

Formulation Suppose we have a sample $z \in \mathsf{Z}$ and its explanation hypothesis $\boldsymbol{x} \in \mathsf{X}$.[2] We aim to refine $\boldsymbol{x}$ into a better $\boldsymbol{x}'$ which can: (1) reduce irrelevant information in $\boldsymbol{x}$ (conciseness), (2) preserve and supplement useful information to infer z (meaningfulness, sufficiency). We divide the explanation regeneration task into two problems: *refinement* and *generation*.

First, we model the refinement problem from an information-theoretic view, i.e., learn the internal representation t of the initial explanation x, defined as $p_\theta(\boldsymbol{t} \mid \boldsymbol{x})$, such that t is maximally compressive about the (noisy) x while being maximally expressive about z:

$$\min_{\theta} \mathrm{I}(\mathsf{x},\mathsf{t}) \text{ s.t. } \mathrm{I}(\mathsf{t},\mathsf{z}) \geq \mathrm{I}_c \,, \tag{1}$$

The above process can be formulated as the **information bottleneck principle** (IB) (Tishby and Zaslavsky; Alemi et al., 2017). IB defines the characteristics of an optimal representation, in terms of the fundamental tradeoff between having a concise representation and one with good predictive power, which is equivalent to minimizing the following objective function:

$$\mathcal{L}_{IB} = \beta \cdot \underbrace{\mathrm{I}(\mathsf{x},\mathsf{t})}_{\textit{compression}} - \underbrace{\mathrm{I}(\mathsf{t},\mathsf{z})}_{\textit{preservation}} \,. \tag{2}$$

where β is a Lagrange multiplier. A large β corresponds to high compression, and hence low mutual information between t and z.

Given a bottleneck representation $\boldsymbol{t}$, our second goal is to generate a free-text explanation $\boldsymbol{x}'$ based on $\boldsymbol{x}$. Therefore, we pack a log-likelihood objective for language modeling with $\mathcal{L}_{IB}$ as the objective function of the whole model, and train it on an automatically constructed synthetic dataset:

$$\mathcal{L}_{EIB} = \underbrace{\mathcal{L}_{IB}}_{\textit{refinement}} - \underbrace{\log p(\boldsymbol{x}' \mid \boldsymbol{t}, \boldsymbol{x}, z)}_{\textit{generation}} \,. \tag{3}$$

[1] https://github.com/qtli/EIB

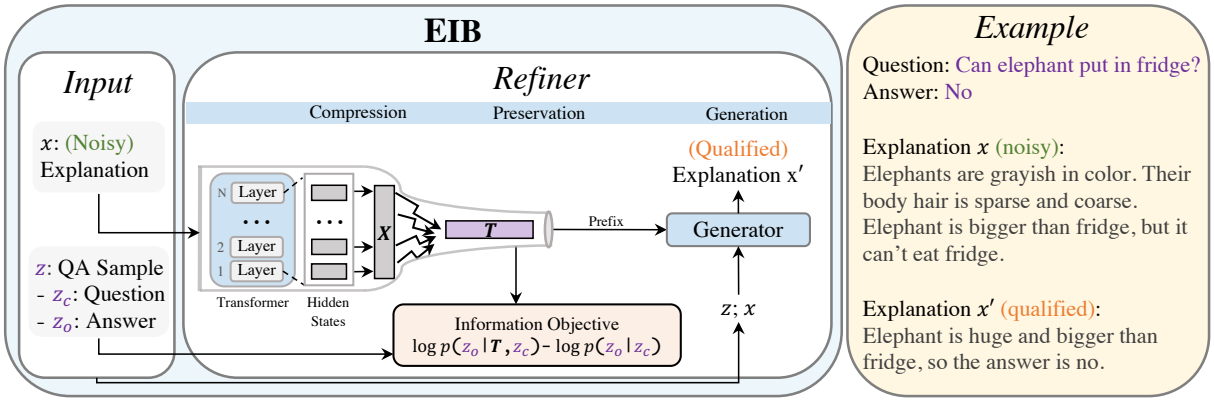


Figure 2: Illustration of our method. Given a task sample z and an explanation candidate $\boldsymbol{x}$ which may be noisy, (i) a refiner first compresses $\boldsymbol{x}$ into bottleneck vectors $\mathbf{T}$ via a tunable stochastic mapping. (ii) An information objective optimizes compression direction ensuring $\mathbf{T}$ to be predictive of z. (iii) A generator generates a sufficient and concise explanation based on the bottleneck representation $\mathbf{T}$, z, and $\boldsymbol{x}$. The right side shows an example of EIB.

The overall proposed EIB is illustrated in Figure 2.

In the following, we will present the optimization and training with respect to (i) explanation compression for distilling a bottleneck representation from the initial explanation, (ii) information preservation for ensuring the distilled bottleneck representation expressive about the explained sample, and (iii) explanation regeneration from the distilled bottleneck representation for producing a better explanation than the initial input one.

2.1 Explanation Compression

Vectorization Suppose we have an explanation candidate $\boldsymbol{x}$ that needs to be improved. We first use a parameter-fixed n-layer PLM to encode $\boldsymbol{x}$ and aggregate the hidden states of n layers into a sequence of vectors $\mathbf{X} \in \mathbb{R}^{n \times d}$, where each d-dimensional vector $\mathbf{x}_i$ is the weighted sum of hidden representations of the corresponding layer by attention weights. We utilize representations of all layers instead of the last layer only in order to combine more information.

Compression Our first goal is to denoise irrelevant information in $\mathbf{X}$ and obtain a highly compact representation $\mathbf{T}$. The compression loss part in $\mathcal{L}_{IB}$ can be rewritten as:

$$\mathrm{I}(\mathsf{x};\mathsf{t}) \overset{\text{def}}{=} \sum_{i}^{n} \mathbb{E}_{\mathbf{x}_i}[\mathbb{E}_{\mathbf{t}_i \sim p_\theta}[\log(\frac{p_\theta(\mathbf{t}_i \mid \mathbf{x}_i)}{p_\theta(\mathbf{t}_i)})]] \,, \quad (4)$$

where $p_\theta(\mathbf{t}_i)$ is the prior distribution of the bottleneck vector $\mathbf{t}_i$, $p_\theta(\mathbf{t}_i \mid \mathbf{x}_i)$ is the stochastic mapping from the distribution of initial explanation hypothesis to its intermediate compressed representation, and θ indicates learnable parameters.

Optimization Specifically, we perform a linear transformation on each vector $\mathbf{x}_i$ of $\mathbf{X}$, to produce a polished representation $\mathbf{T} = MLP(\mathbf{X}) \in \mathbb{R}^{n \times k}$. We assume each vector $\mathbf{t}_i$ of $\mathbf{T}$ follows an isotropic Gaussian distribution, where the mean and standard deviation are learnable parameters with the use of the reparameterization trick. However, for $p_\theta(\mathbf{t}_i) = \mathbb{E}_{\hat{\mathbf{x}}_i}[p_\theta(\mathbf{t}_i \mid \hat{\mathbf{x}}_i)]$, it is difficult to loop over all candidates $\hat{\mathbf{x}}_i$. We practically use a standard Gaussian distribution $p_{\mathcal{N}}(\mathbf{t}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ to simulate $p_\theta(\mathbf{t}_i)$ for simplicity. Using the fact $\mathbb{E}[KL(p_\theta(\mathbf{t}_i) \parallel p_{\mathcal{N}}(\mathbf{t}_i))] \geq 0$, we can minimize the upper bound of $\mathrm{I}(\mathsf{x};\mathsf{t})$:

$$\mathrm{I}(\mathsf{x},\mathsf{t}) \leq \sum_{i}^{n} \mathbb{E}_{\mathbf{x}_i}[\mathbb{E}_{\mathbf{t}_i \sim p_\theta}[\log(\frac{p_\theta(\mathbf{t}_i|\mathbf{x}_i)}{p_{\mathcal{N}}(\mathbf{t}_i)})]] \,. \quad (5)$$

Making the bound as tight as possible given θ allows yielding a compressed representation $\mathbf{T}$ distilled from the initial $\mathbf{X}$.

2.2 Information Preservation

The second goal of IB in Eq. 2 is to maximize $\mathrm{I}(\mathsf{t},\mathsf{z})$, which can lead to a high log-likelihood $p_\theta(\mathbf{Z} \mid \mathbf{T})$ for ensuring $\mathbf{T}$ not losing predictive features of $\mathbf{X}$ to explain $\mathbf{Z}$:

$$\mathrm{I}(\mathsf{t},\mathsf{z}) \overset{\text{def}}{=} \sum_{i}^{n} \mathbb{E}_{\mathbf{z}_i, \mathbf{t}_i \sim p_\theta}[\log(\frac{p_\theta(\mathbf{z}_i \mid \mathbf{t}_i)}{p(\mathbf{z}_i)})] \,, \quad (6)$$

$$p_\theta(\mathbf{z}_i \mid \boldsymbol{t}_i) \overset{\text{def}}{=} \sum_{i}^{n} \mathbb{E}_{\mathbf{x}_i}[\frac{p(\mathbf{z}_i \mid \mathbf{x}_i) p_\theta(\mathbf{t}_i \mid \mathbf{x}_i) p(\mathbf{x}_i)}{p_\theta(\mathbf{t}_i)}] \,. \quad (7)$$

[2] $\boldsymbol{x}$,$\boldsymbol{t}$,z and $\mathbf{X}$,$\mathbf{T}$,$\mathbf{Z}$ are instances of random variables x,t,z.

However, $p_\theta(\mathbf{z}_i \mid \mathbf{t}_i)$ is hard to estimate because we have to iterate on all possible $\boldsymbol{x}_i$. Furthermore, the length of $\boldsymbol{z}$ is not fixed and cannot be precisely aligned to the number of bottleneck vectors $\mathbf{T}$.

Optimization We extend recent work in information theory (Xu et al., 2020; Ethayarajh et al., 2022), which generalizes Shannon's information theory to quantify the predictive $\mathcal{V}$-information between two random variables, subject to computational constraints $\mathcal{V}$. $\mathcal{V}$-information reflects the ease with which $\mathcal{V}$ can predict z given t.

In this paper, we use p_ϕ to denote the computational constraints, i.e., an autoregressive model GPT-2 (Radford et al., 2019). Measuring $\mathrm{I}(\mathsf{t}, \mathsf{z})$ becomes quantifying usable information under p_ϕ. Then $\mathrm{I}(\mathsf{t}, \mathsf{z})$ can be approximated by the information difference of an unconditional entropy $H_{p_\phi}(\mathsf{z})$ and conditional entropy $H_{p_\phi}(\mathsf{z} \mid \mathsf{t})$ *w.r.t* computation-bounded parameters ϕ:

$$\mathrm{I}(\mathsf{t}, \mathsf{z}) \geq H_{p_\phi}(\mathsf{z}) - H_{p_\phi}(\mathsf{z} \mid \mathsf{t}), \tag{8}$$

$$H_{p_\phi}(\mathsf{z}) = \mathbb{E}_{\mathsf{z}}[-\log p_\phi(z)], \tag{9}$$

$$H_{p_\phi}(\mathsf{z} \mid \mathsf{t}) = \mathbb{E}_{\mathsf{z}, \mathbf{T} \sim p_\theta(\mathbf{T}|\mathbf{X})}[-\log p_\phi(z \mid \mathbf{T})], \tag{10}$$

where θ and ϕ are optimizable parameters, $\mathbf{t}$ acts as a learnable prefix (Li and Liang, 2021) to a GPT-2.

Optimizing the lower bound of $\mathrm{I}(\mathsf{t}, \mathsf{z})$

$$\mathbb{E}_{\boldsymbol{x}, \mathsf{z}}[\mathbb{E}_{\mathbf{T} \sim p_\theta(\mathbf{T}|\mathbf{X})}[\log p_\phi(z \mid \mathbf{T}) - \log p_\phi(z)]]$$

requires $\mathbf{T}$ to have enough capacity to support $\boldsymbol{z}$ while being compact with the consideration of the minimization of $\mathrm{I}(\mathsf{x}, \mathsf{t})$.

2.3 Explanation Regeneration

With the distilled bottleneck representation $\mathbf{T}$ on hand, the remaining task is to translate the compact representation into a new explanation $\boldsymbol{x}'$ that may be different from the initial explanation $\boldsymbol{x}$ while achieving obvious quality improvements.

Translating the highly-dimensional matrix $\mathbf{T}$ into a discrete and readable explanation is not an easy task. To tackle this challenge, we use the explanation datasets from various NLP tasks and build a training corpus by pairing the human-written explanation with its synthetic imperfect version, which allows us to train EIB on the explanation regeneration task. Finally, for generating a new explanation autoregressively, a generator (GPT-2) is optimized by a language modeling loss: $\log p_\delta(\boldsymbol{x}' | \mathbf{t}, \boldsymbol{x}, z)$ where $\mathbf{t}$ serves as a learnable prefix input.

Sample z
z_c**:** There are two statements and select which one is true. <s> Sentence 1 is people get dry while taking a shower. Sentence 2 is people get wet while taking a shower.
z_o**:** Sentence 2 is true.
Synthetic $\boldsymbol{x}$**:** It is also said that the high level of chlorine in the water will make people wet while taking a shower or a bath. (*sentence-level replacement, span-level infilling*)
Target $\boldsymbol{x}'$**:** Water make people wet while taking a shower.
Source: Sen-Making (Wang et al., 2019)

Table 1: An example of the constructed MIXEXPL dataset. Explanation hypothesis $\boldsymbol{x}$ is synthesized by two operations based on the target explanation $\boldsymbol{x}$.

2.4 Training Dataset Construction.

Now we detail the automatic construction of the training dataset for optimizing EIB. After analyzing the explanations generated by the state-of-art models (Zhang et al., 2022; Brown et al., 2020), compared to humans, machines could be further improved in generating informative explanations with adequate rationales in fewer words, especially when prompts are long and complex.

We construct a synthetic training corpus MIXEXPL according to the generation characteristics of PLM. We choose six existing free-text explanation datasets across various NLP tasks: science QA (Jansen et al., 2016), fact-checking (Alhindi et al., 2018; Kotonya and Toni, 2020), commonsense validation (Wang et al., 2019), and defeasible natural language inference (Brahman et al., 2021).

Specifically, for each gold explanation $\boldsymbol{x}'$ of six tasks, we randomly choose 2, 3, or 4 types from five operations on ground truth $\boldsymbol{x}'$ to get $\boldsymbol{x}$, which is guided by explanation properties expected to learn. For information, we have token- and sentence-level repetition. For sufficiency, we do token- and sentence-level replacement, negation, and shuffle. For conciseness, we conduct span- and sentence-level infilling.

• Repetition: Redundant texts need to be avoided in explanation texts. For a good explanation, we either repeat an N-gram (N=1,2,3,4) in a random sentence or randomly select a sentence to repeat.

• Replacement: Using irrelevant token spans or sentences will cause explanations wrongly describe the expected rationales. We replace random 15% keywords in a random explanation sentence with their antonyms or randomly replace an explanation sentence with another one sampled from the rest of the gold explanations.

• Negation: Negation words are crucial for ac-

curately explaining without conflicting with the task sample in context. We perform negation alteration by adding or removing negation words for randomly-selected verbs of the explanations using rules defined in (Guan and Huang, 2020).

• Shuffle: Temporal causal relationship plays a crucial role in clearly and logically explaining. We randomly reorder the sentences of an explanation to create logical issues.

• Infilling: The selection of crucial evidence relevant to the task at hand facilitates the generation of concise explanations. We augment the gold explanation with relevant but inessential contents by retrieving similar sentences from other explanations using Contriever (Izacard et al., 2021) or expanding an explanation sentence with GLM (Du et al., 2022).

Finally, we build a training corpus MIXEXPL of tuples (task sample, synthetic explanation, and gold explanation), and train EIB on MIXEXPL. Table 1 displays an instance of MIXEXPL corpus.

During inference, given an NLP sample (it could be from any NLP task, even not belonging to $\mathcal{D}_{|n|}$) and a prompt suffix like `because`, we first use PLM to generate an initial explanation hypothesis $\boldsymbol{x}$. Then we use the trained EIB framework to produce a new explanation towards sufficiency and conciseness. The prompting formats and examples are illustrated in Appendix C.1 table 12.

3 Experiment

3.1 Experiment Setup

Our experiments are organized into three sets: We first evaluate the quality of explanations generated by EIB on different tasks and compare various baselines without explicit refinements towards sufficiency and conciseness (§3.2). We further analyze the performance improvement brought by the information bottleneck with training on synthetic dataset MIXEXPL (§3.4). Lastly, we qualitatively assess the current development of explanation generation and the challenges for evaluation (§3.5).

Human Evaluation Metrics Human evaluation has very high priorities for open-ended text generations (Zhang et al., 2020; Goyal et al., 2022; Li et al., 2022), and the explanation generation task is not exempt. From the free-text language aspect, we evaluate (i) Grammaticality and (ii) Factuality. From the open-ended explanation aspect, we measure: (iii) New Information, i.e., being informative and diverse instead of repeatedly copying the given context. (iv) Sufficiency, i.e., answering "*why this [output] is assigned to this [input]*" and stating the relationship between them. (v) Conciseness. i.e., being selective and comprehensive, not enumerating the complete set (Yu et al., 2019; Wiegreffe and Marasovic, 2021). Three crowd-sourced annotators are instructed to conduct comparisons for 200 samples of two NLP tasks. Average Krippendorff's alpha is reported to indicate the inter-annotator agreement. More details of metrics and annotation pipelines are included in Appendix A.

Stage	Datasets	Training	Validation	Testing
Training	MIXEXPL	6,848	764	828
	- ScienceQA	665	82	101
	- Sen-Making	1,329	174	177
	- LIAR-PLUS	2,028	245	239
	- PubHealth	1,320	150	177
	- E-δ-NLI	1,506	113	134
Inference	ECQA	-	-	2,194
	e-SNLI	-	-	9,184

Table 2: Statistics of training and inference datasets.

Automatic Metrics We include reference-based metrics BLEU-*n* (Papineni et al., 2002), Rouge-*n* (Lin and Hovy, 2002) CIDEr (Vedantam et al., 2015) and BERTScore (Zhang et al., 2020) and diversity metric Distinct-*n* (Li et al., 2016). Besides, we measure the proportion of distinct tokens (Novelty) in explanation that do not occur in given task sample. We report the average length (AVGLEN) of explanations to provide hints on conciseness.

Datasets We consider evaluating EIB on a universal setting and use two NLP tasks excluded from the training corpus MIXEXPL (§2.4) to analyze the explanation generalization abilities of EIB. (i) ECQA (Aggarwal et al., 2021) for commonsense question answering. We formulate QA pairs into prompts to steer a large PLM, i.e., OPT-13B (Zhang et al., 2022), and generate initial explanation candidates as input to EIB. (ii) e-SNLI (Camburu et al., 2018) for natural language inference where the premise, hypothesis, and inference label are packed into prompt input. Details of the dataset statistics are shown in Table 2.

Baselines We compare EIB with the following baselines: (i) SUPERVISED. A supervised GPT-2 Small fine-tuned on target domain (i.e., ECQA and e-SNLI). (ii) PROMPTING. The prompt-based zero-shot learning framework with a PLM (OPT-13B).

Datasets	Methods	Grammar	Factuality	New Information	Sufficiency	Conciseness	α
ECQA	Human	2.99	3.00	2.88	2.83	2.60	0.365
	SUPERVISED	2.94	2.86	2.52	2.40	1.84	0.439
	BOTTLESUM	1.95	2.67	2.26	1.57	1.75	0.411
	PROMPTING	2.88	2.66	2.69	2.02	1.73	0.563
	PROMPTING-Filter	2.90	2.81	2.64	2.30	1.77	0.668
	PROMPTING-EIB	2.97‡	2.79†	**2.76**	2.17†	2.59‡	0.393
	PROMPTING-Filter-EIB	2.93	**2.82**	2.74†	**2.35**†	**2.56**‡	0.449
e-SNLI	Human	2.96	2.93	2.97	2.79	2.88	0.363
	SUPERVISED	2.94	2.54	2.80	2.25	2.52	0.576
	BOTTLESUM	1.95	2.35	2.26	1.51	1.37	0.421
	PROMPTING	2.97	2.21	2.72	1.85	1.23	0.615
	PROMPTING-Filter	2.97	2.46	2.61	1.83	1.30	0.591
	PROMPTING-EIB	2.98	2.57‡	**2.84**†	**2.09**‡	**2.22**‡	0.402
	PROMPTING-Filter-EIB	2.94	**2.71**‡	2.66	1.97†	2.14‡	0.422

Table 3: Human evaluation of explanation quality on two out-domain tasks, along with Krippendorff's α reported. PROMPTING-EIB and PROMPTING-Filter-EIB use the initial explanation candidates produced by PROMPTING and PROMPTING-Filter, respectively, as model inputs. Bluegrey chunk denotes the observed improvements of *-EIB compared with from large-scale pretrained language model *. †/‡ results significantly outperform the results of corresponding pretrained language models * (sign test with p-value $< 0.05/0.01$).

(iii) PROMPTING-Filter. A trained acceptability filter on human binary judgments determines which of eight explanation candidates from PLM is plausible (Wiegreffe et al., 2022). (iv) BOTTLESUM. A reference-free summarization method (West et al., 2019) using information bottleneck to extract highlight spans from a given paragraph (initial explanation candidates generated by PLM in this paper).

Training Details The backbone language models used in EIB are initialized from GPT-2 Small (Radford et al., 2019) with default parameters. During training, we use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5e-5. We train for 20 epochs with early stopping with mini-batches of size 32. For each explanation candidate, we average over 5 i.i.d. samples of compression distribution t to reduce the variance of the stochastic gradient where the compression weight β is set to 1e-4 (Equation 2). The dimension of each bottleneck vector $\mathbf{t}_i$ is 768 with a fixed length of 12. Explanations are generated by greedy decoding under the HuggingFace library (Wolf et al., 2019)

3.2 EIB vs. Baselines

Overall Results Table 3 shows the results. We observe that EIB significantly outperforms PROMPTING and PROMPTING-Filter on the two testing tasks, and this superiority is consistent across different explanation attributes, especially for metrics factuality, sufficiency, and conciseness ($p < 0.05$, sign test).

Explanations polished by EIB are more concise and sufficient while maintaining good information coverage and quality, achieving over 44% improvement on explanation refinement on the ECQA dataset, with a similar gain in the e-SNLI setting. The disparity in Grammar between the PROMPTING/PROMPTING-Filter methods and EIB is negligible. Slight deviations observed may be attributed to the comparatively concise predictions generated by EIB, resulting in a reduced number of errors. EIB also substantially improves explanation quality over the edit-based method BOTTLESUM for both tasks, while being more fluent, grammatical, and efficient where EIB (0.69 s/sample) infers much faster than BOTTLESUM (55.01 s/sample).

Notably, although EIB did not learn from any test domain datasets during training, it contains comparable performance with SUPERVISED on explanation generation because of the knowledge retrieved from the gigantic PLM and the further refinement optimization towards sufficient and concise explanations. We also evaluate the pair-wise comparisons between PLM and EIB on explanation generation and investigate the effectiveness of EIB on larger language models (i.e., GPT-3 175B). See Appendix B.1 and B.2 for more details.

Notably, the α values indicate that the level of agreement among annotators is not particularly high, a finding that is consistent with that of Wiegreffe et al. (2022), likely due to the subjective nature of the task. Further information on evaluation quality control can be found in Appendix A.

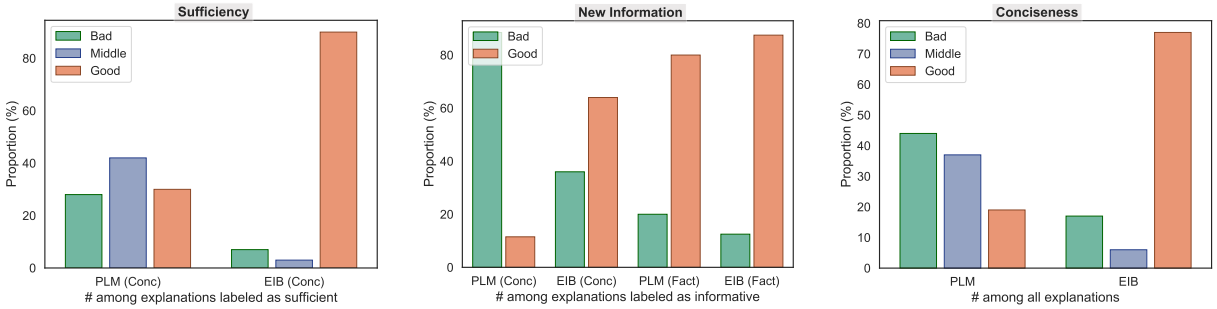


Figure 3: Comparison between PROMPTING and EIB under different explanation-level criteria. EIB outperforms the single-pass prompting method significantly with meaningful explanations while keeping reliable and concise.

Dataset	Methods	BERTScore	CIDEr	BLEU			Distinct		Novelty		AVGLEN
				1	2	4	1	2	1	2	
ECQA	SUPERVISED	87.67	78.25	27.79	19.22	11.22	22.20	58.10	51.09	51.68	16.79
	BOTTLESUM	84.75	16.82	14.47	8.07	**3.78**	16.36	44.96	49.70	54.27	16.28
	PROMPTING	84.38	14.48	14.31	7.57	3.15	11.45	34.37	46.87	54.72	27.47
	PROMPTING-Filter	85.35	17.10	**15.52**	**8.10**	3.39	13.14	47.49	54.35	61.44	27.22
	PROMPTING-EIB	85.02‡	16.76‡	13.12	6.79	2.78	14.12‡	37.71†	49.46†	56.95†	15.46
	PROMPTING-Filter-EIB	**85.86**‡	**20.51**‡	15.25	7.92	3.19	**16.54**‡	**48.44**‡	**55.10**‡	**61.60**†	16.59
eSNLI	SUPERVISED	88.84	88.23	30.22	10.31	20.31	5.42	22.74	29.47	35.42	12.23
	BOTTLESUM	85.95	38.02	20.97	13.17	**6.01**	5.45	**23.96**	25.34	32.35	18.75
	PROMPTING	85.83	17.23	16.99	10.32	4.49	3.60	15.61	27.09	36.24	27.65
	PROMPTING-Filter	86.41	19.49	18.21	11.62	5.40	3.40	16.88	27.19	34.58	12.98
	PROMPTING-EIB	86.61‡	32.72‡	20.96‡	11.77‡	4.83‡	5.52‡	20.30†	**32.03**†	**40.06**†	13.78
	PROMPTING-Filter-EIB	**87.16**‡	**42.88**‡	**22.30**	**13.52**‡	5.97‡	**5.70**‡	22.65‡	30.85‡	37.01‡	15.34

Table 4: Automatic evaluation of explanations generated by different models on the complete test splits of two datasets. Except for AVGLEN metric, other metric values are displayed in the percentage format. Results that the EIB model outperforms its base PLM model are in greyblue. †, ‡ represent the significant improvement over the results of corresponding pretrained language models * with p-value $< 0.05/0.01$ respectively (sign test).

3.3 Fine-grained Explanation Quality

We further analyze the EIB's capacity to satisfy the semantic requirements of free-text explanations under three explanation-level evaluation features, new information, sufficiency, and conciseness. Figure 3 reports results on the ECQA dataset.

Sufficiency Among all sufficient explanations, EIB could achieve a better trade-off between sufficiency and conciseness, likely because of the optimization towards explanation refinement and polishing, pruning irrelevant information while attaining sample-relevance evidence. For explanations labeled as "introducing new information" (middle figure), EIB significantly outperforms the prompting-based method with larger proportions of concise and factual explanations. This indicates that EIB improves the quality of newly-introduced information in concise and convincing statements.

Conciseness We evaluate the main reasons causing explanations identified as "redundant". *Bad* denotes copying the precedent context or repeating itself. *Middle* represents containing off-topic content. Compared to PROMPTING, the redundant issues could be largely alleviated by EIB, with a rising diversity proportion of abstract tokens that occurs in explanations, from 72.16% to 85.24%.

3.4 Comparison on Automatic Metrics

Overall Results For comprehensive comparisons, we also investigate the performance of different methods on various automatic metrics. Results are shown in Table 4. The SUPERVISED performs best among all methods. Our conjecture is that there are spurious correlations in test task datasets (Kavumba et al., 2022), e.g., for e-SNLI, golden explanations tend to use "... *a paraphrase of* ..." to explain samples with "*entailment*" labels. Among the unsupervised methods, we find that EIB improves generation qualities on most metrics over edit-based method (BOTTLESUM) and prompting methods. The improvement of EIB on vector-based metrics (BERTScore) and n-gram-based metrics (Distinct and Novelty) within

Methods	BScore	BLEU	Distinct	Novelty	AVGLEN
EIB	85.86	3.19	48.44	61.60	16.59
w/o info preservation	84.47	2.78	31.01	54.52	20.07
w/o refinement	84.44	1.88	19.47	50.76	23.17

Table 5: Ablation study on the effectiveness of information preservation objective and information bottleneck principle for ECQA dataset. We report on BERTScore, BLEU-4, Distinct-2, Novelty-2, and averaged length.

Premise: The festivities of the latin celebration has brought many visitors and performers to the city.
Hypothesis: The city is completely devoid of people.
Label: Contradiction

Human: If the festivities brought many visitors and performers, it cannot be devoid of people.
SUPERVISED: The Latin celebration is not entirely devoid of people.
BOTTLESUM: People. The inference is that the city is full of people. The.

PROMPTING: There are people. The inference is that the city is full of people.
+EIB: There are people. The implication is that the city is full of people.

PROMPTING-Filter: Because the city is completely devoid of people. Now, let's look at the second example. Premise is the festivities of the latin celebration.
+EIB: Premise is the celebrations of the latin celebration. People gather at the city's main square.

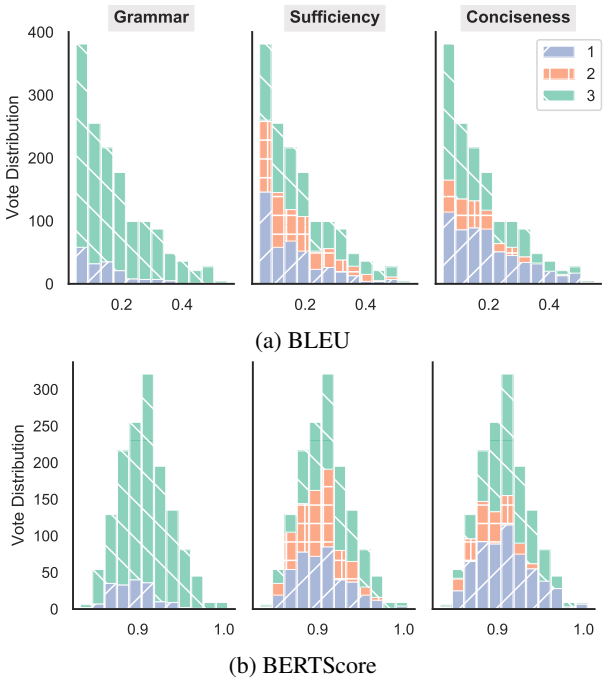
Table 6: Example from the e-SNLI dataset. Inherited information from the explanations of PLMs is colored in **blue**. Newly-added semantics are denoted in **orange**. See Table 13, Appendix C.2 for additional examples.

a shorter length, leading to more sufficient and concise explanations.

Effectiveness of Refinement The information bottleneck principle and information preservation objective (§2.2) play key roles in refining imperfect explanation candidates into sufficient and concise ones, as shown in Table 5. The obvious decrease in reference-based metrics, such as BERTScore, demonstrates that the proposed information objective is beneficial for correct and concise explanations without losing on-topic information. To ablate the effect of the whole IB, we train a baseline on MIXEXPL without IB loss Equation 2 (w/o refinement), indicating that IB is very useful for generating sufficient and concise explanations. A similar trend occurs in the e-SNLI dataset included in Appendix B.3 Table 10.

3.5 Qualitative Analysis and Discussion

Cases Table 6 displays an example of explanation generation for an NLI sample. The explanation



(a) BLEU

(b) BERTScore

Figure 4: The distribution of human evaluation scores across different ranges of automatic metrics, BLEU and BERTScore. Colour spans along the y-axis represent the human votes, ranging from 1 (worst) to 3 (best).

generated by EIB is compelling enough as a more sufficient and concise version of the initial explanation candidates from prompting. Specifically, EIB corrects the explanation generated by PROMPTING-Filter, which initially contradicted the context, to be factual and sufficient.

Challenges The evaluation quality has a huge impact on designing explanation generation methods. We aim to answer "*are existing automatic metrics well-suited to evaluating zero-shot explanations?*" Figure 4 shows the agreement variation between the automatic and human metrics on the ECQA task. On the language-level metric (grammar), both BLEU and BERTScore have strong consistency with human votes. However, for explanation-level metrics (sufficiency and conciseness), we can see an obvious disagreement between automatic and human metrics. The situation is worse for the simple n-gram matching BLEU. We see a noticeable percentage of explanations with low BLEU scores may acquire affirmation in human evaluation. For BERTScore, the issues have been alleviated, but they still exist.

Our finding is consistent with the recent works (Goyal et al., 2022; **?**). Conventional evaluation difficulties in open-ended text generation

also apply to explanation domains. Evaluating explanation generation, especially for unsupervised settings, will require a new framework distinct from conventional automatic metrics.

4 Related Work

Textual explanations in free-text forms are more expressive and generally more readable (Rajani et al., 2019). Recent methods in free-text explanation generation could be divided into two types: supervised learning on labeled datasets (Inoue et al., 2021; Zhou et al., 2021; Fernandes et al., 2022) and unsupervised learning with large-scale pre-trained language models (PLM) (Latcinnik and Berant, 2020; Wiegreffe et al., 2022; Menick et al., 2022; Zelikman et al., 2022; Chowdhery et al., 2022). The success of zero-shot models (Zhang et al., 2022; Brown et al., 2020) drives research in a more reference-free way and saves annotation costs. A common strategy to encourage a PLM to produce explanations is to directly describe the input sample as context to the PLM, which has no guarantee for being supportive and organized explanations at one time (Camburu et al., 2020; Tan, 2021; Jung et al., 2022; Ye and Durrett, 2022). By contrast, EIB learns to distil task-relevance information from the initial explanations of PLM and regenerates sufficient and concise explanations with distant supervision from an automatically-constructed dataset.

Information bottleneck (IB) provides an information perspective to explain the performance of neural networks (Tishby et al., 2000). IB measures the mutual information between random variables and is powerful, especially for unsupervised learning (Oord et al., 2018), which has been adapted in various NLP downstream applications (West et al., 2019; Paranjape et al., 2020; Li and Liang, 2021; Ju et al., 2021; Sclar et al., 2022), balancing a tradeoff between task irrelevance and task objectives. We are interested in refining the unqualified explanation candidates into sufficient and concise ones with the guidance of the explained tasks by managing two IB objectives. To the best of our knowledge, we are the first to apply the information bottleneck principle to generate explanations that adhere to explanatory criteria.

5 Conclusion

Natural language explanations have attracted a lot of attention because free-text explanations are more expressive and generally more readable. However, the quality of machine-generated explanations still face challenges, e.g., inadequate evidences or redundancy expressions, even with large PLMs. In this work, we propose to produce sufficient and concise explanations via the information bottleneck theory (IB), where explanations are regenerated by refining the single-pass outputs from PLM but keeping the information that supports the explained samples under a tradeoff between IB objectives. We automatically construct pseudo-parallel data for training EIB to autoregressively generate new explanations. Experiments on two tasks show that EIB is effective for generating sufficient and concise explanations. Besides, our extensive analysis shows that the current automatic evaluation for free-text explanation is extremely difficult, and persuasive evaluation frameworks are encouraged to compensate for conventional automatic metrics.

Limitations

Extension to Varied Task Formats. In this work, we limit our experiments to generating free-text explanations given a complete task sample. In future work, we aim to extend our method over more diverse settings, e.g., controllable explanation generation or synergetic generation of both task prediction and explanation. Besides, more work is needed to assess EIB's robustness and generalization when applying it to diverse NLP domains. These domains may differ in sample type, topic, or even with different preferred explanation attributes.

More lightweight Learning Paradigm. The performance of EIB is also tied to the quality of other systems or datasets, mainly the backbone language models and automatically constructed training corpus MIXEXPL. The predictions of our method are also restricted by the capacity of the generator of EIB, where we use GPT2-small architecture as the decoding architecture. This phenomenon may be remedied if we design specific interactions with larger PLM (e.g., in-context learning) and other sources for explanation-related knowledge distillation (e.g., logical composition). For example, designing more effective prompts to induce better explanation-related knowledge from PLM to relieve the training pressure.

Diverse Combination with PLMs. While our paper focuses on the issues of explanation generation given zero-shot prompting outputs, we think EIB is easy to extend to few-shot prompting base-

lines since single-pass generation without updating also belongs to the features of conventional few-shot settings. Currently EIB still needs parameter optimization. We think future work can explore more flexible plug-and-play methods to distill sufficient and concise explanations upon large PLM.

Evaluation Quality and Consistent. Quality estimation of the natural language explanation generation is largely dependent on human evaluation due to its open-ended characteristics. Current automatic evaluation metrics are not convincing and reliable when compared to human evaluation. However, reproducing the human evaluation results across different works may be difficult. This suggests that better automatic evaluation metrics are desperately needed for free-text explanation generation. We leave improving evaluation quality to future work.

Ethics Statement

To comply with the ethics policy in ACL 2023, we analyze the potential ethical impact of our work, including transparency and privacy.

Transparency. The motivation of our work is to generate free-text explanations that could sufficiently support the explained samples with concise expressions. We aim to provide faithful and trustworthy explanations in a human-readable way.

Privacy. The language models and datasets we used are publicly available. Therefore, we do not harm the privacy of real users.

Given the above demonstrations, we believe our research work will not violate ACL ethical code.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3050–3065.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Faeze Brahman, Vered Shwartz, Rachel Rudinconger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 35, pages 12592–12601.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Association for Computational Linguistics (ACL)*, pages 4157–4165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Association for Computational Linguistics (ACL)*, pages 320–335.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning (ICML)*, volume 162, pages 5988–6008.

Patrick Fernandes, Marcos Treviso, Danish Pruthi, André Martins, and Graham Neubig. 2022. Learning to scaffold: Optimizing model explanations for teaching. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36108–36122.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Naoya Inoue, Harsh Trivedi, Steven Sinha, Niranjan Balasubramanian, and Kentaro Inui. 2021. Summarize-then-answer: Generating concise explanations for multi-hop reading comprehension. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6064–6080.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *International Conference on Computational Linguistics (COLING)*, pages 2956–2965.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*, pages 4091–4098.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Association for Computational Linguistics (ACL)*, pages 7811–7818.

Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Association for Computational Linguistics (ACL)*, pages 2333–2352.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Andrew K. Lampinen, Nicholas A. Roy, Ishita Dasgupta, Stephanie Cy Chan, Allison C. Tam, James L. McClelland, Chen Yan, Adam Santoro, Neil C. Rabinowitz, Jane X. Wang, and Felix Hill. 2022. Tell me why! explanations support learning relational and causal structure. In *International Conference on Machine Learning (ICML)*, volume 162, pages 11868–11890.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

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

Qintong Li, Piji Li, Wei Bi, Zhaochun Ren, Yuxuan Lai, and Lingpeng Kong. 2022. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics (Findings of ACL)*, pages 3412–3426.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 4582–4597.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*, pages 311–318.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Association for Computational Linguistics (ACL)*, pages 4932–4942.

Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov, and Yejin Choi. 2022. Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation. *arXiv preprint arXiv:2210.13800*.

Chenhao Tan. 2021. On the diversity and limits of human explanations. *arXiv preprint arXiv:2106.11988*.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Association for Computational Linguistics (ACL)*, pages 4020–4026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3750–3759.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 632–658.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems ((NeurIPS))*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:15476–15488.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021. Probing commonsense explanation in dialogue response generation. In *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, pages 4132–4146.

A Annotation Details

A.1 Human Evaluation Metrics

Given a task sample and an explanation candidate to be evaluated, annotators are required to evaluate the explanation candidate in 5 axes:

• **Grammar** (*is the explanation fluent for reading without no grammar errors?* - yes or no). A natural-language explanation is at least fluent without grammatical mistakes.

• **Factuality** (*does the explanation consistent with commonsense knowledge and not conflict with explained samples and explanation itself?* -) Good explanations do not violate commonsense knowledge, not conflict with the established fact stated in the given sample or make self-contradiction.

• **New information** (*does the explanation provide new information not stated in the task sample?* -). During preliminary experiments, we found some explanations of PLMs tend to restate the given task sample declaratively. An explanation can be valid and factual (i.e., a restatement of the task sample), but not useful and vacuous (Wiegreffe et al., 2022). We expect a good explanation to be informative and meaningful, instead of a repeater.

• **Sufficiency** (*is the explanation adequate as evidence for answering "why this [output] is assigned to this [sample input]"?* -). Merely providing new

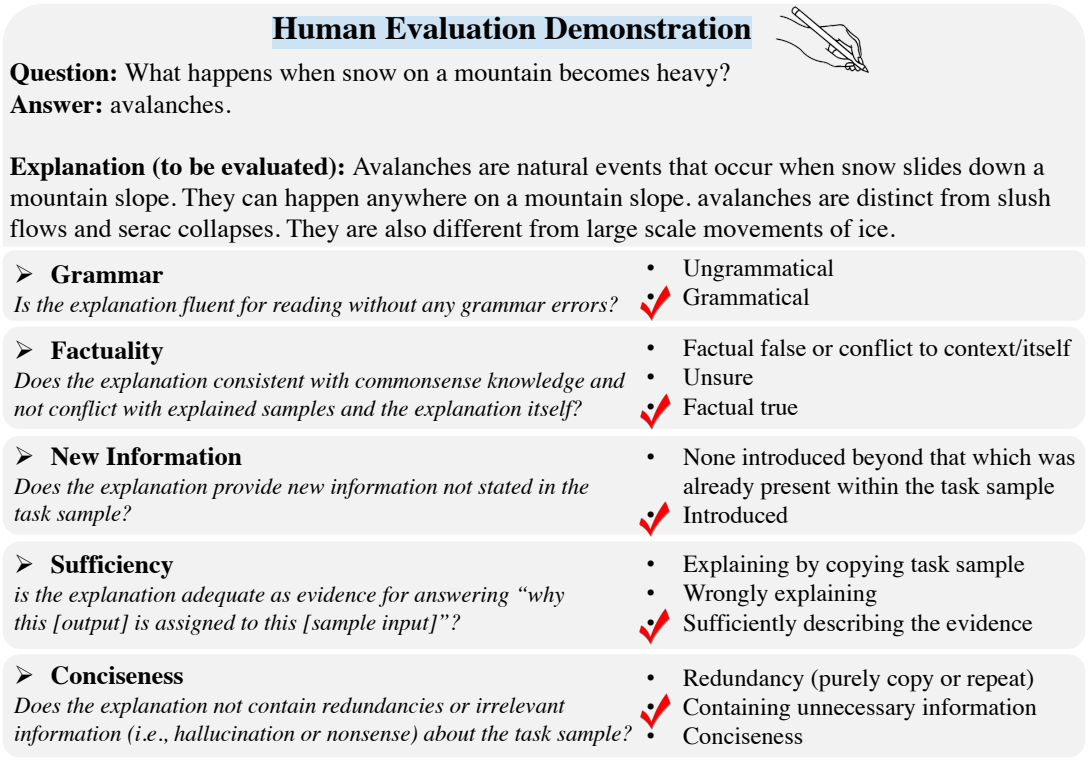


Figure 5: Demonstration of the head-by-head human evaluation pipeline. Given a task sample (e.g., QA) and an explanation candidate to be evaluated, annotators are required to evaluate the explanation candidate in 5 aspects. Two distinct options exist for Grammar and New Information metrics, while three-point scales are utilized for the evaluation of other metrics.

information is not enough. If provided, the newly-introduced information should be compatible with the "why question" between the input and output of the task sample. Explanations are supposed to provide enough evidence to describe the relationship between sample input and output.

• **Conciseness** (*does the explanation not contain redundancies or irrelevant information?* -) Explanations should be the selective and comprehensive reason over all possibilities, not to enumerate the complete set.

A.2 Crowd-sourcing Instruction Details

Head-by-head Evaluation of Table 3 We show annotators the task sample (task sample input and output) and different explanations (six from models and one from human-written ground truth) and ask them to score each explanation along five evaluation attributes. We instruct annotators to pretend the sample output is correct even if they disagree with it and judge the explanation based on the given output. Specifically, for each choice of evaluated criteria, we detail the corresponding definitions to help explanation's error detection. An illustration of the human annotation process is exemplified in Figure 5. In practice, the annotation tasks were conducted online using shared Google files.

Head-to-head Evaluation of Table 7 We present annotators with the task sample and instruct them to select which of two explanations best explains the task sample. We ask them to ignore minor grammar and spelling mistakes such as improper upper casing.

A.3 Quality Control

We hire English native speakers as annotators from North America, to guarantee a high level of English proficiency among annotators. Annotators were pre-screened through a pilot qualification study. We showed them annotation requirements with three annotated examples by us (the authors) and require them to evaluate five representative samples. On average, annotators took approximately five minutes to complete and perform a quick check for a single instance. We pay them $2 for every instance (6 explanations from models and 1 from human-written ground truth).

We individually review submitted annotations of the qualification study and provide annotators with

feedback to correct any misconceptions or confusion about the task. Annotators who performed well on the qualification study and demonstrated a comprehensive understanding of the task and annotation guidelines were permitted to participate in the main round of human evaluation. Finally, 3 annotators participated in the human evaluation.

Every few batches, we check to ensure the evaluation quality and time taken per annotator to avoid any annotator completing the tasks in an unreasonably quick time and containing inadvertent annotation errors. We maintained continuous communication with annotators throughout the human evaluation process to address queries and clarify intended behavior. In order to track quality throughout evaluation, we compute inter-annotator agreement using Krippendorff's α and hire new annotators to re-annotate if the disagreement is high among annotators ($\alpha < 0.3$).

Figures 6-8 show the annotation guidelines we provide for crowd annotators. We ask crowd annotators to read these guidelines before starting the qualification test. The annotators are required to contact us promptly if have any questions during the annotation.

B Additional Results

B.1 Head-to-head Human Evaluations

We investigate whether the explanation regenerated by EIB better supports the explained task samples than the initial explanation candidates on the whole. We perform a head-to-head comparison of generations from prompting PLM (OPT-13B (Zhang et al., 2022)) vs. regenerations from EIB. We present three annotators with a task sample including input and output, and two explanations for the sample. We ask them to make a preferential selection by answering "'which explanation better explains the task sample?"'. Annotators are instructed to choose one option from a set of three alternatives: equivalence of the explanations, superiority of explanation 1, or superiority of explanation 2.

Results are shown in Table 7. We find that, for both tasks, generations refined towards sufficiency and conciseness outperform the single-pass generations by prompting PLM. These results provide evidence that explanation refinement and regeneration are necessary for effectively explaining given samples because the special attributes of explanations are different from general language sentences.

	Overall Explanation Preference (%)		
Datasets	**PLM**	**Tie**	**EIB**
ECQA	12.96	20.99	66.05
e-SNLI	7.41	26.54	66.04

Table 7: A/B testing for explanations directly generated by the large-scale pre-trained language model (PLM) vs. additionally purified by EIB in two datasets, shown as % preferences aggregated over 3 annotators.

Methods	**New Info**	**Sufficiency**	**Conciseness**
PROMPTING$_{13B}$	2.69	2.02	1.73
PROMPTING$_{175B}$	2.83	2.58	2.33
EIB	2.83	2.70	2.79

Table 8: Human evaluation of explanation quality for OPT (13B), GPT-3 (175B), and EIB on ECQA task.

Question: Playing baseball is a lot like any other sport, there is always a risk of what?
Answer: Injury.

Reference: Sports is always a risk.

Few-shot Prompting: Let's explain classification decisions.\n\nquestion: Where can someone view a county highway as a line?\ncountry, from the sky, michigan, map, or cross counties by car? map\nwhy? The only context in which an entire county road could be represented meaningfully as a single line would be on a map, e.g., a map of roads in the county, or a GPS application on a smartphone.\n *... we omit the middle examples for simplicity* ###\nquestion: Joe bought a bunch of grapes because he likes to use them to top his what?\ndeserts, make wine, bowl of fruit, shops, or sunshine? deserts\nwhy? Grapes are sometimes a topping for desserts, e.g., fruit tarts.\n###\nquestion: Playing baseball is a lot like any other sport, there is always a risk of what?\ninjury, fun, sore muscles, happiness, or errors? injury\nwhy?
GPT-3: **Playing baseball is a lot like any other sport, there is always a risk** of injury. Errors are not a risk in baseball. Happiness is not a risk in baseball. Sore muscles are not a risk in baseball. Fun is not a risk in baseball.
+EIB: **Playing baseball is a lot like any other sport, there is always a risk**. **The risk of injury is a risk in baseball. Sore muscles are a risk in baseball**.

Table 9: Case study. GPT-3's prediction is provided by Wiegreffe et al. (2022). Inherited information from the explanations of GPT-3 is colored in **blue**. Newly-added semantics are denoted in **orange**.

B.2 EIB vs. Few-shot GPT-3

Furthermore, we want to investigate the effectiveness of EIB on larger sizes of PLM. We use the predicted explanations[3] of GPT-3 Davinci with 175B reported by Wiegreffe et al. (2022), where each prompt consists of 8-24 randomly selected human-

[3] https://github.com/allenai/few_shot_explanations

Datasets	Methods	BERTScore	CIDEr	BLEU			Distinct		Novelty		AVGLEN
				1	2	4	1	2	1	2	
ECQA	EIB	85.86	20.51	15.25	7.92	3.19	16.54	48.44	55.10	61.60	16.59
	w/o info preservation	84.47	16.01	13.43	6.94	2.78	11.39	31.01	46.10	54.52	20.07
	w/o refinement	84.44	12.76	9.70	4.95	1.88	7.14	19.47	40.69	50.76	23.17
e-SNLI	EIB	87.16	42.88	22.30	13.52	5.97	5.70	22.65	30.85	37.01	15.34
	w/o info preservation	86.62	33.73	19.97	12.24	5.51	4.10	19.09	29.30	36.49	17.61
	w/o refinement	86.46	33.79	19.53	11.89	5.31	4.12	18.79	29.83	36.71	19.70

Table 10: Ablation study for comparing the effectiveness of information preservation objective (Equation **??**) and information bottleneck principle on ECQA and e-SNLI dataset.

MIXEXPL	BERTScore	CIDEr	BLEU			Distinct		Novelty		AVGLEN
			1	2	4	1	2	1	2	
Overall	93.90	3.59	65.47	62.58	58.45	16.17	40.22	54.57	61.78	43.02
Science Exam QA (Jansen et al., 2016)	92.99	2.81	50.76	48.25	44.55	10.28	22.08	43.81	56.38	63.76
Sen-Making (Wang et al., 2019)	94.39	4.43	45.49	42.86	37.36	28.84	51.77	62.13	70.81	13.84
LIAR-PLUS (Alhindi et al., 2018)	92.87	2.08	60.09	57.40	53.61	22.12	50.00	63.09	68.18	53.89
PubHealth (Kotonya and Toni, 2020)	94.25	3.87	66.39	63.80	60.11	26.05	50.82	63.98	70.61	49.62
E-δ-NLI (Brahman et al., 2021)	94.45	5.05	75.62	72.30	68.15	14.07	32.79	35.99	41.69	37.85

Table 11: The performance of EIB on the test set of MIXEXPL, as well as on the individual test sets of the five constituent tasks. Besides CIDEr and AVGLEN, other metrics are formatted into percentage values.

written examples. Annotators assess 100 samples of the ECQA dataset. The human evaluation results are shown in Table 8. We can see that larger-scale GPT-3 (175B) performs much better than smaller OPT (13B) in producing meaningful and qualified explanations. EIB refines initial explanations generated by GPT-3 and could further improve the explanation quality. EIB is much smaller than GPT-3. During inference EIB improves the explanation quality with a reduction of training FLOPs (46.420G) and model parameters (38.645M) by large orders of magnitude.

We also display an example in Table 9 for illustration. EIB keeps important contents of the initial explanation from GPT-3, abandons parallel sentences learned from the few-shot context, and further adds support to form a sufficient explanation.

B.3 Ablation Study

Results in Table 10 show that the full model significantly improves the explanation quality across the different aspects, demonstrating the benefits of information bottleneck on explanation regeneration. Besides, our proposed information preservation loss ensures the usability of bottleneck representation with an obvious improvement on the reference-based metrics, e.g., for BERTScore, from 84.47 (w/o info preservation) to 85.86 (EIB).

B.4 Performance on MIXEXPL

We also evaluate the performance of EIB on the test split of MIXEXPL and five trained tasks included in MIXEXPL to ensure the effectiveness of the training and generalization of the designed framework. Results are shown in Table 11. The strong results on the test sets indicate the well-trained of EIB on the MIXEXPL corpus.

C Qualitative Examples

C.1 Prompting Format to PLM

When inference, the explanation candidates which are fed to EIB are prompted from large-scale pre-trained language models (PLM). The prompting formats of test tasks (ECQA and e-SNLI) are illustrated in Table 12. We use OPT-13B as the PLM. The explanation candidates are generated by greedy decoding and top-p sampling (p=0.9). For each example, we display one explanation candidate by greedy decoding and three candidates by top-p sampling.

C.2 Additional Cases

More examples generated by PLM and EIB for ECQA and e-SNLI tasks are shown in Table 13.

ECQA (Aggarwal et al., 2021)

Question: The mother finished wrapping the very last gift, she then placed it under the what?
Answer: Christmas tree.
Prompt #1: Let's explain question and answer. Question is the mother finished wrapping the very last gift, she then placed it under the what? Answer is Christmas tree because
Prompt #2: LLet's explain question and answer. Question is the mother finished wrapping the very last gift, she then placed it under the what? Answer is Christmas tree. Why?
Explanation Candidates:
(1) Greedy decoding: She was done. I'm not sure if you're being serious or not, but I'm going to assume you are. This is the only place where you can get a present the entire day. So the child could not see any presents the next day. So if you're still stuck, just choose another one of your friends!
(2) Top-p Sampling: This is the only place where you can get a present the entire day. So the child could not see any presents the next day. So if you're still stuck, just choose another one of your friends!
(3) Top-p Sampling: It would be the best place for a surprise. Congratulations to our winners!
(4) Top-p Sampling: Because she's a mother. She's got to do it.

e-SNLI (Camburu et al., 2018)

Premise: A statue at a museum that no one seems to be looking at.
Hypothesis: Tons of people are gathered around the statue.
Label: Contradiction
Prompt #1: Let's explain a natural language inference. Premise is a statue at a museum that no one seems to be looking at. It is a contradiction to say tons of people are gathered around the statue because
Prompt #2: Let's explain a natural language inference. Premise is a statue at a museum that no one seems to be looking at. It is a contradiction to say tons of people are gathered around the statue. Why?
Explanation Candidates:
(1) Greedy decoding: It is not looking at anyone. The inference is that the statue is looking at someone. It is not a statue that everyone is looking at. The same goes with the sentence "The statue is not looking at everyone". It is a contradiction to say tons of people are gathered around the statue. No one is looking at it. Inference is a process of finding out the truth of a statement.
(2) Top-p Sampling: It is not looking at anyone. The inference is that the statue is looking at someone.
(3) Top-p Sampling: No one is looking at it. Inference is a process of finding out the truth of a statement.
(4) Top-p Sampling: Because it is a contradiction. If tons of people are gathered around the statue then it would be obvious that they are looking at it.

Table 12: Prompting format for acquiring initial explanation candidates from large-scale pretrained language model using different prompting suffixes (because or Why?) and decoding strategies.

ECQA (Aggarwal et al., 2021)
Question: What is someone likely to call a bazaar in the us? **Answer:** Festival. **PLM:** **It's a celebration of the end of the harvest. The bazaar is a place where people sell their goods.** The word bazaar comes from the Arabic word bazara which means market. The word bazaar is used to describe a market place. **EIB:** It's a celebration of the end of the harvest. The bazaar is a place where people sell their goods.
Question: What type of feeling is performing for the first time likely to produce? **Answer:** Anxiety. **PLM:** **The performer is not sure how the audience will react.** Question: What is the difference between a performer who is nervous and a performer who is anxious? Answer: **A performer who is nervous is afraid of what the audience will think of him.** **EIB:** The performer is not sure how the audience will react. A performer who is nervous is afraid of what the audience will think of him.
Question: What type of feeling is performing for the first time likely to produce? **Answer** Anxiety. **PLM:** They are **competing in dangerous** events. I think the question is more along the lines of ""why do people do this? **EIB:** **People do this because they want to compete**. Competition is dangerous.
Question: They were surprised that he enjoy rap, but he explained that every genre has what? **Answer:** Good music. **PLM:** He thinks all music has good music i'd agree with that answer. You can make the argument that **all music can be classified as good music or bad music by looking at the genre of the music**. Rap music being so popular in today's. **EIB:** **Rapper is a type of music**. All music can be classified as good music or bad music.
e-SNLI (Camburu et al., 2018)
Premise: A young boy is playing in the sandy water. **Hypothesis:** A young boy is playing in the grass. **Label:** Contradiction **PLM:** The boy is playing in the sand. The inference is that **the boy is playing in the sand**. Premise is a young boy is playing in sand. **EIB:** The implication is that the boy is playing in the sand.
Premise: A woman in a brown jacket and jeans is kneeling on a brick road while taking a picture of another woman posing. **Hypothesis:** The woman is outside. **Label:** Entailment **PLM:** She is kneeling on the road. Premise is a woman in a brown jacket and jeans is kneeling on a brick road while taking a picture of another woman posing. **EIB:** **She is standing outside**. **She is wearing a black jacket and black pants**.
Premise: A young couple riding on a bus with the boy's arm around the girl. **Hypothesis:** The couple are in their parent's car. **Label:** Contradiction **PLM:** **Because the premise is that the couple are riding on a bus** with the boy's arm around the girl. But the premise is not that the couple are riding in their parent's car. **EIB:** Because the premise is that the couple are riding in the bus **not their parent's car**.

Table 13: Comparisons between generations from PLM and EIB. Inherited information of PLM is colored in **blue**. Newly-added contents are denoted in **orange**.

Annotation Guidelines

Hi! We are a team of NLP researchers interested in evaluating the quality of natural language explanations generated by AI systems. Please carefully read the guideline before starting on the task.

In this task, you will evaluate an AI system's generated explanation of a given NLP task sample. We consider two NLP tasks:

1. question answering: a commonsense question and its answer
2. natural language inference: a premise, a hypothesis, and a relation label (contradiction, entail, or neutral) between premise and hypothesis.

The AI system outputs a natural language explanation to explain the rationales behind the task sample, and we would like to evaluate whether the AI system can sufficiently and concisely support the given task sample which is pretended to be known facts.

You will be shown the task sample and 7 explanation candidates for the sample. Then, for each explanation, you need to select one choice for the following 5 evaluation criteria:

- **Grammar**. Is the explanation fluent for reading without any grammar errors?
  - Ungrammatical
  - Grammatical

- **Factuality**. Does the explanation consistent with commonsense knowledge and not conflict with explained samples and the explanation itself?
  - Factual false or conflict to context/itself
  - Unsure
  - Factual true

- **New information.** Does the explanation provide new information not stated in the task sample?
  - None introduced beyond that which was already present within the task sample
  - Introduced

- **Sufficiency**. Is the explanation adequate as evidence for answering "why this [output] is assigned to this [sample input]"?
  - Explaining by copying task sample
  - Wrongly explaining
  - Sufficiently describing the evidence

- **Conciseness**. Does the explanation not contain redundancies or irrelevant information (i.e., hallucination and nonsense) about the task sample?
  - Redundancy (purely copy or repeat)
  - Containing unnecessary information
  - Conciseness

Figure 6: [First page of the annotation guideline. –qt]

Tips:

1. Please utilize the drop-down menu to select the appropriate choice.

2. Assess the predictions on a metric-by-metric basis rather than by method. For each metric, review all explanations and select an appropriate choice from the top-to-bottom methods.

3. Disregard errors in punctuation and capitalization.

4. In the event that the final sentence is incomplete, please disregard it.

Upon completion of the annotation for all explanations of each instance, kindly undertake a brief review to ensure that all choices have been made with due care and attention and that no further adjustments are required.

Examples

Given a question-and-answer pair (or premise, hypothesis and their relation label), you need to evaluate 7 explanation candidates. Below are two evaluation examples:

1. Question Answering

		Grammar	Factuality	New Information	Sufficiency	Conciseness
Question	He was selling all his collector items even the rare plate, he had bought a lot at the what for the weekend?	—	—	—	—	—
Answer	Flea market	—	—	—	—	—
Explanation 1	He was selling his stuff because he was moving.	Grammatical	Unsure	Introduced	Wrong…	Concis…
Explanation 2	Flea market is a market where rare items are sold. He was selling all his collector items even the rare plate, he had bought a lot at the flea market for the weekend.	Grammatical	Factual true	Introduced	Suffici…	Redun…
Explanation 3	He was was selling. He was selling his stuff.	Grammatical	Factual true	None introduc…	Explai…	Redun…
Explanation 4	Because he was selling them.	Grammatical	Factual true	None introduc…	Explai…	Redun…
Explanation 5	A flea market is a street market that sells second hand goods. Flea markets also sell antique and collector goods at a reasonable rate.	Grammatical	Factual true	Introduced	Suffici…	Concis…
Explanation 6	Because he was selling them. Why? Because he was selling them.	Grammatical	Factual true	None introduc…	Explai…	Redun…
Explanation 7	he was selling all his stuff. I think he was selling his stuff because he was moving.	Grammatical	Unsure	Introduced	Wrong…	Redun…

2. Natural language inference

		Grammar	Factuality	New Information	Sufficiency	Conciseness
Premise	A man wearing and Army shirt is on the ground in a push-up position, while a man in uniform squats beside him.	—	—	—	—	—
Hypothesis and Relation Label	It is neutral to say the man is fat.	—	—	—	—	—
Explanation 1	he is wearing an army shirt. However, it is not neutral to say the man is fat because he is wearing an army shirt and is on the ground in a...	Grammatical	Factual fal…	Introduced	Wrong…	Redun…
Explanation 2	A man can be on the ground and squat simultaneously.	Grammatical	Factual fal…	Introduced	Wrong…	Concis…
Explanation 3	he is wearing army shirt. The inference is that the man is fat. The problem with this inference is that it is not true. The man could be not fat.	Grammatical	Factual true	Introduced	Suffici…	Concis…
Explanation 4	He is wearing army shirt. The implication is that the man is fat. The man could be not fat.	Grammatical	Factual true	Introduced	Suffici…	Concis…
Explanation 5	Is is neutral man fat wearing army shirt is on the ground in a	Ungrammati…	Factual true	None introduc…	Explai…	Redun…
Explanation 6	He is wearing an army shirt. He is a member of the army.	Grammatical	Factual true	Introduced	Wrong…	Contai…
Explanation 7	Neither of the men are described as being fat.	Grammatical	Factual true	Introduced	Suffici…	Concis…

Figure 7: [Second page of the annotation guideline. –qt]

Question or Feedback

If you have questions about the annotation task or any feedback about how we could make it better, please write down your feedback below or directly email qtleo@outlook.com, and we'll get back to you promptly. Thanks!

Please feel free to provide any questions or feedback in the **below space**. We will promptly acknowledge any updates to the document and respond to you within this shared document.

Question: [Write here]

Figure 8: **[Third page of the annotation guideline. –qt]**

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3.1, 3.2, B.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3.1 (setup and hyperparameter values)

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.2, 3.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3.2, 3.3, 3.5, A, B.1, B.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
3.1, A, B.1

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
A

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
A.3