A Trusted Multi-View Evidential Fusion Framework for Commonsense Reasoning

Shuo Yang
School of Computer Science and Cyber Engineering, Guangzhou University
Guangzhou, China
yangshuodelove@gmail.com

Abstract

While deep learning models are powerful, they have limitations in tasks that require commonsense reasoning, as these tasks often involve interpreting information that may not be directly available in the input. Providing evidence has been proven to significantly enhance performance in commonsense reasoning tasks. However, there are various perspectives on evidence, including natural language explanations generated by pre-trained language models, facts derived from world knowledge like text corpora and knowledge bases, and rationales extracted from the input context. Hence, it is crucial to determine how to estimate the confidence degree of different evidence and how to combine them reliably. To address these challenges, this study proposes a trusted multi-view evidential fusion framework for reliable commonsense reasoning tasks that dynamically assesses the confidence of evidence and combines different views of evidence in a trustworthy manner. The proposed method is applied to three commonsense question-answering benchmarks, demonstrating that this approach can effectively reason with multi-view evidence and can compete with state-of-the-art performance.

Keywords: Multi-view learning, Evidential deep learning, Commonsense reasoning

1. Introduction

Many deep neural networks (DNNs) usually use softmax function to convert the continuous activations of the output layer to class probabilities. Softmax outputs a point estimate as parameter set of a categorical distribution (Sensoy et al., 2018). It reports high confidence even for incorrect prediction without the associated uncertainty. This can be illustrated by a toy example. Upon the multiple-choice question given in Fig. 1, the options provided are not accurate. When boiling water in a kettle, the correct outcome is that water will turn into steam. However, softmax may smugly provide a high estimation for choice B, which is contrary to the common sense. Furthermore, it is challenging to instill commonsense reasoning (CSR) abilities in DNNs to achieve humanoid reasoning (Narang et al., 2020; Zhou et al., 2020). Although NNs can effectively learn from a vast number of examples, human CSR, in reality, often does not require extensive example learning (Wang et al., 2023b; Rajani et al., 2019). Instead, humans acquire the knowledge required for CSR through their daily living.

At present, while many natural language models (LMs) possess CSR capabilities, it remains unclear to what extent evidence from different sources has an impact on the performance of reasoning. In CSR, current methods have yet to fully explore how multi-view and even conflicting pieces of evidence should be reasonably integrated to facilitate credible reasoning. For example, Aggarwal et al. (2021) and Rajani et al. (2019) utilize commonsense question-answering samples and standard explanations as input to train an LM. Their findings demonstrate that incorporating explanations for CSR as general knowledge into neural networks (NNs) can improve performance on commonsense question answering (CQA). However, DeYoung et al. (2020) highlighted the importance of comprehensiveness in evaluating the faithfulness of model reasoning and showed that single source evidence may not be sufficient.

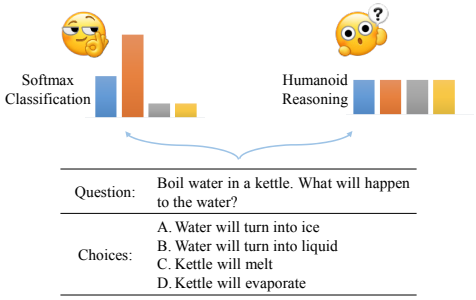


Figure 1: A motivation example.

As demonstrated by DeYoung et al. (2020), even when the answer is explicitly present in the rationale of some input samples, the model may select incorrect options. Moreover, the quality of different views of evidence varies across data samples. In this study, evidential quality is defined as the extent to which the evidence dominates model decision-making or provides efficacy for a reasonable prediction. Thus, CSR needs to integrate multiple sources of evidence (Narang et al., 2020).

This study presents a novel approach to CSR

LREC-COLING 2024, pages 1722–1733
20-25 May, 2024.

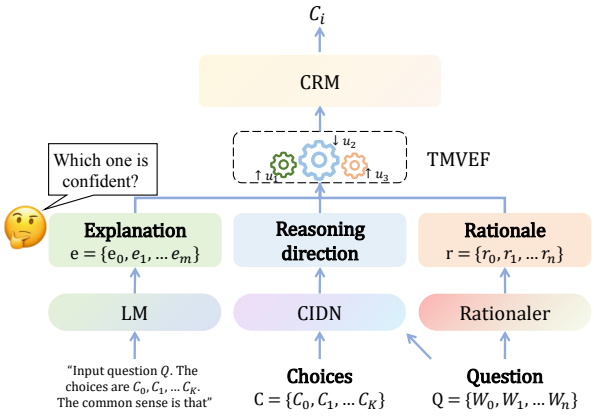


Figure 2: Overview of the proposed CRM under TMVEF framework. It consists of three components: (1) the explanation generation, (2) the Rationaler, and (3) the contextually independent directional network (CIDN). The explanation generation component takes the question Q concatenated with its answer choices $C_0, C_1, \cdots C_k$ as prompt to generate explanations. The Rationaler component is responsible for identifying tokens as key clues in Q. The CIDN is designed to forecast the direction of reasoning. The generated explanations, extracted rationales, and the reasoning direction are then fed into the downstream CRM under the proposed TMVEF framework to fuse different views evidence based on their uncertainties (u_i).

tasks that leverages multiple sources of evidence without prioritizing any of them, while dynamically evaluating their uncertainties. The proposed framework, called trusted multi-view evidential fusion (TMVEF), treats different evidence as multiple views, comprehensively evaluates the confidence degree of each view of evidence and reliably combines them. The TMVEF can produce an overall uncertainty (Antorán et al., 2020; Van Amersfoort et al., 2020; Charpentier et al., 2020; Sensoy et al., 2018) by using subjective logic theory (Jøsang, 2016) and use interpretable rules for evidential fusion. Based on TMVEF, an overview of the proposed *commonsense reasoning model* (CRM) is provided in Fig. 2.

The main contributions are as follows. This study (i) identifies a crucial issue with current CRMs in which they may fail to accurately evaluate the reliability of different views of evidence and combine them in a trustworthy manner. (ii) proposes a novel fusion framework that achieves a credible combination of multi-view evidence and trusted decision-making for CSR tasks, which benefit from the uncertainty from each view and evidence-level integration with the Dempster-Shafer theory (Le Hegarat-Mascle et al., 1997) in a learnable manner. (iii) proposes a rationale extractor that identifies important snippets from the input context (IC) as one view of reasoning evidence. (iv) proposes a CIDN network that forecasts the direction of reasoning to cope with the absence of reliable explicit evidence (e.g., rationales or explanations).

2. Related Work

Linguistic reasoning needs reasonable text interpretation which requires priori commonsense knowledge (Zhao et al., 2023; Bhargava and Ng, 2022; Lin et al., 2020a). One of the important ways to obtain commonsense knowledge is through commonsense text generation (Aglionby and Teufel, 2022; Choi, 2022). For example, Wei et al. (2022), Marasovic et al. (2022), Zelikman et al. (2022) and Lampinen et al. (2022) using LMs to generate explanations help improve the performance of CSR. However, DeYoung et al. (2020) pointed out that single view evidence may not be comprehensive for model reasoning. Because some commonsense knowledge is implicitly presented in the data (e.g., important snippets in the context) (Li et al., 2022), it is necessary to mine and integrate evidence from different views. Empirical experiments also demonstrated that the model using multiple modal information perform better than the model using only single modal information (Ma et al., 2023). However, evidence from different sources may vary in terms of reliability and effectiveness for reasoning. Therefore, it is necessary to explore faithful multi-view fusion strategies for reliable evidence consolidation (Kiela et al., 2018). Late fusion strategies are commonly adopted techniques, including decision averaging (Wang et al., 2019), decision voting (Barezi and Fung, 2019), and weighted approach (Zhang et al., 2023), to consolidate decision outputs from various sources.

Recently, research on interpretable reasoning (Wang et al., 2023a; Sensoy et al., 2018) has received widespread attention. For example, Sensoy et al. (2018) proposed a method for measuring the classification uncertainty based on single-view evidence. Eq. (1) uses subjective logic theory to simultaneously model belief masses b_k^e $(k = 1, 2, \ldots K)$ and the overall uncertainty u^e based on the regular output of NNs.

$$u^e + \sum_{k=1}^{K} b_k^e = 1 \tag{1}$$

where u^e represents the overall uncertainty of evidence e, and b_k^e represents the belief mass assigned to category k. The advantage is that the model can pay much more attention to the output from the view with low uncertainty and pay less attention to the view with high uncertainty. For evidence *e*, the subjective logic can associate evidence supporting different class labels

$\mathbb{E}^e = [e_1^e, e_2^e, \dots e_K^e]$ with the parameter of a Dirichlet distribution $\boldsymbol{\alpha^e} = [\alpha_1^e, \alpha_2^e, \dots \alpha_K^e]$ using $\alpha_k^e = e_k^e + 1$.

$$S^e = \sum_{i=1}^{K} (e_i^e + 1) = \sum_{i=1}^{K} (\alpha_i^e), \tag{2}$$

$$b_k^e = \frac{e_k^e}{S^e} = \frac{\alpha_k^e - 1}{S^e} \tag{3}$$

$$u^e = \frac{K}{S^e} \tag{4}$$

where S^e controls the overall strength of the distribution. A smaller S^e results in a more peaked distribution that concentrates probability around certain category, while a larger S^e generally leads to a distribution that is more spread out.

Different from prior studies, this study proposes an end-to-end CRM under TMVEF with interpretable evidential fusion rules for reliable CSR. Furthermore, multi-task learning strategy is used to optimize the parameters of the model, which not only considers the loss from each view of evidence but includes the overall loss after evidential fusion as well.

3. Trusted Multi-View Evidential Fusion Framework

Drawing from subjective logic theory (Sensoy et al., 2018), this study proposes a TMVEF framework. The framework is designed to meet two criteria to be considered "trusted": (i) the resulting decision is transparent and easily interpretable, and (ii) for cases that the model is unable to manage confidently, the framework assigns belief mass and uncertainty as degrees of confidence to illustrate the rationality of model reasoning. Unlike softmax, which provides results without interpretation, this approach allows for a more nuanced understanding of the decision-making process by simultaneously modeling the probability (belief mass) of each class and overall uncertainty.

3.1. Rationale Extractor: Rationaler

In this study, rationale is defined as snippets of input text at the token level. The goal of rationale extraction is to identify important snippets in the input text that are crucial for predicting the desired output (Chen et al., 2022; Glockner et al., 2020; Lehman et al., 2019). This study proposes a novel rationale extractor called *Rationaler*. It automatically highlights key points in the input text, which serve as essential evidence for reliable CQA. Because a rationale consists of different snippets with diverse distance intervals in the IC, and the relationships between the snippets may be parallel, progressive, or temporal, the BERT model with bidirectional encoding and a left-to-right LSTM with unidirectional dependency are integrated as a joint encoder to enhance the model's ability to handle different contexts and to enhance the robustness of the model.

Because rationale extraction involves identifying key tokens in an input sequence, this process is analogous to object detection in images or videos, where the goal is to identify and segment important objects. Thus, the negative cases dominate the overall loss. This imbalance may cause the model to become optimized and biased toward predicting negative tokens and to neglect positive tokens that are expected to be selected; therefore, model training may also be inefficient, as most snippets are easy negatives that do not provide useful learning signals. Additionally, for some sentences in the dataset, it is difficult to highlight important snippets. These sentences are referred to as hard-to-detect samples (HDS). As the number of HDS is relatively small compared with the number of easily detected samples (EDS), HDS account for a small proportion of the overall loss. However, EDS may not contribute to useful learning signals, and the model may not pay enough attention to HDS, which would lead to model degeneration when predictions are made on HDS in test datasets.

To address the challenge of positive/negative and HDS/EDS imbalance that is commonly encountered in computer vision, this study adopts the focal loss (FL) (Lin et al., 2020b). This loss function mitigates the disparities between the number of positive and negative tokens and overcome difficulties in identifying HDS in a small number of samples.

$$\mathrm{FL} = \begin{cases} -\alpha(1-p)^\gamma \log(p) & \text{if } y = 1 \\ -(1-\alpha)p^\gamma \log(1-p) & \text{if } y = 0 \end{cases} \tag{5}$$

where α controls the proportion of positive and negative tokens to resolve the imbalance between positives and negatives. γ controls the attenuation level of sample losses of EDS to resolve the imbalance between HDS/EDS. In the experiments, $\alpha = 0.9$ and $\gamma = 2$.

3.2. Multi-View Evidential Fusion Rule

Drawing on Dempster's combination rule for two independent sets of masses (Han et al., 2023, 2021), this study introduces an *evidential fusion rule* for the calculation of the joint belief mass for class k (b_k^J) and the overall joint uncertainty u^J to integrate two views of evidence e_1 and e_2.

$$\begin{aligned} b_k^J &= \frac{1}{1-C}\left(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1\right) \\ u^J &= \frac{1}{1-C} u^1 u^2 \end{aligned} \tag{6}$$

where $C = \sum_{i \neq j} b_i^1 b_j^2$ refers to the amount of conflicts between e_1 and e_2. Eq. (6) ensures that: (i)

when both e_1 and e_2 have high uncertainty (i.e., large u^1 and u^2), the final prediction has a low belief mass (i.e., small b_k^J), because if all u^i are relatively large, all b_k are relatively small due to (1) and (6); (ii) if e_1 and e_2 have high belief mass, the final prediction may have high confidence, or the uncertainty of the final prediction may increase because they can support different categories; (iii) when e_1 (or e_2) has high uncertainty and e_2 (or e_1) has low uncertainty, the final prediction depends on e_2 (or e_1). Due to the transitivity and commutativity of the evidential fusion rule, (6) is easily generalized to fuse $N\,(N \geq 3)$ different views of evidence.

Furthermore, by applying subjective logic theory, the joint Dirichlet strength, evidence, and parameters of the Dirichlet distribution after fusion are as follows:

$$S^J = K/u^J,\ e_k^J = b_k^J \times S^J,\ and\ \alpha_k^J = e_k^J + 1 \quad (7)$$

Compared with softmax, the evidential fusion rule is a more reliable method for combining different views of evidence. This is because it not only provides belief masses that are closely related to category probability, but it also provides an overall uncertainty (uncertainty mass). Furthermore, the evidential fusion rule maintains the evidential-based uncertainty estimation, as shown in (6) and (7), which facilitates easy computation, trustworthy decision-making, and intuitive human interpretation. In contrast, softmax-based evidential fusion outcomes are unreliable and unexplainable by adding and averaging the prediction probabilities of different views of evidence. This is because softmax-based models may overfit the data, and the probability distribution forcibly outputted for each piece of evidence may not necessarily be a true distribution.

3.3. Evidential Deep Learning for Multi-View Evidence

The CSR datasets in the experiments were designed for a multiple-choice task in which each sample had only one true label. Therefore, cross-entropy (CE) loss was utilized to compute the classification error between the target and the output logits.

Using the evidential fusion rule and subjective logic theory, it is easy to obtain the value of parameter $\boldsymbol{\alpha} = \left\{\alpha^1, \alpha^2 \ldots \alpha^v, \alpha^J\right\}$ of the Dirichlet distribution for v views of evidence and multi-view joint evidence after fusion. Thus, by applying the probability density function of the Dirichlet distribution to the CE loss, the loss function for a single view of evidence is:

$$\mathcal{L}_{ce}\left(\boldsymbol{\alpha_i^v}\right) = \int \left[\sum_{j=1}^{K} -y_{ij}\log(p_{ij})\right] \frac{1}{\mathrm{B}(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} dp_i \quad (8)$$

where K denotes the number of categories, $y_i.$ is the true label distribution, and p_{ij} represents the probability of classifying sample i into category j. As $\mathcal{L}_{ce}\left(\boldsymbol{\alpha_i^v}\right)$ decreases, $p_{ij'}$ tends to approach 1 if j' is the true class label. This causes $\alpha_{ij'}$ to gradually increase due to the relationship $p_{ij'} = \alpha_{ij'}/S$. This increase in $\alpha_{ij'}$ results in a greater amount of supporting evidence with $e_{ij'} = \alpha_{ij'} - 1$. Therefore, decreasing $\mathcal{L}_{ce}\left(\boldsymbol{\alpha_i^v}\right)$ generates more pieces of evidence for the correct label.

Let us delve into how (8) incorporates uncertainty into the loss function and why it benefits the model updates. Firstly, because the overall uncertainty $u = K/S$ and K is a constant, u and S are inversely proportional. As $S = \sum_k \alpha_k$, u is inversely proportional to α_k. Thus, $p_{ij}^{\alpha_{ij}-1} \propto p_{ij}^{1/u}$. When u is small, $1/u$ becomes large, and $p_{ij}^{1/u}$ tends toward 0. This indicates that the contribution of samples with high confidence to the total loss is relatively minor, and therefore they provide little useful information for model optimization as their prediction results are already precise or certain. In contrast, when a sample's prediction has a large u, $1/u$ tends toward 0, and $p_{ij}^{1/u}$ approaches 1. Consequently, samples with low confidence have a significant impact on the total loss and influence the parameter optimization of the model. Hence, the model is optimized toward reducing the uncertainty of all samples and making the overall predictions more reliable.

Although (8) ensures more evidence for the correct labels, it does not reduce the amount of evidence for the wrong labels. It keeps $p_{ij'}$ closer to the true label $y_{ij'} = 1$ without continuously reducing p_{ik} to $y_{ik} = 0\ (k \neq j')$. Hence, (8) fails to address situations in which evidence has a high belief mass but supports different categories. This is because evidence can be mistakenly assigned to incorrect categories. However, in a single-label classification task, only one category is correct. To eliminate evidence that supports incorrect labels, KL divergence is incorporated into the loss function.

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}} &= y_i + (1 - y_i) \odot \boldsymbol{\alpha_i} \\ &= [e_0 + 1, e_1 + 1, \ldots 1_{j'}, \ldots e_{K-1} + 1]^T\end{aligned} \quad (9)$$

$$\begin{aligned}&\mathrm{KL}\left[\mathrm{D}\left(p_i \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}\right) || \,\mathrm{D}\left(p_i \mid \mathbf{1}\right)\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}\right) + \sum_{j=1}^{K}\left(\tilde{\alpha}_{ij} - 1\right) \times \\ &\quad \left[\psi\left(\tilde{\alpha}_{ij}\right) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)\right]\end{aligned} \quad (10)$$

$$\mathcal{L}\left(\boldsymbol{\alpha_i^v}\right) = \mathcal{L}_{ce}(\boldsymbol{\alpha_i^v}) + \lambda\,\mathrm{KL}\left[\mathrm{D}\left(p_i \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}\right) || \,\mathrm{D}\left(p_i \mid \mathbf{1}\right)\right] \quad (11)$$

where $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}$ is an adjusted Dirichlet parameter with correct evidence removed and incorrect evidence

kept. ψ is the digamma function. Γ denotes the natural logarithm of the gamma function. In (10), $\mathbf{1}$ is an all ones vector with the same size of $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}$. The KL divergence as a part of the loss function moves the adjusted Dirichlet distribution $\mathrm{D}\left(p_i \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}\right)$ close to the uniform Dirichlet distribution $\mathrm{D}\left(p_i \mid \mathbf{1}\right)$, which forces the number of pieces of evidence for incorrect labels toward 0. In (11), $\lambda \ \ (0 < \lambda < 1)$ is a balancing coefficient that gradually increases from 0 toward 1 during the training process. This is because if KL divergence gets early attention, the correct label may not receive sufficient evidence, which would lead to almost equal evidence for all categories. Because $p_{ij} = \alpha_{ij}/S = (e_{ij}+1)/S$, when the values of e_{ij} $(j = 1, \ldots, K)$ are similar, the probabilities of the categories exhibit a uniform distribution.

To tackle situations where all views of explicit evidence are in low quality or unreliable (i.e., high uncertainty), the CIDN network is proposed to predict a directional variable as implicit evidence. This study assumes that critical logic for similar causal inference is directionally consistent in the reasoning space and is context-independent. Therefore, a Transformer-based network is constructed to forecast the direction of reasoning, and integrating the direction with the context as a third view of evidence. This approach is based on the consideration that while contexts and scenes under the analogous commonsense reasoning can be diverse, the underlying inference logic or reasoning direction maintains a certain level of stability. To train the CIDN effectively, this study employs contrastive loss $\mathcal{L}_{cl}(q_i)$, which aims to maximize the similarity between the integrated commonsense question context q with reasoning direction d and the correct option c^+, that is $q + d \rightsquigarrow c^+$, while minimizing the similarity with all other options c^-.

$$\mathcal{L}_{cl}(q_i) = -\log \frac{\exp(\mathsf{sim}(q_i^d, c_i^+)/\tau)}{\sum_{c_i^-} \exp(\mathsf{sim}(q_i^d, c_i^-)/\tau)} \tag{12}$$

where q_i^d denotes the sum of feature representations of q and d for sample i.

To harness multiple views of evidence and enhance the overall performance in CSR, this study employs a multi-task learning approach.

$$\mathcal{L}_{total} = \sum_{i=1}^{N} \left[\mathcal{L}\left(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{J}}\right) + \sum_{v=1}^{V} \mathcal{L}\left(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{v}}\right) + \mathcal{L}_{cl}(q_i) \right] \tag{13}$$

where $\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{J}}$ and $\mathcal{L}\left(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{J}}\right)$ represent joint parameters of the Dirichlet distribution and the global loss after evidential fusion, respectively. $\sum_{v=1}^{V} \mathcal{L}\left(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{v}}\right)$ denotes the sum of individual losses for each view of evidence.

Table 1: Evaluate the performance of models that perform hard rationale extraction. All models, except those marked with (u), are supervised at the token level of rationale. The symbol † denotes results obtained from DeYoung et al. (2020). Additionally, the human agreement on rationale extraction is included as a performance reference.

	Accuracy	F1
e-SNLI		
Lei et al. (2016) †	-	0.692
Lei et al. (2016) (u) †	-	0.379
Bert-To-Bert (2020) †	-	0.701
Rationaler	0.731	0.711
Human †	0.812±0.15	0.799±0.13
CoS-E v1.0		
Lei et al. (2016) †	0.477	0.331
Lei et al. (2016) (u) †	0.476	0.000
Bert-To-Bert (2020) †	0.344	0.519
Rationaler	0.640	0.535
Human †	0.626±0.32	0.654±0.32

4. Experimental Results and Discussion

Experiments are conducted on three datasets: CoS-E v1.0, CoS-E v1.11 (Rajani et al., 2019), and e-SNLI (Camburu et al., 2018).

4.1. Experiment for Rationale Extraction

In Table 1, various models that perform the discrete selection of rationales are evaluated by matching predicted rationales with reference rationales. The proposed *Rationaler* surpasses the performance of other methods by at least 1% in terms of token-level F1 score.

Due to the fact that CoS-E is built from crowdsourcing, which adds diversity of perspective to the dataset and inevitably introduces noise, data cleaning (DC) strategies are employed to control the quality of rationale extraction during the training stage: (i) this study removes all samples in which all the tokens in the input question are labeled as rationales, as the model may not be able to determine which part of the input sequence is most relevant to the classification target; (ii) this study removes all samples in which only one token belonging to the group of stop words is labeled as the rationale, as the token may not provide any additional useful information to the downstream classification task. Table 2 presents an example that compares the number of rationales before and after DC.

After pretraining the *Rationaler* model on the Movie Viewer dataset (DeYoung et al., 2020), this study attempts to transfer it to the CoS-E dataset,

Table 2: Number of rationales compared before and after DC on CoS-E v1.0.

CoS-E	# of rationales before DC	# of rationales after DC
Train	43665	28710
Val	5460	3495

but no significant improvement is observed in performance. Experiments also incorporate additional layers such as Bi-LSTM, GRU, and multiple linear layers after the BERT model in *Rationaler*, but none of these modifications result in any obvious improvement. Furthermore, in some instances, *Rationaler* does not select any tokens in the input sequences as rationales, indicating that there may be insufficient important information pertaining to classification objectives in those samples.

4.2. Experiment on Commonsense Reasoning

4.2.1. Evaluation of CSR Performance by Single-View Evidence

Table 3 presents a comparison of the performances of different views of evidence used in CSR. It shows the results of experiments that were conducted using the BERT baseline with questions, only explicit evidence (i.e., explanations or rationales), or both question and evidence as input. For both the CoS-E and e-SNLI datasets, human-annotated open-ended explanations and human-extracted rationales are utilized as ground-truth evidence. To provide a comprehensive evaluation, this study also experimented with generated explanations and extracted rationales at both the training and inference stages. Specifically, ChatGLM-6B (Zeng et al., 2023; Du et al., 2022) and *Rationaler* are used to generate explanations for and to extract rationales from CQA samples, respectively. In the cases where *Rationaler* does not select any tokens as rationales for a sample, the entire document (i.e., sample sentence) is used as the rationale. Moreover, the quality of explanation generated by four large LMs (LLMs) are evaluated by executing CQA task taking only explanation as input to BERT model, as shown in Table 4.

Table 3 indicates that using manually annotated evidence generally results in better performance than using extracted or generated evidence, which highlights the importance of high-quality evidence in improving the performance of CSR. Previous research by Camburu et al. (2018) also reported a 2% decline in accuracy on e-SNLI when their model was trained on human explanations and tested on generated explanations alone. In Table 3, using *only rationale* or *only explanation* refers to replacing the question with the rationale or explanation as the input to the CRM model, which enables us to measure the performance of different views of evidence. Moreover, the accuracy of using only the rationale (or explanation) is lower than that of using the question and evidence together *quest.+ral.* (or *quest.+expl.*), which indicates that the rationale and explanation can only provide a basis for guiding CRMs and cannot replace the question completely.

Table 3 shows that, in most cases, using *quest.+expl.* (or *only explanation*) performs significantly better than using *quest.+ral.* (or *only rationale*). The reason for this is that an explanation usually summarizes the general laws of things' development, which can provide the necessary association between the question and the answer. On the other hand, although rationales highlight important snippets in the question, they may not be sufficient to guide the model toward the correct answer (DeYoung et al., 2020). Thus, generalized evidence is more favorable for CSR.

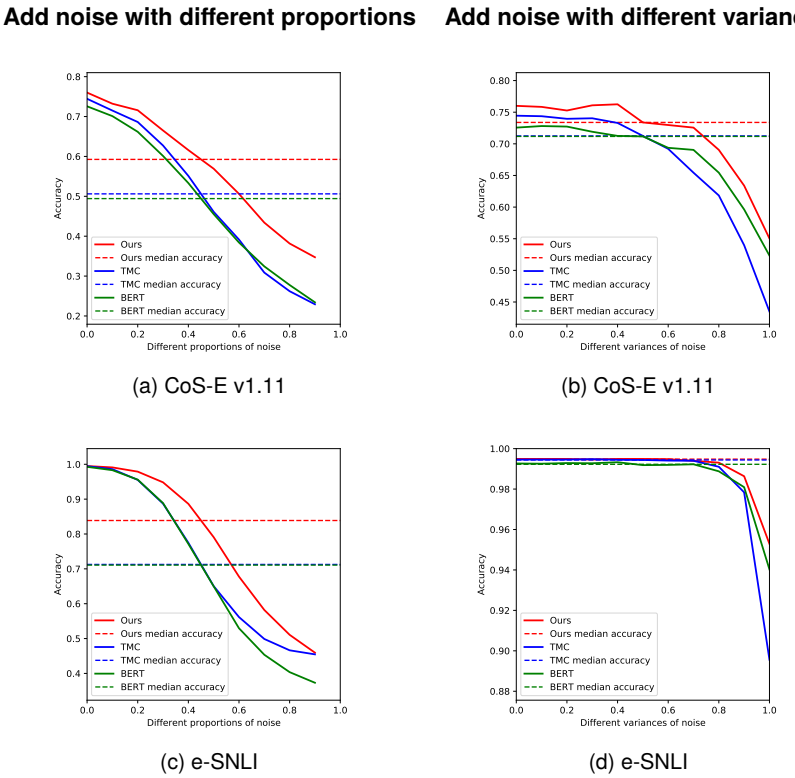
4.2.2. Evaluation of Effectiveness of Uncertainty-based Fusion

Tables 5 and 6 display that simply splicing the rationale and explanation as unified evidence (*+ral.+expl.*) as the input of BERT cannot guarantee a gain in accuracy compared with using single-view evidence. The proposed method bears out the trusted fusing different views of evidence reduces overall uncertainty, raises utilization rate of valid evidence, and finally enhances reasoning performance. This study also evaluates the quality of the extracted rationale and generated explanation by the human evaluation (HE) procedure on all datasets. Although the explanation generated by the LM is not better than that given by humans, the model learned to capture some of the built-in characteristics.

To further validate the effectiveness of the proposed method in fusing diverse views, this study adds different proportions of noise or Gaussian noise with different standard variances (*i.e.*, $\sigma = 0.1 \sim 1.0$) to a random view for each sample. As shown in Fig. 3, when the data is free of noise (i.e., noisy proportion or standard variance is 0.0), the method achieves competitive results. The accuracy of all methods decrease during adding noise to the data. Benefiting from the uncertainty-based fusion, the method perceives the noisy view, thereby limiting the representation of views containing noise and highlighting the effectiveness of noiseless views in the final prediction.

Table 3: Evaluating CSR performance using single-view evidence. The results are obtained on the CoS-E dev and the e-SNLI test datasets, where evidence is used during both training and evaluation. The symbols *+ral.* / *+expl.* refer to using a rationale or an explanation as single-view evidence, respectively. The ground truth (GT) involves using a rationale or an explanation that is annotated by humans, while the N-GT means evidence is generated or extracted using the LM model or the proposed *Rationaler*.

	Acc. on CoS-E v1.11		Acc. on CoS-E v1.0		Acc. on e-SNLI	
	N-GT	GT	N-GT	GT	N-GT	GT
only quest.	53.73		62.95		88.95	
only ral.	40.54	42.01	52.11	54.42	71.59	92.76
only expl.	52.99	65.11	63.60	76.84	77.36	97.47
quest. + ral.	54.43	53.40	60.11	63.16	**89.69**	97.21
quest. + expl.	**55.12**	**72.89**	**66.15**	**82.00**	89.34	**98.53**



(a) CoS-E v1.11 (b) CoS-E v1.11

(c) e-SNLI (d) e-SNLI

Figure 3: Performance comparison with different proportions (a, c) and variances (b, d) of noise.

Table 4: Comparing the quality of explanations generated by four LLMs using zero shot learning. #param denotes scale of model parameters.

LLM (#param)	CoS-E v1.11	e-SNLI
Openai-GPT (110M)	27.85	35.07
GPT2-medium (345M)	44.64	43.45
GPT2-large (774M)	47.01	50.22
ChatGLM (6B)	52.99	77.36

4.2.3. Uncertainty Estimation

To evaluate the estimated uncertainty, this study visualizes the distribution of in-/out-of-distribution samples in terms of uncertainty. The original samples and the samples with noise are considered as in-distribution and out-of-distribution samples, respectively. Specifically, this study replaces tokens of a random single view or all views with random words from the vocabulary (Fig. 4), or adds Gaussian noise with different standard variances (Fig. 5), to 50% of the test samples. The experimental results on CoS-E and e-SNLI are shown in the Figs. 4 and 5. From the results, the following observations are drawn: (1) Datasets with higher prediction accuracy (e.g., e-SNLI) are usually associated with lower uncertainty. (2) In contrast, datasets with lower accuracy are usually associated with higher

Table 5: Enhancing CSR performance by evidential fusing multi-view evidence. *PreSoTA* represents the best score previously achieved on each dataset by single model (Xu et al., 2021; Narang et al., 2020). *+ral.+expl.* means multi-view evidence is concatenated. *+TMVEF* involves the fusion of different views of evidence under the TMVEF framework. TMC means a fusion approach by extending EDL (Sensoy et al., 2018) with TMVEF but removing CIDN that forecasts the direction of reasoning. † denotes results obtained from Xu et al. (2021), Narang et al. (2020) and Rajani et al. (2019).

	Acc. on CoS-E v1.0			Acc. on e-SNLI		
	N-GT	GT	HE	N-GT	GT	HE
PreSoTA†	-	83.70	-	-	91.60	-
Human†	-	80.40	16.00	-	89.00	78.00
WT5-Base†	-	59.40	-	-	90.90	-
WT5-11B†	-	82.70	30	-	92.30	**90**
EDL (multi-view)	63.89	83.05	-	87.88	99.23	-
TMC (EDL+TMVEF)	66.78	83.47	-	**89.59**	99.46	-
BERT (quest. + ral. + expl.)	66.56	82.74	-	89.50	99.27	-
Ours (quest. + ral. + expl. + TMVEF)	**67.00**	**85.37**	**33.33**	89.54	**99.49**	61.11

Table 6: Results on CoS-E v1.11 dev-random-split. BERT (MV) means using BERT alongside majority voting to train multiple BERT models on multi-view evidence and then combine their predictions. † denotes results obtained from Zelikman et al. (2022) and Wei et al. (2022).

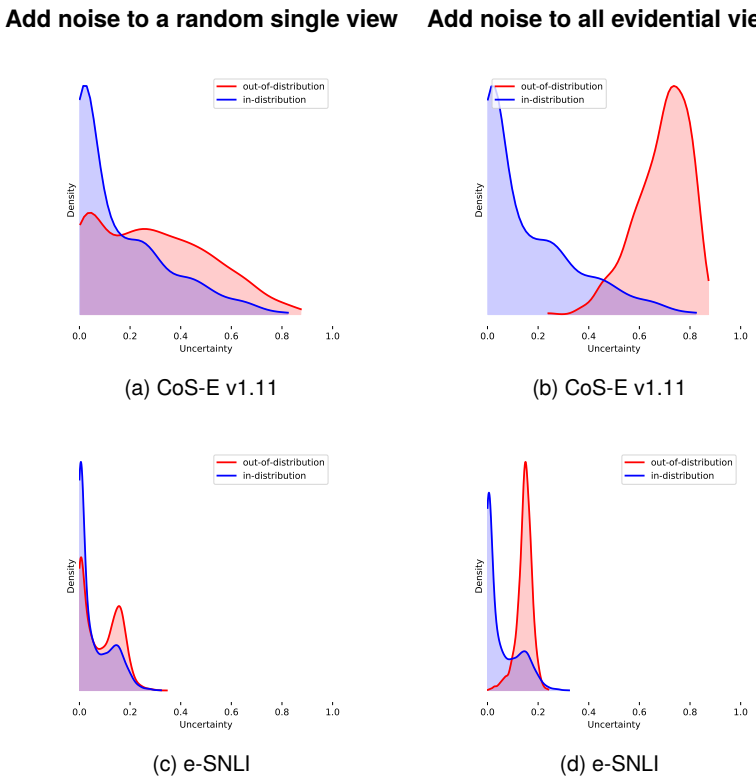
	Acc. on CoS-E v1.11			
	N-GT	GT	#param	throughput (#item/s)
EDL	53.32	70.29	108M	43.68
TMC	58.56	74.45	108M	41.66
BERT	57.00	72.74	108M	**173.63**
BERT (MV)	54.27	72.56	108M	154.99
STaR†	-	72.50	6B	-
GPT3 (CoT)†	-	73.50	175B	-
Ours	**59.79**	**76.00**	108M	27.15

uncertainty. (3) Much higher uncertainties are usually estimated for out-of-distribution samples. (4) As the noise expands from a random single view to all views or from small variance to high variance, the uncertainty of the data will increase. These observations suggest that the proposed method is effective at perceiving uncertainty due to its reasonable decision by assigning relatively lower uncertainty to in-distribution samples and higher uncertainty to out-of-distribution samples. Fig. 6 compares the method with TMC in terms of providing much more accurate predictions as the predicted uncertainty decreases. This implies trusted inferences are supported by the output of the model, which include classification accuracy and uncertainty.

5. Conclusion

This study proposes a TMVEF framework for enhancing the performance of CSR tasks through the reliable combination of different views of evidence. The TMVEF has three merits, including (i) directly modeling uncertainty as a metric to dynamically evaluate evidential quality, (ii) associating the belief masses of different classes with the Dirichlet distribution and optimizing the CRM by the uncertainty of evidence, and (iii) combining different views of evidence in an interpretable and theoretically supported manner. After evidential fusion, an overall uncertainty and a joint belief mass can be calculated to represent the credibility of the final prediction.

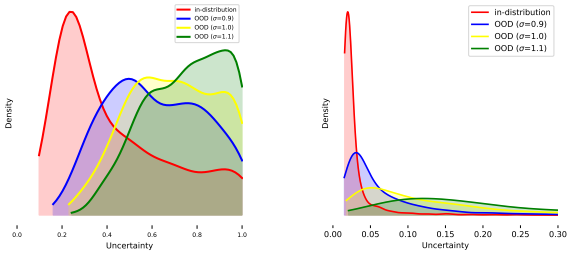
The experiments yielded the following conclusions: (i) Different views of evidence provide varying degrees of support for CSR. (ii) A single view of evidence may not be fully utilized by the model, and different views of evidence may conflict with each other by underpinning different categories. (iii) Concatenating evidence with input questions may not always improve CQA performance on any dataset due to noise, which may distract the downstream classifier from capturing key clues for reliable reasoning. (iv) The proposed TMVEF framework combines different views of evidence in an interpretable and reliable manner, which reduces uncertainty, improves valid evidence utilization, and ultimately enhances inference performance. These findings suggest that although many NLP models approach human performance, further research is needed to better understand how and why they make their predictions. Hopefully, this study on different views of evidence and the fusion method can provide valuable insights for CSR and facilitate further ex-



(a) CoS-E v1.11 (b) CoS-E v1.11

(c) e-SNLI (d) e-SNLI

Figure 4: Density of uncertainty on 50% noisy test data and 50% normal test data.



(a) CoS-E v1.11 (b) e-SNLI

Figure 5: Density of uncertainty with different variances of noise.

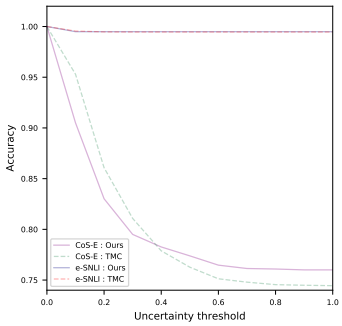


Figure 6: Accuracy with uncertainty thresholding on CoS-E v1.11 and e-SNLI.

ploration of model interpretability.

6. Limitations

This study proposes the TMVEF for CSR. TMVEF combines uncertain or even conflicting evidence in an interpretable and learnable manner and employs multi-task learning strategy to enhance the overall performance for CQA. From experiments, TMVEF is more effective in situations with low-confidence prediction. For example, the performance gain on e-SNLI is not obvious like that of CoS-E. This is because all views of evidence in e-SNLI are already reliable and sufficient (see Table 3), and further evidential fusion would not bring additional benefits. Furthermore, the TMVEF framework does not add model parameters (see Table 6), but due to the need to calculate the uncertainty of various views of evidence and execute evidential fusion, the time cost increases. In addition, a rationale extractor *Rationaler* is proposed to identify important snippets in the input context. It has the limitation that only handle with sentence-level rationale extraction. As applying the proposed framework to document-level datasets, new models will be investigated for extracting rationales at higher level of granularity.

7. Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No.: 61802079), the Guangzhou Science and Technology Project, China (Grant No.: 202102020564), the Guangzhou University Grant (No.: 69-62396702) and the China Scholarship Fund. Thanks to the referees for their comments, which helped improve this paper considerably.

8. Bibliographical References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21)*, pages 3050–3065, virtual. Association for Computational Linguistics.

Guy Aglionby and Simone Teufel. 2022. Faithful knowledge graph explanations in commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, pages 10811–10817, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. 2020. Depth uncertainty in neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, pages 1–15, Red Hook, NY, USA. Curran Associates Inc.

Elham J. Barezi and Pascale Fung. 2019. Modality-based factorization for multimodal fusion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 260–269, Florence, Italy. Association for Computational Linguistics.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, volume 36, pages 12317–12325, virtual. AAAI Press.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Proceddings of the 32nd Conference on Neural Information Processing Systems (NIPS'18)*, volume 31, pages 1–11, Montréal, Canada. Curran Associates, Inc.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2020. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, pages 1–12, Red Hook, NY, USA. Curran Associates Inc.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'22)*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 4443–4458, virtual. Association for Computational Linguistics.

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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, virtual. Association for Computational Linguistics.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted multi-view classification. In *International Conference on Learning Representations (ICLR'21)*, pages 1–11, virtual.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2023. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transaction on Pattern Analysis and Machine Intelligence.*, 45(2):2551–2566.

Audun Jøsang. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, volume 4. Springer.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18)*, volume 32, pages 5198–5204, Louisiana, USA. AAAI Press.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

S. Le Hegarat-Mascle, I. Bloch, and D. Vidal-Madjar. 1997. Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):1018–1031.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'19)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, volume 36, pages 10947–10955, virtual. AAAI Press.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020a. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, virtual. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020b. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Xuan Ma, Xiaoshan Yang, and Changsheng Xu. 2023. Multi-source knowledge reasoning graph network for multi-modal commonsense inference. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(4):1–17.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, pages 1–16.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, page 3183–3193, Red Hook, NY, USA. Curran Associates Inc.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, pages 1–11, virtual.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023a. PINTO: Faithful language reasoning using prompted-generated rationales. In *International Conference on Learning Representations (ICLR'23)*, pages 1–17, Kigali Rwanda.

Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. 2019. Multi-view clustering via late fusion alignment maximization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'19)*, pages 3778–3784, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. Cat: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, page 13111–13140, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Thirty-sixth Conference on Neural Information Processing Systems (NIPS'22)*, volume 35, pages 24824–24837, Louisiana, USA. Curran Associates, Inc.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207, virtual. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Thirty-sixth Conference on Neural Information Processing Systems (NIPS'22)*, volume 35, pages 15476–15488, Louisiana, USA. Curran Associates, Inc.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR'23)*, pages 1–56, Kigali Rwanda.

Tiejian Zhang, Xinwang Liu, Lei Gong, Siwei Wang, Xin Niu, and Li Shen. 2023. Late fusion multiple kernel clustering with local kernel alignment maximization. *IEEE Transactions on Multimedia*, 25:993–1007.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 14883–14896, Toronto, Canada. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI'20)*, pages 9733–9740, New York, USA. AAAI Press.

9. Language Resource References

Camburu, Oana-Maria and Rocktäschel, Tim and Lukasiewicz, Thomas and Blunsom, Phil. 2018. *e-SNLI*. Curran Associates, Inc. PID https://github.com/OanaMariaCamburu/e-SNLI.

Rajani, Nazneen Fatema and McCann, Bryan and Xiong, Caiming and Socher, Richard. 2019. *CoS-E*. Association for Computational Linguistics. PID https://github.com/salesforce/cos-e.