Emphasising Structured Information: Integrating Abstract Meaning Representation into LLMs for Enhanced Open-Domain Dialogue Evaluation

Anonymous ACL submission

Abstract

Automatic open-domain dialogue evaluation has attracted increasing attention, yet remains challenging due to the complexity of assessing response appropriateness. Traditional evaluation metrics, typically trained with true positive and randomly selected negative responses, tend to assign higher scores to responses that share greater content similarity with contexts. However, adversarial negative responses, despite possessing high lexical overlap with contexts, can be semantically incongruous. Consequently, existing metrics struggle to evaluate such responses effectively, resulting in low correlations with human judgments. While recent studies have demonstrated the effectiveness of Large Language Models (LLMs) for open-domain dialogue evaluation, they still face challenges in handling adversarial negative examples. We propose a novel evaluation framework that integrates Abstract Meaning Representation (AMR) enhanced domain-specific language models (SLMs) with LLMs. Our SLMs explicitly incorporate AMR graph information through a gating mechanism for enhanced semantic representation learning, while both SLM predictions and AMR knowledge are integrated into LLM prompts for robust evaluation. Extensive experiments on open-domain dialogue evaluation tasks demonstrate the superiority of our method compared to state-of-the-art baselines, particularly in discriminating adversarial negative responses. Our framework achieves strong correlations with human judgments across multiple datasets, establishing a new benchmark for dialogue evaluation. Our code and data are publicly available.

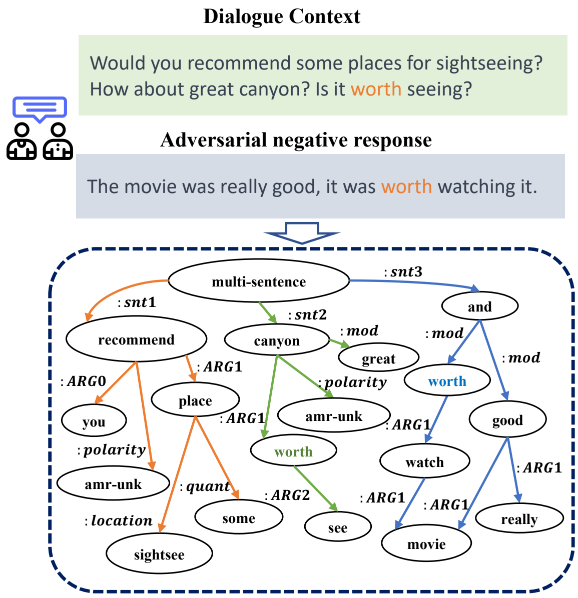

Figure 1: AMR graphs for the conversational context and response. The semantic relationship of the word "worth" appearing in both context and response is captured through distinct colored representations in their respective AMR graphs.

1 Introduction

Open-domain dialogue systems have garnered substantial attention owing to their broad applicability (Zhang et al., 2021; Liu et al., 2023) across various domains, including personal medical assistance and biomedical telecommunications (Sai et al., 2020; Yang et al., 2024). Traditional evaluation approaches, such as n-gram-based metrics (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) and embedding-based metrics (Zhang et al., 2020), assess the semantic similarity between response candidates and gold references. these methods correlate poorly with human evaluation due to their limited capacity to incorporate conversational context (Liu et al., 2016).

While recent advances in trainable evaluation frameworks (Lowe et al., 2017; Tao et al., 2018) have improved context-response relationship modeling, they face fundamental limitations stemming from their training . These models, typically trained with true positive and randomly sampled negative examples, tend to assess responses primarily through surface-level content similarity. Although some approaches have attempted to address this by incorporating adversarial examples (Sai et al., 2020; Gupta et al., 2021), they either require extensive pre-training on large-scale conversational corpora or demand adaptation to specific datasets, incurring substantial computational overhead. Moreover, their exclusive reliance on surface-form features compromises robustness when evaluating adversarial examples that deviate from the training distribution.

The vulnerability to adversarial attacks further compounds this challenge. Jin et al. (2019) demonstrated that even simple synonym substitutions can lead to misclassification in text analysis tasks. For instance, a positive review stating "*The characters, cast in impossibly contrived situations, are totally estranged from reality*" would be misclassified as negative when minimally modified to "*The characters, cast in impossibly engineered circumstances, are fully estranged from reality*", despite maintaining semantic equivalence.

Recent advances in Large Language Models (LLMs) have shown promise in dialogue evaluation (Liu et al., 2023; Kocmi and Federmann, 2023; Chiang and yi Lee, 2023). However, these models still exhibit suboptimal performance when evaluating adversarial negative responses. To address these limitations, we propose integrating LLMs with domain-specific language models (SLMs) enhanced by Abstract Meaning Representation (AMR) graph information, specifically aimed at improving evaluation robustness for adversarial examples. AMR graphs serve as powerful tools for capturing dialogue system states and providing complementary semantic knowledge (Bai et al., 2021; Bonial et al., 2020). Consider the following example: given the context "*Would you recommend some places for sightseeing? How about great canyon? Is it worth seeing?*", and an adversarial negative response "*The movie was really good, it was worth watching it*", existing metrics might erroneously classify this as positive due to lexical overlap. AMR graphs help address this by modeling semantic relationships between concepts (e.g., "worth" and "canyon") through explicit edge relations (e.g., ":mod" and ":ARG1").

Our approach introduces an AMR graph-enhanced SLM that effectively identifies adversarial negative examples in open-domain dialogue. The framework integrates both the SLM's predictions and AMR graph information into the LLM's prompt, creating a robust automatic evaluator that leverages domain-specific knowledge during inference. The SLM architecture comprises two key components: sentence and graph encoders. The sentence encoder processes surface-form knowledge from conversational contexts and responses, while the graph encoder models AMR structural information, capturing both conceptual elements and their interrelations. These complementary representations are unified through a sophisticated gating mechanism and optimised via contrastive learning, encouraging alignment between textual and structural features for positive context-response pairs. The final evaluation integrates both the SLM's prediction score and AMR graph information into the LLM's prompt.

Comprehensive empirical evaluation across three public datasets demonstrates our model's superior performance compared to state-of-the-art baselines, including LLM-based methods. Our key contributions include:

Our contributions can be summarised as follows:

- The framework to integrate AMR graph information into open-domain dialogue evaluation through a novel combination of enhanced SLMs and LLMs.
- A dual-representation approach that leverages both surface-form and semantic graph information, with LLM capabilities enhanced by SLM predictions and AMR knowledge.
- Comprehensive experimental results demonstrating substantial improvements over existing methods, particularly in evaluating challenging adversarial negative responses.

2 Related Work

Dialogue Evaluation Metrics. Traditional n-gram-based metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), compute lexical overlap between response candidates and gold references. More sophisticated embedding-based metrics, such as Extrema (Forgues and Pineau, 2014) and BERTScore (Zhang et al., 2020), first project responses and references into high-dimensional semantic spaces before calculating their similarity. However, both approaches have shown limited efficacy in evaluating open-domain dialogue systems (Liu et al., 2016).

Regarding trainable metrics, RUBER (Tao et al., 2018) evaluates response quality by measuring semantic similarity between the generated response, dialogue context, and ground truth reference. Sai et al. (2020) introduced DEB, which leverages a BERT model pre-trained on large-scale Reddit conversations. While effective, the computational cost of pre-training on extensive datasets makes this approach less practical. Similarly, Mask-and-fill (Gupta et al., 2021) employs a Speaker-Aware BERT architecture (Gu et al., 2020) to enhance dialogue understanding, though it requires dataset-specific adaptation before fine-tuning. Zhang et al. (2021) developed MDD-Eval for cross-domain dialogue evaluation, but this method necessitates human annotations and additional training data while failing to address adversarial negative examples.

LLM-based Evaluators. The emergence of Large Language Models (LLMs) has enabled new approaches to dialogue evaluation. Fu et al. (2023) developed GPTScore, leveraging pre-trained language models for multi-aspect, customizable evaluation without task-specific training. Wang et al. (2023) empirically validated the effectiveness of LLM-based evaluation approaches. Kocmi and Federmann (2023) demonstrated the utility of GPT models in machine translation evaluation. Liu et al. (2023) introduced G-Eval, employing GPT-4 across multiple generation tasks including dialogue response, text summarization, data-to-text generation, and machine translation. Chan et al. (2023) proposed ChatEval, a multi-agent debate framework that surpasses single-LLM evaluators in performance. However, these LLM-based approaches have yet to be applied to evaluating adversarial negative responses incorporating non-textual domain knowledge.

3 Methodology

3.1 Task Description

Our model operates on input tuples consisting of a dialogue context $\mathcal{C}$, a response $\mathcal{R}$, and their corresponding AMR graphs $\mathcal{G}_\mathcal{C}$ and $\mathcal{G}_\mathcal{R}$. The primary objective of the SLM component is to perform binary classification, predicting a label $\mathcal{Y} \in \{0, 1\}$ for each response, where 0 and 1 denote negative and positive responses, respectively.

The SLM generates a classification confidence score defined as:

$$\text{Score}_{\text{SLM}} = P(\mathcal{Y} \mid \mathcal{C}, \mathcal{R}, \mathcal{G}_\mathcal{C}, \mathcal{G}_\mathcal{R}) \tag{1}$$

The derived confidence score, in conjunction with the semantic structural information encoded in AMR graphs $\mathcal{G}_\mathcal{C}$ and $\mathcal{G}_\mathcal{R}$, is incorporated into the LLM's prompt. This integration enables the LLM to leverage both statistical confidence and explicit semantic knowledge for more robust open-domain dialogue evaluation.

3.2 Overall Architecture

Figure 2 illustrates the comprehensive architecture of our proposed framework, which seamlessly integrates SLM and LLM components. The SLM architecture incorporates a dual-encoder design: a sequence encoder for processing textual information and a graph encoder specialized in AMR graph representation learning. The complementary representations from these encoders are dynamically balanced through an adaptive gating mechanism, which modulates the information flow from both sources.

To optimise the alignment between textual and structural representations, particularly for positive response pairs, we employ a contrastive learning strategy during the training phase. This approach minimizes the representational distance between sentence and graph embeddings for semantically coherent pairs, while maintaining appropriate separation for negative examples.

The final evaluation framework leverages both the SLM's classification confidence score $\text{Score}_{\text{SLM}}$ and the structured AMR graph information, which are systematically integrated into the LLM's prompt through a carefully designed template. This multi-modal integration enables the LLM to synthesize both statistical and semantic evidence for more robust dialogue evaluation.

3.3 Sequence Encoder

The sequence encoder employs a standard Transformer architecture (Vaswani et al., 2017) to process the input dialogue components. Given a dialogue context $\mathcal{C}_i = \{w_1, w_2, \ldots, w_\mathcal{C}\}$ and a response $\mathcal{R}_i = \{w_1, w_2, \ldots, w_\mathcal{R}\}$, where w_i denotes the i-th token and $\mathcal{C}$, $\mathcal{R}$ represent respective sequence lengths, the encoder generates a sentence representation $\mathbf{H}_S$. The encoding process can be formally expressed as:

$$\mathbf{H}_S = \text{SeqEncoder}(\mathcal{C}, \mathcal{R}) \tag{2}$$

$$h_i = \sum_{j=1}^{\mathcal{C}+\mathcal{R}} \alpha_{ij} \left(W^H h_j\right) \tag{3}$$

$$\alpha_{ij} = \text{Attention}\left(h_i, h_j\right) \tag{4}$$

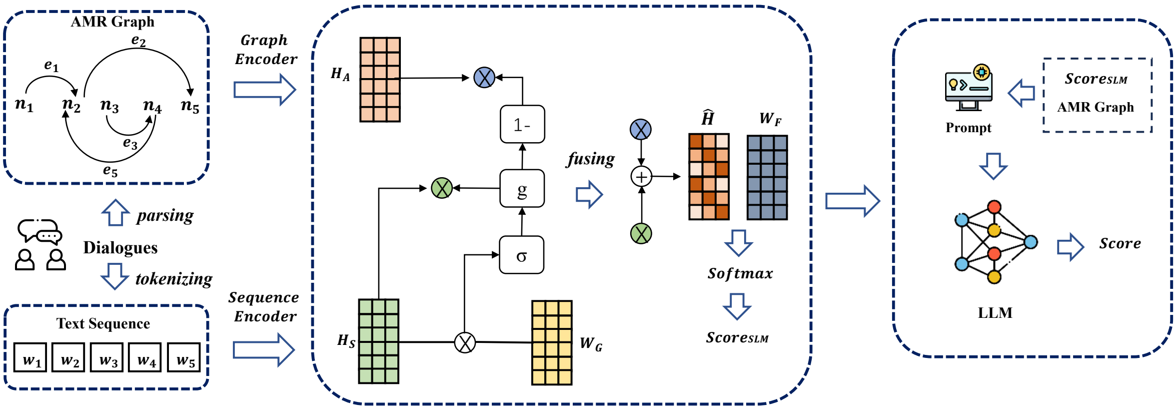

Figure 2: The architecture of the proposed model. The left part is the SLM architecture, containing two encoders and the gate mechanism for encoding and fusing the sequence and AMR graph information of context-response pairs. The right part is the LLM where the prompt contains the prediction score of the SLM and AMR graph information.

where $\mathbf{H}_S = \{h_1, h_2, \ldots, h_{\mathcal{C}+\mathcal{R}}\}$ represents the sequence of hidden states and W^H denotes the transformation matrix.

3.4 Graph Encoder

For modeling AMR graph structures, we utilise the Graph Transformer (Zhu et al., 2019), an extension of the standard Transformer that specialises in graph-structured data. An AMR graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ comprises nodes $\mathcal{V}$ and edges $\mathcal{E}$, where each edge $e \in \mathcal{E}$ is represented as a triple $\langle n_i, r_{ij}, n_j \rangle$ denoting the relation r_{ij} between nodes n_i and n_j. The graph encoding process is defined as:

$$\mathbf{H}_A = \text{GraphEncoder}(\mathcal{V}, \mathcal{E}) \tag{5}$$

$$h_i' = \sum_{j=1}^{M} \hat{\alpha}_{ij} \left(W^V h_j' + W^R \boldsymbol{r}_{ij} \right) \tag{6}$$

where $\mathbf{H}_A = \{h_1', h_2', \ldots, h_M'\}$ represents the graph embeddings, and W^V, W^R are learnable transformation matrices.

The graph attention mechanism, which distinguishes the Graph Transformer from standard Transformers, is computed as:

$$\hat{\alpha}_{ij} = \frac{\exp\left(\hat{e}_{ij}\right)}{\sum_{m=1}^{M} \exp\left(\hat{e}_{im}\right)}$$

$$\hat{e}_{ij} = \frac{\left(W^Q h_i'\right)^T \left(W^K h_j' + W^R \boldsymbol{r}_{ij}\right)}{\sqrt{d}} \tag{7}$$

where W^Q, W^K are transformation matrices and d is the dimensionality of the hidden states.

3.5 Aggregation Gate

To effectively combine the complementary information from both sequence and graph representations, we implement an adaptive gating mechanism. Given the sentence representation $\mathbf{H}_S$ and graph representation $\mathbf{H}_A$, the gate value g_i is computed as:

$$g_i = \sigma\left(W^G \mathbf{H}_S + b_g\right) \tag{8}$$

$$\hat{\mathbf{H}} = g_i \mathbf{H}_S + (1 - g_i)\, \mathbf{H}_A \tag{9}$$

where W^G, b_g are learnable parameters, and $\hat{\mathbf{H}}$ represents the final fused representation.

3.6 Training objectives and Evaluation

The fused representation $\hat{\mathbf{H}}$ is used to predict the classification probability for the context-response pair:

$$\text{Score}_{\text{SLM}} = \text{softmax}\left(W^F \hat{\mathbf{H}} + b_f\right) \tag{10}$$

The training objective combines classification and contrastive learning:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_C \tag{11}$$

$$\mathcal{L}_{cls} = -\log P(\mathcal{Y} = 1 \mid \hat{\mathbf{H}}) \tag{12}$$

The contrastive loss $\mathcal{L}_C$, inspired by Henderson et al. (2017), facilitates alignment between sentence and graph representations:

$$\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N} \frac{e^{\mathrm{sim}\left(\mathbf{H}_S^+, \mathbf{H}_A^+\right)}}{\sum_j e^{\mathrm{sim}\left(\mathbf{H}_S^-, \mathbf{H}_A^-\right)}} \tag{13}$$

where $\mathbf{H}_S^+$, $\mathbf{H}_A^+$ denote positive pair representations and $\mathbf{H}_S^-$, $\mathbf{H}_A^-$ represent negative pairs.

The final evaluation score integrates the SLM prediction score $\mathrm{Score}_{\mathrm{SLM}}$ and AMR graph information $\mathcal{G}$ through the LLM's prompt.

$$\mathrm{Score} = \mathrm{LLMs}(\mathrm{Score}_{\mathrm{SLM}}, \mathcal{G}) \tag{14}$$

4 Experiments

4.1 Dataset

We conduct experiments on three widely-recognised open-domain dialogue datasets: **DailyDialog++** (Sai et al., 2020), **PersonaChat** (Zhang et al., 2018), and **TopicalChat** (Gopalakrishnan et al., 2019). DailyDialog++ is particularly noteworthy as it is the sole publicly available dataset containing human-crafted adversarial negative responses. Each context is paired with three types responses: five positive responses, five random negative responses, and five adversarial negative responses.

For PersonaChat and TopicalChat, which lack human-created adversarial responses in their original forms, we utilise the augmented datasets from (Zhao et al., 2024). These enhanced datasets feature 2,000 conversational contexts, each accompanied by five positive responses and adversarial negative counterparts.

4.2 Experimental Settings

The preprocessing of AMR graph structures involves multiple stages. Initially, we employ the *amrlib* library (Cai and Lam, 2020) to transform each context-response pair into its corresponding AMR graph representation. Following the methodology outlined in (Song et al., 2020), we subsequently process these graphs using the AMR simplifier (Konstas et al., 2017). This procedure include the error-checking and therefore yields refined and accurate AMR graphs. For the LLM component, we utilise `GPT-3.5-turbo` and `GPT-4-1106`.

4.3 Baselines

For the word-overlap and embedding-based metrics, we select widely used ones in generative dialogue systems, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). For the learning-based metrics, We compare our method with DEB (Sai et al., 2020), USR (Mehri and Eskenazi, 2020), Mask-and-fill (Gupta et al., 2021), and MDD-Eval (Zhang et al., 2021). Additionally, we select G-Eval (Liu et al., 2023) and LLM-Eval (Lin and Chen, 2023) as the LLM-based baseline metrics.

4.4 Evaluation Set and Human Annotation

To rigorously assess our proposed metric, we establish a comprehensive evaluation protocol comprising two distinct sets: a *Standard Set* and an *Adversarial Set*.

Dataset Construction The Standard Set encompasses positive and random negative responses, with 400 context-response pairs sourced from each of DailyDialog++, PersonaChat, and TopicalChat datasets, totalling 1,200 samples. The random negative responses are selected from different dialogue turns to ensure contextual diversity. The Adversarial Set, designed to evaluate robustness against challenging examples, contains an additional 400 context-response pairs per dataset, featuring positive and adversarial negative responses. In aggregate, our evaluation corpus comprises 2,400 context-response pairs.

Evaluation Criteria Following the metrics proposed by Zhong et al. (2022), we assess responses across four dimensions: (1) Naturalness: The degree to which a response is naturally written; (2) Coherence: The extent to which the content of the output is well-structured, logical, and meaningful; (3) Engagingness: The degree to which the response is engaging; and (4) Groundedness: The extent to which a response is grounded in facts present in the context.

Human Annotation Three qualified human evaluators, each holding at least a master's degree in Computer Science and demonstrating full professional English proficiency, independently rated each context-response pair. Assessments were conducted using a 5-point Likert scale, where higher scores indicate superior quality. The final human annotation score for each aspect was derived by averaging across all evaluators. To ensure annotation reliability, we computed the Inner-Annotator Agreement (IAA) using Cohen's Kappa coefficient (Cohen, 1960). The achieved average IAA score of 0.64 between annotator pairs indicates substantial agreement (0.6-0.8), validating the ro-

Metrics	Standard Set Pearson's ρ	Standard Set Spearman's τ	Adversarial Set Pearson's ρ	Adversarial Set Spearman's τ
BLEU-1	0.1841 (0.1620)	0.1825 (0.1623)	0.2064 (0.1321)	0.2102 (0.9274)
BLEU-2	0.1881 (0.1928)	0.1772 (0.3928)	0.1540 (0.3937)	0.1969 (0.3921)
BLEU-3	0.1847 (0.4265)	0.1835 (0.3521)	0.1543 (0.4336)	0.1973 (0.2292)
BLEU-4	0.1980 (0.2552)	0.1787 (0.8398)	0.1598 (0.6175)	0.1844 (0.7698)
ROUGE-1	0.2183 (0.4698)	0.2026 (0.7390)	0.2305 (0.9120)	0.2141 (0.4276)
ROUGE-2	0.2055 (0.9153)	0.1911 (0.1263)	0.1516 (0.5291)	0.1693 (0.5201)
ROUGE-L	0.2183 (0.1028)	0.2034 (0.1928)	0.2377 (0.0183)	0.2271 (0.1912)
METEOR	0.1804 (0.1018)	0.1561 (0.1793)	0.1342 (0.1123)	0.1034 (0.5443)
BERTScore	0.2517 (0.3556)	0.2658 (0.2369)	0.2016 (0.3430)	0.2230 (0.2561)
DEB	0.3236 (0.0630)	0.2856 (0.2382)	0.3492 (0.0622)	0.3406 (0.8098)
USR	0.2636 (0.0206)	0.2482 (0.8432)	0.2297 (0.0624)	0.2760 (0.1892)
Mask-and-fill	0.1904 (0.1732)	0.2056 (0.0975)	0.2604 (0.1320)	0.2895 (0.0460)
MDD-Eval	0.2813 (0.0610)	0.2424 (0.8223)	0.2982 (0.4162)	0.2792 (0.0218)
G-Eval (`GPT-3.5`)	0.3418 (0.0106)	0.3325 (0.0190)	0.3294 (0.2327)	0.3412 (0.2272)
G-Eval (`GPT-4`)	0.4321 (0.0001)	0.4312 (0.0071)	0.4298 (0.0225)	0.4528 (0.0021)
LLM-Eval (`GPT-3.5`)	0.3548 (0.0211)	0.3723 (0.0190)	0.3501 (0.3712)	0.3421 (0.0762)
LLM-Eval (`GPT-4`)	0.4315 (0.0206)	0.4621 (0.0172)	0.4691 (0.2355)	0.4528 (0.5632)
Ours(`w/o LLM`)	0.3575 (0.0442)	0.3646 (0.0347)	0.3492 (0.0620)	0.3545 (0.0215)
Ours (`GPT-3.5 w/o AMR`)	0.4590 (0.0241)	0.4592 (0.0539)	0.4623 (0.2327)	0.4745 (0.2342)
Ours (`GPT-3.5 w/o SLM`)	0.4782 (0.1242)	0.4723 (0.0119)	0.4898 (0.2237)	0.4902 (0.0938)
Ours (`GPT-3.5`)	0.4890 (0.0001)	0.4873 (0.0019)	0.4955 (0.1237)	0.4920 (0.0462)
Ours (`GPT-4 w/o AMR`)	0.5290 (0.2421)	0.5392 (0.0129)	0.5212 (0.2375)	0.5522 (0.5632)
Ours (`GPT-4 w/o SLM`)	0.5426 (0.0106)	0.5701 (0.0019)	0.5521 (0.8375)	0.5209 (0.9472)
Ours (`GPT-4`)	**0.5693 (0.0021)**	**0.5927 (0.0043)**	**0.5628 (0.0116)**	**0.5826 (0.0025)**

Table 1: Pearson and Spearman correlations with human judgments on the DailyDialog++ dataset. The number figures in parentheses are p-values.

Metrics	Standard Set Pearson's ρ	Standard Set Spearman's τ	Adversarial Set Pearson's ρ	Adversarial Set Spearman's τ
BLEU-1	0.2063 (0.9228)	0.2152 (0.6538)	0.1764 (0.2243)	0.1663 (0.0335)
BLEU-2	0.1951 (0.7401)	0.1823 (0.1361)	0.1405 (0.3621)	0.1619 (0.1422)
BLEU-3	0.1680 (0.3465)	0.1941 (0.8264)	0.1375 (0.2103)	0.1676 (0.3456)
BLEU-4	0.2002 (0.2836)	0.1930 (0.1712)	0.1253 (0.0924)	0.1543 (0.8927)
ROUGE-1	0.2130 (0.4942)	0.2159 (0.3892)	0.2075 (0.5918)	0.2198 (0.1984)
ROUGE-2	0.2016 (0.0183)	0.2023 (0.9172)	0.1832 (0.1830)	0.2073 (0.1983)
ROUGE-L	0.2103 (0.9028)	0.2034 (0.9283)	0.2027 (0.9278)	0.2236 (0.9183)
METEOR	0.1997 (0.0183)	0.1768 (0.0918)	0.1439 (0.9214)	0.1705 (0.4028)
BERTScore	0.2865 (0.2357)	0.2721 (0.2568)	0.2254 (0.5914)	0.2643 (0.6019)
DEB	0.3653 (0.0241)	0.3434 (0.8346)	0.3512 (0.0301)	0.3706 (0.8398)
USR	0.3466 (0.0392)	0.3456 (0.1343)	0.3681 (0.0462)	0.3859 (0.1846)
MDD-Eval	0.3481 (0.0619)	0.3410 (0.1802)	0.3735 (0.1503)	0.3601 (0.9348)
Mask-and-fill	0.3093 (0.1812)	0.3105 (0.8013)	0.3764 (0.3153)	0.3613 (0.2203)
G-Eval (`GPT-3.5`)	0.4891 (0.0923)	0.4874 (0.0122)	0.4551 (0.0410)	0.4610 (0.0512)
G-Eval (`GPT-4`)	0.5241 (0.0131)	0.5313 (0.0424)	0.5123 (0.0112)	0.5513 (0.0253)
LLM-Eval (`GPT-3.5`)	0.4648 (0.1821)	0.4573 (0.9181)	0.4450 (0.7163)	0.4614 (0.7817)
LLM-Eval (`GPT-4`)	0.5321 (0.8127)	0.5392 (0.7161)	0.5269 (0.9221)	0.5258 (0.9271)
Ours(`w/o LLM`)	0.3668 (0.0044)	0.3784 (0.0037)	0.3954 (0.0060)	0.3911 (0.0055)
Ours (`GPT-3.5 w/o AMR`)	0.5007 (0.0032)	0.4998 (0.0008)	0.5011 (0.0237)	0.5105 (0.0047)
Ours (`GPT-3.5 w/o SLM`)	0.5118 (0.0024)	0.5068 (0.0038)	0.5199 (0.0007)	0.5187 (0.0005)
Ours(`GPT-3.5`)	0.5517 (0.0044)	0.5209 (0.0002)	0.5204 (0.0053)	0.5225 (0.0057)
Ours (`GPT-4 w/o AMR`)	0.6199 (0.0001)	0.6127 (0.0004)	0.6178 (0.0017)	0.6004 (0.0028)
Ours (`GPT-4 w/o SLM`)	0.6267 (0.0021)	0.6299 (0.0003)	0.6245 (0.0047)	0.6309 (0.0145)
Ours (`GPT-4`)	**0.6598 (0.0021)**	**0.6604 (0.0023)**	**0.6526 (0.0013)**	**0.6612 (0.0046)**

Table 2: Pearson and Spearman correlations with human judgments on the PersonaChat dataset.

Metrics	Standard Set Pearson's ρ	Standard Set Spearman's τ	Adversarial Set Pearson's ρ	Adversarial Set Spearman's τ
BLEU-1	0.2102 (0.2993)	0.1982 (0.8628)	0.1444 (0.0203)	0.1553 (0.0032)
BLEU-2	0.1721 (0.7761)	0.1772 (0.3132)	0.1295 (0.4321)	0.1439 (0.5402)
BLEU-3	0.1577(0.1357)	0.1642 (0.1854)	0.1225 (0.0203)	0.1328 (0.0341)
BLEU-4	0.1482 (0.2901)	0.1503(0.1709)	0.1323 (0.0203)	0.1228 (0.3265)
ROUGE-1	0.2050 (0.4808)	0.2144 (0.0371)	0.1752 (0.2839)	0.1788 (0.6052)
ROUGE-2	0.2005 (0.0956)	0.2027 (0.1231)	0.1835 (0.4462)	0.2028 (0.2302)
ROUGE-L	0.2197 (0.4980)	0.2011 (0.3924)	0.1908 (0.2993)	0.2335 (0.7158)
METEOR	0.1857 (0.1314)	0.1576 (0.4371)	0.1518 (0.8903)	0.1685 (0.4094)
BERTScore	0.2555 (0.6227)	0.2542 (0.9268)	0.2194 (0.1936)	0.2558 (0.2032)
DEB	0.3255 (0.0152)	0.3306 (0.0470)	0.3419 (0.0158)	0.3668 (0.0812)
USR	0.3466 (0.0045)	0.3428 (0.1257)	0.3338 (0.0478)	0.1706 (0.0462)
MDD-Eval	0.3277 (0.0245)	0.3398 (0.2784)	0.3869 (0.3478)	0.3557 (0.0254)
Mask-and-fill	0.2998 (0.0458)	0.3052 (0.0025)	0.3668 (0.1069)	0.3627 (0.0044)
G-Eval (GPT-3.5)	0.4995 (0.0025)	0.4754 (0.0011)	0.4774 (0.0069)	0.4688 (0.0098)
G-Eval (GPT-4)	0.5314 (0.0028)	0.5055 (0.0015)	0.4995 (0.0057)	0.5022 (0.0064)
LLM-Eval (GPT-3.5)	0.4837 (0.0001)	0.4798 (0.0004)	0.4512 (0.0007)	0.4799 (0.0004)
LLM-Eval (GPT-4)	0.5008 (0.0022)	0.5096 (0.0036)	0.5178 (0.0019)	0.5257 (0.0007)
Ours(w/o LLM)	0.3602 (0.0011)	0.3599 (0.0004)	0.3611 (0.0017)	0.3587 (0.0023)
Ours (GPT-3.5 w/o AMR)	0.5022 (0.0001)	0.5120 (0.0009)	0.5118 (0.0025)	0.5099 (0.0002)
Ours (GPT-3.5 w/o SLM)	0.5172 (0.0025)	0.5099 (0.0065)	0.5112 (0.0004)	0.5101 (0.0051)
Ours(GPT-3.5)	0.5200 (0.0051)	0.5115 (0.0007)	0.5127 (0.0057)	0.5110 (0.0001)
Ours (GPT-4 w/o AMR)	0.6274 (0.0001)	0.6266 (0.0019)	0.6198 (0.1237)	0.5207 (0.0272)
Ours (GPT-4 w/o SLM)	0.6470 (0.0021)	0.6482 (0.0031)	0.6398 (0.0004)	0.6402 (0.0054)
Ours (GPT-4)	**0.6641 (0.0002)**	**0.6603 (0.0002)**	**0.6598 (0.0007)**	**0.6674 (0.0003)**

Table 3: Pearson and Spearman correlations with human judgments on the TopicalChat dataset.

bustness of our human evaluation framework.

5 Results

5.1 Evaluation Performance on Standard Set

We evaluate our model against the baselines by analysing the correlation between automated evaluation scores and human judgements across three datasets. The results presented in Table 1 to Table 3 reveal that n-gram and embedding-based baselines, which compute word overlap or semantic similarity between gold references and responses, demonstrate weak positive correlations with human annotations across two datasets. Amongst the n-gram baselines, ROUGE-L exhibits the strongest correlation. The embedding-based approach, BERTScore, whilst outperforming the n-gram baselines, still achieves suboptimal performance when compared with more sophisticated metrics. Learning-based metrics, which consider the contextual relationship between dialogue pairs, demonstrate superior overall performance. Specifically, Mask-and-fill and USR achieve better correlations than n-gram baselines, whilst DEB and MDD-Eval secure higher correlations among these approaches. Regarding LLM-based methods, G-Eval and LLM-Eval demonstrate the strongest performance across all three datasets, establishing themselves as the leading baselines.

Our method in its basic configuration (Ours w/o LLM) achieves moderately positive correlations across the three datasets (less than 0.4). However, when integrating SLM with LLM, our approach achieves the highest overall performance on both Pearson and Spearman correlations across all datasets. Notably, our GPT-4 variant exhibits superior performance compared to all baselines. Through ablation studies examining the effectiveness of SLM and AMR graphs, we observe that Ours (w/o SLM) outperforms Ours (w/o AMR), which combines only LLM and SLM components, thereby validating the effectiveness of incorporating AMR graphs in open-domain dialogue evaluation.

5.2 Evaluation Performance on Adversarial Set

To evaluate our method's capability in evaluating adversarial negative examples, we conduct comparative analyses against baseline approaches on the adversarial set. Tables 1 to 3 present the correlation results between automated metrics and human judgements.

The n-gram and embedding-based metrics exhibit weakly positive correlations with human judgements, primarily due to their inherent limitation of solely comparing gold references with response candidates, without considering the contextual relationships that characterise adversarial examples. Regarding learning-based approaches, USR demonstrates limited robustness against adversarial negative examples, showing only weak positive correlations with human judgements. In contrast, MDD-Eval, Mask-and-fill, and DEB achieve notably stronger performance across both Pearson and Spearman correlations. LLM-based methods establish themselves as the strongest baseline approaches, demonstrating superior performance in handling adversarial examples.

Our proposed metric consistently surpasses all baseline approaches across both correlation metrics. Specifically, Ours(GPT-4) achieves strong correlations on the adversarial set, significantly outperforming the strongest baseline G-Eval. Similar improvements are observed in Spearman correlations across the three datasets. The ablation analysis further substantiates the benefits of our integrated approach: Ours(w/o AMR) shows notably lower correlations, demonstrating that the incorporation of AMR graph information significantly enhances the model's ability to evaluate adversarial examples. These results comprehensively validate the effectiveness of integrating AMR graph-enhanced SLM with LLMs for robust open-domain dialogue evaluation.

Context:	Hi kevin, how was your year at college? It was great! How was your year? It was good. Do you have a **girlfriend** at **school**?
Response:	Are you still in touch with any of your old **school friends**?
Context:	Would you recommend some **places** for sightseeing? How about great canyon? Is it worth seeing?
Response:	Singapore is reportedly a very exciting **place** to live.
Context:	I need change for the **machines**? You need to put **50** cents into the washer **machine** and a dollar into the dryer. So what do I need to do?
Response:	In our factory, there are **50** electrical **machines**.

Table 4: Samples of context-response pairs. The bold words represent the overlapping words.

5.3 Case Study

To demonstrate the effectiveness of AMR graphs in identifying adversarial negative responses, we present several illustrative examples in Table 4.

Model	Accuracy
BERT Regressor	75.92
RUBER	76.50
DEB	82.04
Mask-and-fill	85.27
Ours (SLM)	**86.81**
Ours (- w/o GM)	86.22
Ours (- w/o CL)	86.46
Ours (- w/o GM, CL)	85.64
Graph Transformer	84.73
Sentence Transformer	83.81

Table 5: Ablation study on Dailydialog++ dataset.

These cases highlight instances where responses were incorrectly classified as "positive" without AMR graph information, but were accurately identified as "negative" when incorporating semantic structural knowledge from AMR graphs. This analysis underscores the crucial role of AMR-derived semantic information in enhancing the model's discriminative capability for challenging adversarial examples. We also analyse the attention heatmap of Graph Transformer and Sentence Transformer in Appendix A.1

5.4 Ablation Study

We evaluate our SLM's classification performance on the DailyDialog++ testset. As shown in Table 5, our SLM surpasses all baselines and demonstrating the effectiveness of incorporating AMR graph information. Ablation studies reveal that removing either the Graph Transformer or Sentence Transformer components of SLM leads to decreased performance, with the Graph Transformer alone performing marginally better than the Sentence Transformer. While removing the contrastive learning (CL) or gating mechanism (GM) shows minimal impact, the removal of AMR information results in the most significant performance drop, highlighting its crucial role in dialogue evaluation.

6 Conclusion

In this paper, we presents a novel evaluation framework for open-domain dialogue systems that integrates AMR graph-enhanced SLMs with LLMs. Comprehensive experimental results across multiple datasets demonstrate that our method consistently outperforms existing approaches, including state-of-the-art LLM-based methods, in the challenging task of open-domain dialogue evaluation.

Ethics Statement

Our proposed evaluation metric enhances the assessment of open-domain dialogue systems through AMR integration and contrastive learning. While the framework effectively addresses the one-to-many nature of dialogue evaluation, it may occasionally assign high scores to inappropriate responses. We recommend careful screening of training data and implementation of content filters before deployment.

Limitations

Despite demonstrating robust performance, our method primarily focuses on semantic dependencies between context and response. Following Howcroft et al. (2020), we acknowledge that human evaluation encompasses multiple attributes beyond semantic relationships. Future work should explore disentangling these various attributes to enhance model interpretability and evaluation comprehensiveness.

References

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. *ArXiv*, abs/2105.10188.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Claire Bonial, L. Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David R. Traum, and Clare R. Voss. 2020. Dialogue-amr: Abstract meaning representation for dialogue. In *International Conference on Language Resources and Evaluation*.

Deng Cai and Wai Lam. 2020. Amr parsing via graph-sequence iterative inference. *ArXiv*, abs/2004.05572.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gabriel Forgues and Joelle Pineau. 2014. Bootstrapping dialog systems with word embeddings.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *ArXiv*, abs/2302.04166.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH*.

Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu, Zhenhua Ling, Zhiming Su, and Si Wei. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey P. Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

David M. Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *INLG*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *European Association for Machine Translation Conferences/Workshops*.

Ioannis Konstas, Srini Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural amr: Sequence-to-sequence models for parsing and generation. In *Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *ArXiv*, abs/1603.08023.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634.

Ryan Lowe, Michael Noseworthy, Iulian Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *ArXiv*, abs/1708.07149.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Ananya B. Sai, Akash Kumar Mohankumar, Siddharth Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Linfeng Song, Ante Wang, Jinsong Su, Yue Zhang, Kun Xu, Yubin Ge, and Dong Yu. 2020. Structural information preserving for graph-to-text generation. *ArXiv*, abs/2102.06749.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *ArXiv*, abs/2303.04048.

Bohao Yang, Chen Tang, and Chenghua Lin. 2024. Improving medical dialogue generation with abstract meaning representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11826–11830. IEEE.

Chen Zhang, L. F. D'Haro, Thomas Friedrichs, and Haizhou Li. 2021. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *AAAI Conference on Artificial Intelligence*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. 2024. Slide: A framework integrating small and large language models for open-domain dialogues evaluation. *arXiv preprint arXiv:2405.15924*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiehan Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better amr-to-text generation. In *Conference on Empirical Methods in Natural Language Processing*.

A More Experimental Results and Analysis

A.1 Attention Visualisation Analysis

We analyse the attention patterns of both Sentence and Graph Transformers of the SLM through visualisation of their attention heatmaps for the context-response pair shown in Figure 3.

The Sentence Transformer exhibits strong attention weights between overlapping tokens in context and response. As illustrated in Figure 3 (top), tokens such as "school" and "friend" in the response show high attention scores with their counterparts "school" and "girlfriend" in the context. In contrast, the Graph Transformer, which incorporates entity relationships through AMR structures, demonstrates different attention patterns. Figure 3 (bottom) shows that these lexically similar tokens receive lower attention weights, indicating the model's ability to capture semantic differences beyond surface-level similarities.

B Prompt Templates

B.1 Prompt for Engagingness evaluation

```
Rate the dialogue response.
Use the prediction probability from the SLMs and AMR graphs of the conversation pair to aid your judgment.
Note: Please take the time to fully read and understand the dialogue response.
How dull/interest is the text of the dialogue response? (on a scale of 1-5, with 1 being the lowest)
Input:
Conversation Context:
Response:
AMR Graph:
SLM score:

Evaluation Form (Score ONLY):
Engagingness:
```

B.2 Prompt for Naturalness evaluation

```
Rate the dialogue response.
Use the prediction probability from the SLMs and AMR graphs of the conversation pair to aid your judgment.
Note: Please take the time to fully read and understand the dialogue response.
To what extent the response is naturally written (on a scale of 1-5, with 1 being the lowest)
Input:
Conversation Context:
Response:
AMR Graph:
SLM score:

Evaluation Form (Score ONLY):
Naturalness:
```

B.3 Prompt for Coherence evaluation

```
Rate the dialogue response.
Use the prediction probability from the SLMs and AMR graphs of the conversation pair to aid your judgment.
Note: Please take the time to fully read and understand the dialogue response.
To what extent the response is well-structured, logical, and meaningful (on a scale of 1-5, with 1 being the lowest)
Input:
Conversation Context:
Response:
AMR Graph:
SLM score:

Evaluation Form (Score ONLY):
Coherence:
```

B.4 Prompt for Groundedness evaluation

```
Rate the dialogue response.
Use the prediction probability from the SLMs and AMR graphs of the conversation pair to aid your judgment.
Note: Please take the time to fully read and understand the dialogue response.
To what extent the response is grounded in facts present in the context (on a scale of 1-5, with 1 being the lowest)
Input:
Conversation Context:
Response:
AMR Graph:
SLM score:

Evaluation Form (Score ONLY):
Groundedness:
```

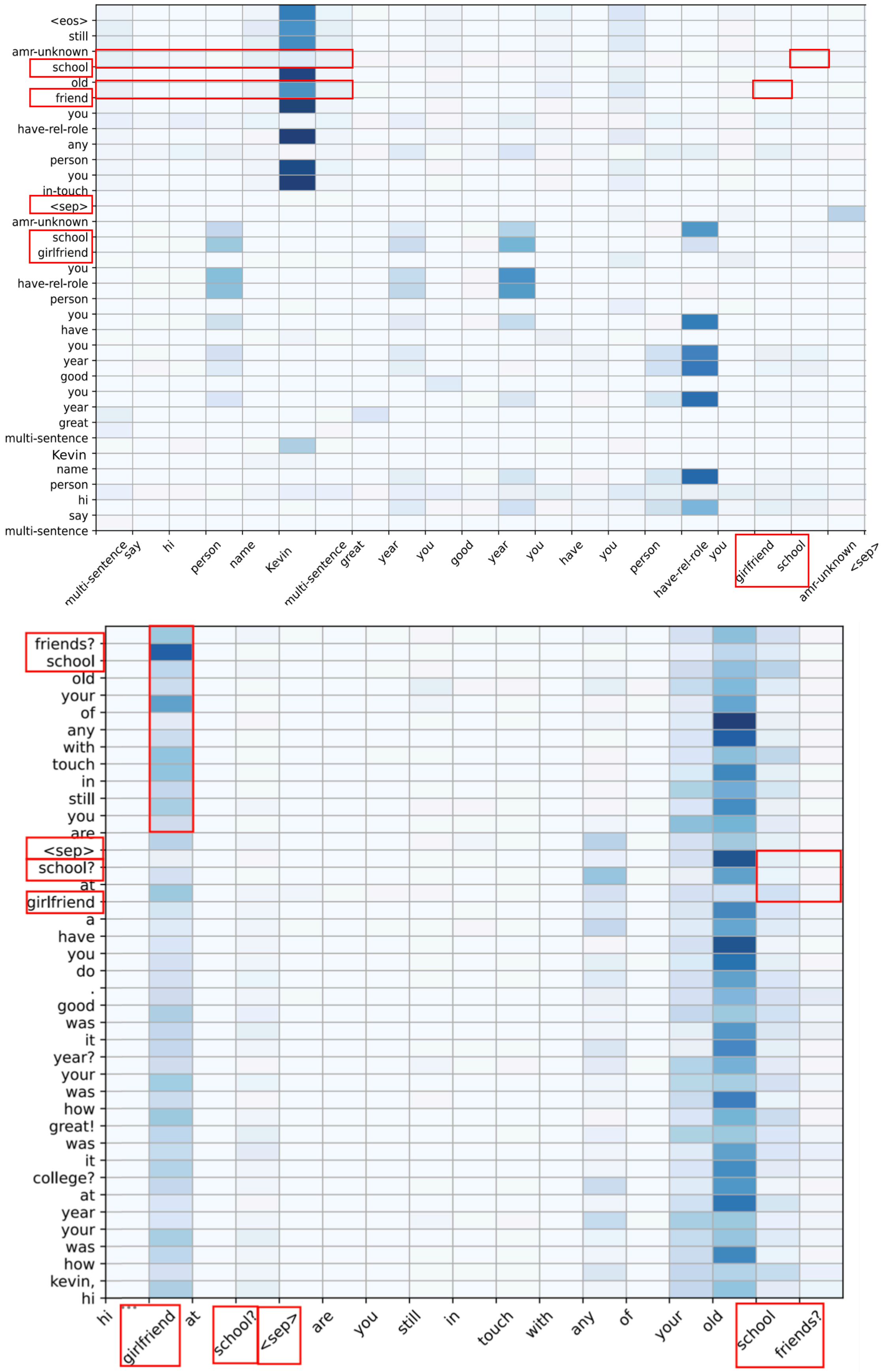

Figure 3: Attention pattern visualisation for context-response analysis. Top: Graph Transformer attention heatmap showing semantic-aware attention distribution. Bottom: Sentence Transformer attention heatmap highlighting lexical-level attention patterns. Overlapping tokens between context and response (*friends* and *school*) demonstrate distinct attention behaviours in the two encoders.