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 effectively evaluate such responses, 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 017 face challenges in handling adversarial negative examples. We propose a novel evaluation framework that integrates Abstract Meaning Representation (AMR) enhanced domain-021 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 evalu-027 ation. Extensive experiments on open-domain dialogue evaluation tasks demonstrate the superiority of our method compared to state-ofthe-art baselines. Our comprehensive ablation studies reveal that AMR graph information contributes substantially more to performance improvements. Our framework achieves strong correlations with human judgments across multiple datasets, establishing a new benchmark for dialogue evaluation. Our code and data are 037 publicly available.

1 Introduction

042

Open-domain dialogue systems have garnered substantial attention owing to their broad applicability (Zhang et al., 2021; Liu et al., 2023) across

Dialogue Context



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.

various domains, including personal medical assistance and biomedical telecommunications (Sai et al., 2020; Yang et al., 2024b). 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 nega-

tive examples, tend to assess responses primarily through surface-level content similarity. Although some approaches have attempted to address 062 this by incorporating adversarial examples (Sai et al., 2020; Gupta et al., 2021), they either require extensive pre-training on large-scale conver-065 sational corpora or demand adaptation to specific datasets, incurring substantial computational overhead. Moreover, their exclusive reliance on surfaceform features compromises robustness when evaluating adversarial examples that deviate from the training distribution. The vulnerability to adversarial attacks further compounds this challenge. Jin 072 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.

061

090

100

101

103

104

105

106 107

108

109

110

111

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 graphenhanced 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.

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

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:

- A novel framework integrating AMR graph information into open-domain dialogue evaluation through a dual-representation approach that combines specialized SLMs with LLMs.
- A comprehensive evaluation methodology across four distinct criteria (Naturalness, Coherence, Engagingness, and Groundedness), with detailed performance breakdowns demonstrating consistent improvements across all dimensions.
- Extensive experimental results demonstrating substantial improvements over existing methods including reasoning-focused LLMs, with ablation studies revealing that AMR graph information contributes 7.4% more to performance than SLM score alone.

2 **Related Work**

Dialogue Evaluation Metrics. Traditional ngram-based metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and ME-TEOR (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

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

semantic spaces before calculating their similarity. However, both approaches have shown limited efficacy in evaluating open-domain dialogue systems (Liu et al., 2016).

161

162

163

164

166

167

168

169

170

171

172

173

174

175

177

178

179

181

203

206

210

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-andfill (Gupta et al., 2021) employs a Speaker-Aware BERT architecture (Gu et al., 2020) to enhance dialogue understanding, though it requires dataset-176 specific adaptation before fine-tuning. Zhang et al. (2021) developed MDD-Eval for cross-domain dialogue evaluation, but this method necessitates hu-180 man annotations and additional training data while failing to address adversarial negative examples.

LLM-based Evaluators. The emergence of Large 182 Language Models (LLMs) has enabled new ap-183 proaches 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. 187 (2023) empirically validated the effectiveness of LLM-based evaluation approaches. Kocmi and Fe-189 dermann (2023) demonstrated the utility of GPT models in machine translation evaluation. Liu et al. (2023) introduced G-Eval, employing GPT-4 across 192 multiple generation tasks including dialogue re-193 sponse, text summarization, data-to-text generation, and machine translation. Chan et al. (2023) 195 proposed ChatEval, a multi-agent debate framework that surpasses single-LLM evaluators in performance. However, these LLM-based approaches 198 have yet to be applied to evaluating adversarial neg-199 ative responses incorporating non-textual domain knowledge.

Methodology 3

Task Description 3.1

Our model operates on input tuples consisting of a dialogue context 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:

$$Score_{SLM} = P(\mathcal{Y} \mid \mathcal{C}, \mathcal{R}, \mathcal{G}_{\mathcal{C}}, \mathcal{G}_{\mathcal{R}})$$
(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 Score_{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.

The complementary nature of SLM and LLM integration stems from their distinct capabilities: while the SLM excels at encoding structured graph information through specialized transformers, LLMs offer superior contextual reasoning but lack native graph processing abilities. As shown in our attention analysis in Appendix A.2, the SLM's graph encoder can identify semantic inconsistencies in adversarial examples that may be missed by text-only representations. By combining these approaches, our framework leverages both struc-



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.

tured semantic knowledge and advanced reasoning capabilities.

3.3 Sequence Encoder

260

262

273

274

275

276

277

278

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

$$\mathbf{H}_{S} = \operatorname{SeqEncoder}(\mathcal{C}, \mathcal{R})$$
(2)

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

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

where $\mathbf{H}_S = \{h_1, h_2, \dots, h_{\mathcal{C}+\mathcal{R}}\}$ 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:

$$\mathbf{H}_{A} = \operatorname{GraphEncoder}(\mathcal{V}, \mathcal{E}) \tag{5}$$

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

289

291

292

293

294

296

297

298

300

301

302

303

304

305

309

where $\mathbf{H}_A = \{h'_1, h'_2, \dots, h'_M\}$ 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:

$$\hat{\alpha}_{ij} = \frac{\exp\left(\hat{e}_{ij}\right)}{\sum_{m=1}^{M} \exp\left(\hat{e}_{im}\right)}$$
²⁹⁵

$$\hat{e}_{ij} = \frac{\left(W^Q h'_i\right)^T \left(W^K h'_j + W^R \boldsymbol{r}_{ij}\right)}{\sqrt{d}} \quad (7)$$

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:

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

$$\hat{\mathbf{H}} = g_i \mathbf{H}_S + (1 - g_i) \mathbf{H}_A \tag{9}$$
 307

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

314

3.6 Training objectives and Evaluation

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

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

The training objective combines classification and contrastive learning:

040

319

323

326

329

330

331

332

334

336

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

The contrastive loss \mathcal{L}_C facilitates alignment between sentence and graph representations:

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

$$\mathcal{L}_C = -\frac{1}{N} \sum_{i=1}^N \frac{e^{\sin(\mathbf{H}_S^+, \mathbf{H}_A^+)}}{\sum_j e^{\sin(\mathbf{H}_S^-, \mathbf{H}_A^-)}} \qquad (13)$$

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 $Score_{SLM}$ and AMR graph information \mathcal{G} through the LLM's prompt.

$$Score = LLMs(Score_{SLM}, \mathcal{G})$$
 (14)

4 Experiments

4.1 Dataset

We conduct experiments on three widelyrecognised open-domain dialogue datasets: **Daily-Dialog++** (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. The SLM is trained on the DailyDialog++ dataset, which comprises 9,259 dialogue contexts in the training set, 1,028 in the validation set, and 1,142 in the test set. 348

349

350

351

352

354

355

356

357

358

360

361

362

365

367

368

369

371

372

373

374

375

376

377

378

379

380

381

383

384

385

386

387

389

390

391

392

393

394

395

397

4.3 Baselines

(11)

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), Maskand-fill (Gupta et al., 2021), and MDD-Eval (Zhang et al., 2021). Additionally, we select G-Eval (Liu et al., 2023), QWQ-32B (Team, 2025), Qwen2.5-7B (Yang et al., 2024a), and LLM-Eval (Lin and Chen, 2023) as the LLM-based metrics. For Qwen2.5-7B, we fine-tuned it on 12,000 both text and AMR structured dialogue examples from all three datasets, ensuring no overlap with evaluation sets.

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 ag-

	Standa	ard Set	Advers	arial Set
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	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
QwQ-32B	0.3915 (0.0123)	0.3876 (0.0142)	0.3783 (0.0224)	0.3861 (0.0182
Qwen2.5-7B	0.3687 (0.0152)	0.3702 (0.0134)	0.3557 (0.0213)	0.3674 (0.0198
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.

	Standa	ard Set	Adversa	arial Set
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	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 (0.892)
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
QwQ-32B	0.5027 (0.0124)	0.5006 (0.0132)	0.4778 (0.0215)	0.4827 (0.0164
Qwen2.5-7B	0.4792 (0.0146)	0.4734 (0.0129)	0.4623 (0.0218)	0.4707 (0.0173
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 (0.927)
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.

	Standa	ard Set	Advers	arial Set
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	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)
QwQ-32B	0.5092 (0.0118)	0.4836 (0.0125)	0.4887 (0.0208)	0.4824 (0.0152)
Qwen2.5-7B	0.4927 (0.0137)	0.4703 (0.0114)	0.4702 (0.0211)	0.4678 (0.0167)
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.

398 gregate, our evaluation corpus comprises 2,400399 context-response pairs.

Correlation Computation For reporting our ex-400 perimental results, we compute correlation between 401 automated scores and human judgments separately 402 for each of the four criteria (Naturalness, Coher-403 ence, Engagingness, and Groundedness). The re-404 ported values in Tables 1-3 represent the average 405 correlations across all four dimensions. This ap-406 proach follows standard practices in dialogue eval-407 uation research (Mehri and Eskenazi, 2020). A 408 detailed breakdown of performance across individ-409 ual criteria is provided in Appendix C. 410

Human Annotation Three qualified human eval-411 uators, each holding at least a master's degree 412 in Computer Science and demonstrating full pro-413 fessional English proficiency, independently rated 414 each context-response pair. Assessments were con-415 ducted using a 5-point Likert scale, where higher 416 scores indicate superior quality. The final human 417 annotation score for each aspect was derived by 418 averaging across all evaluators. To ensure annota-419 tion reliability, we computed the Inner-Annotator 420 421 Agreement (IAA) using Cohen's Kappa coefficient (Cohen, 1960). The achieved average IAA 422 score of 0.64 between annotator pairs indicates 423 substantial agreement (0.6-0.8), validating the ro-424 bustness of our human evaluation framework. 425

5 Results

5.1 Evaluation Performance on Standard Set

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

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 strong performance across all three datasets. We also evaluated reasoningfocused LLMs including QwQ-32B (via direct prompting without AMR) and Qwen2.5-7B (finetuned with structured data for 5 epochs). These models perform slightly better than GPT-3.5 across all datasets. Similarly, the fine-tuned Qwen2.5-7B (0.3687/0.3702) outperforms GPT-3.5, demonstrating the potential of specialized reasoning models.

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

487

488

489

490

491

492

493

494

495

496

497

498

499

501

502

505

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, including the reasoning-focused LLMs. 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, with reasoningfocused models like QwQ-32B and fine-tuned Qwen2.5-7B showing improved performance over standard GPT-3.5. However, despite these improvements, these reasoning-focused LLMs still fall short of our full approach, suggesting that explicit semantic structure through AMR graphs provides complementary information that enhances evaluation capabilities beyond what these models can derive from text alone.

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. 506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

5.3 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.

When comparing Ours (w/o AMR) and Ours (w/o SLM) variants, we observe that removing AMR graph information leads to a more significant performance drop than removing the SLM score, confirming that the structured semantic knowledge encoded in AMR graphs contributes more to performance improvements. However, the full model combining both components achieves the best results, demonstrating that the SLM's specialized architecture for processing graph information and the LLM's reasoning capabilities operate synergistically rather than redundantly.

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 556

557 Our proposed evaluation metric enhances the assessment of open-domain dialogue systems through AMR integration and contrastive learning. While 559 the framework effectively addresses the one-to-560 many nature of dialogue evaluation, it may oc-561 562 casionally assign high scores to inappropriate responses. We recommend careful screening of training data and implementation of content filters before deployment.

Limitations 566

567

570

571

572

573

576

577

578

579

580

581

582

583

584

590

591

592

593

594

596

597

598

602

Despite demonstrating robust performance, our method primarily focuses on semantic dependen-568 cies 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 com-574 prehensiveness.

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 graphsequence 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.

dialog systems with word embeddings.	605
Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei	606
Liu. 2023. Gptscore: Evaluate as you desire. ArXiv,	607
abs/2302.04166.	608
Karthik Gopalakrishnan, Behnam Hedayatnia, Qin-	609
lang Chen, Anna Gottardi, Sanjeev Kwatra, Anu	610
Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür.	611
2019. Topical-Chat: Towards Knowledge-Grounded	612
Open-Domain Conversations. In INTERSPEECH.	613
Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu,	614
Zhenhua Ling, Zhiming Su, and Si Wei. 2020.	615
Speaker-aware bert for multi-turn response selec-	616
tion in retrieval-based chatbots. Proceedings of the	617
29th ACM International Conference on Information	618
& Knowledge Management.	619
Prakhar Gupta, Yulia Tsvetkov, and Jeffrey P. Bigham.	620
2021. Synthesizing adversarial negative responses	621
for robust response ranking and evaluation. In Find-	622
ings.	623
David M. Howcroft, Anya Belz, Miruna Clinciu, Dimi-	624
tra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon	625
Mille, Emiel van Miltenburg, Sashank Santhanam,	626
and Verena Rieser. 2020. Twenty years of confusion	627
in human evaluation: Nlg needs evaluation sheets	628
and standardised definitions. In INLG.	629
Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter	630
Szolovits. 2019. Is bert really robust? a strong base-	631
line for natural language attack on text classification	632
and entailment. In AAAI Conference on Artificial	633
Intelligence.	634
Tom Kocmi and Christian Federmann. 2023. Large lan-	635
guage models are state-of-the-art evaluators of trans-	636
lation quality. In European Association for Machine	637
Translation Conferences/Workshops.	638
Ioannis Konstas, Srini Iyer, Mark Yatskar, Yejin Choi,	639
and Luke Zettlemoyer. 2017. Neural amr: Sequence-	640
to-sequence models for parsing and generation. In	641
Annual Meeting of the Association for Computational	642
Linguistics.	643
Chin-Yew Lin. 2004. Rouge: A package for automatic	644
evaluation of summaries. In Annual Meeting of the	645
Association for Computational Linguistics.	646
Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval:	647
Unified multi-dimensional automatic evaluation for	648
open-domain conversations with large language mod-	649
els. In Proceedings of the 5th Workshop on NLP for	650
Conversational AI (NLP4ConvAI 2023), pages 47–	651
58, Toronto, Canada. Association for Computational	652
Linguistics.	653
Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Nose-	654
worthy, Laurent Charlin, and Joelle Pineau. 2016.	655
How not to evaluate your dialogue system: An empir-	656
ical study of unsupervised evaluation metrics for dia-	657
logue response generation. ArXiv, abs/1603.08023.	658

Gabriel Forgues and Joelle Pineau. 2014. Bootstrapping

604

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen

abs/2303.16634.

abs/1708.07149.

827.

Computational Linguistics.

ArXiv, abs/2102.06749.

reinforcement learning.

you need. In NIPS.

study. ArXiv, abs/2303.04048.

Xu, and Chenguang Zhu. 2023. G-eval: Nlg evalua-

tion using gpt-4 with better human alignment. ArXiv,

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,

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-

the Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddharth

Arora, and Mitesh M. Khapra. 2020. Improving di-

alog evaluation with a multi-reference adversarial

dataset and large scale pretraining. Transactions of

the Association for Computational Linguistics, 8:810-

Xu, Yubin Ge, and Dong Yu. 2020. Structural in-

formation preserving for graph-to-text generation.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan.

Qwen Team. 2025. Qwq-32b: Embracing the power of

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

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui,

Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jian-

hong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang,

Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng

Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and

Zihan Qiu. 2024a. Qwen2.5 technical report. arXiv

Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary

2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In AAAI.

Linfeng Song, Ante Wang, Jinsong Su, Yue Zhang, Kun

Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of*

- 66
- 00/
- 66
- 66 66
- 66
- 668 669 670 671 672
- 67
- 6

676 677

- 678 679 680
- 6
- 684 685

686 687

- 68
- 69

69

69

6

698

70

701

702 703 704

705 706

707 708 709

7

- 710 711
- 712 preprint arXiv:2412.15115.

Bohao Yang, Chen Tang, and Chenghua Lin. 2024b. 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. 713

714

715

717

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

- 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*.
- 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 Case Study

745

746

747

751

754

758

759

761

762

764

770

771

773

774

775

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

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 ma- chine and a dollar into the dryer. So what do I need to do?
Response:	In our factory, there are 50 electrical ma- chines .

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

A.2 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 contextresponse 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.

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.

778

779

B Prompt Templates

B.1 Prompt for Engagingness evaluation

Rate the dialogue response.	780
Use the prediction probability from the	781
SLMs and AMR graphs of the conversation	782
pair to aid your judgment.	783
Note: Please take the time to fully read	784
and understand the dialogue response.	785
How dull/interest is the text of the	786
dialogue response? (on a scale of 1-5,	787
with 1 being the lowest)	788
Input:	789
Conversation Context:	790
Response:	791
AMR Graph:	792
SLM score:	793
	794
Evaluation Form (Score ONLY):	795
Engagingness:	796
	797
B.2 Prompt for Naturalness evaluation	798
Rate the dialogue response.	799

Use the prediction probability from the 800 SLMs and AMR graphs of the conversation 801 pair to aid your judgment. 802 Note: Please take the time to fully read 803 and understand the dialogue response. 804 To what extent the response is naturally 805 written (on a scale of 1-5, with 1 being 806 the lowest) 807 Input: 808 Conversation Context: 809 Response: 810 AMR Graph: 811 SLM score: 812 813 Evaluation Form (Score ONLY): 814

- Naturalness: 815 816 **B.3** Prompt for Coherence evaluation 817 Rate the dialogue response. 818 Use the prediction probability from the 819 SLMs and AMR graphs of the conversation 820 pair to aid your judgment. Note: Please take the time to fully read 822 and understand the dialogue response. 823 824 То what extent the response well-structured, logical, and meaningful 826 (on a scale of 1-5, with 1 being the lowest) Input: Conversation Context: Response: 830 AMR Graph: 831 832 SLM score: 833 Evaluation Form (Score ONLY): Coherence: 836 837 **B.4 Prompt for Groundedness evaluation** 838 Rate the dialogue response. Use the prediction probability from the SLMs and AMR graphs of the conversation pair to aid your judgment. 841 Note: Please take the time to fully read 842 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: 847 Conversation Context: Response: AMR Graph: SLM score: 851 Evaluation Form (Score ONLY): 853 Groundedness: 854 855 С Performance Breakdown by
 - **Evaluation Criteria**

858 Tables 6, 7, and 8 present the correlation results broken down by individual evaluation criteria (Nat-859 uralness, Coherence, Engagingness, and Groundedness) for each dataset. This detailed analysis reveals that our method consistently outperforms 862

baselines across all criteria, with particularly no-863 table improvements in Coherence and Grounded-864 ness for adversarial examples. This pattern aligns 865 with our expectation that AMR graph information 866 would be especially beneficial for capturing seman-867 tic inconsistencies that affect contextual appropri-868 ateness. 869

870

871

872

873

874

875

D **Example Prompt with AMR Graph** and SLM Score

is

In Table Table D, we provide a concrete example of how the SLM score and AMR graph information are incorporated into the LLM prompt for evaluation.

		Standa	ard Set			Advers	arial Set	
		ralness		erence		ralness		erence
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ
BLEU-1	0.1793	0.1769	0.1924	0.1873	0.2012	0.2054	0.2138	0.2168
BLEU-2	0.1836	0.1723	0.1947	0.1835	0.1487	0.1921	0.1598	0.2035
BLEU-3	0.1802	0.1788	0.1906	0.1893	0.1496	0.1925	0.1599	0.2026
BLEU-4	0.1932	0.1741	0.2053	0.1843	0.1549	0.1798	0.1648	0.1906
ROUGE-1	0.2126	0.1976	0.2266	0.2087	0.2248	0.2088	0.2392	0.2207
ROUGE-2	0.2001	0.1861	0.2134	0.1971	0.1462	0.1641	0.1574	0.1753
ROUGE-L	0.2129	0.1984	0.2258	0.2092	0.2318	0.2216	0.2459	0.2343
METEOR	0.1762	0.1521	0.1866	0.1612	0.1296	0.1002	0.1392	0.1071
BERTScore	0.2452	0.2593	0.2607	0.2742	0.1959	0.2176	0.2084	0.2301
DEB	0.3152	0.2778	0.3356	0.2954	0.3402	0.3321	0.3617	0.3517
USR	0.2568	0.2419	0.2730	0.2567	0.2234	0.2693	0.2365	0.2845
Mask-and-fill	0.1852	0.2007	0.1969	0.2124	0.2532	0.2827	0.2684	0.2986
MDD-Eval	0.2739	0.2354	0.2918	0.2505	0.2907	0.2723	0.3083	0.2886
G-Eval (GPT-3.5)	0.3326	0.3243	0.3512	0.3406	0.3201	0.3326	0.3411	0.3523
QwQ-32B	0.3822	0.3793	0.4023	0.3962	0.3692	0.3774	0.3898	0.3963
Qwen2.5-7B	0.3595	0.3615	0.3782	0.3796	0.3462	0.3584	0.3672	0.3779
G-Eval (GPT-4)	0.4219	0.4210	0.4458	0.4435	0.4183	0.4421	0.4443	0.4654
LLM-Eval (GPT-3.5)	0.3459	0.3638	0.3662	0.3834	0.3405	0.3335	0.3613	0.3528
LLM-Eval (GPT-4)	0.4210	0.4520	0.4452	0.4751	0.4573	0.4415	0.4842	0.4662
Ours(w/o LLM)	0.3486	0.3557	0.3686	0.3754	0.3398	0.3452	0.3612	0.3658
Ours (GPT-3.5 w/o AMR)	0.4474	0.4476	0.4725	0.4723	0.4504	0.4628	0.4765	0.4880
Ours (GPT-3.5 w/o SLM)	0.4663	0.4605	0.4922	0.4856	0.4767	0.4775	0.5055	0.5044
Ours (GPT-3.5)	0.4768	0.4751	0.5035	0.5006	0.4826	0.4793	0.5107	0.5057
Ours (GPT-4 w/o AMR)	0.5158	0.5258	0.5449	0.5545	0.5086	0.5384	0.5367	0.5685
Ours (GPT-4 w/o SLM)	0.5292	0.5559	0.5586	0.5862	0.5391	0.5079	0.5681	0.5361
Ours (GPT-4)	0.5550	0.5779	0.5863	0.6093	0.5485	0.5680	0.5801	0.5991
. ,	Enga	gingness	Grour	dedness	Enga	gingness	Grour	ndedness
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ
			real son sp	opeanian 57		opearman s /	r curson s p	Spearman's /
BLEU-1	· ·	•		•		•		•
BLEU-1 BLEU-2	0.1824	0.1809	0.1823	0.1849	0.2059	0.2091	0.2087	0.2095
BLEU-2	0.1824 0.1868	0.1809 0.1759	0.1823 0.1873	0.1849 0.1773	0.2059 0.1527	0.2091 0.1963	0.2087 0.1548	0.2095 0.1967
BLEU-2 BLEU-3	0.1824 0.1868 0.1832	0.1809 0.1759 0.1825	0.1823 0.1873 0.1848	0.1849 0.1773 0.1834	0.2059 0.1527 0.1537	0.2091 0.1963 0.1968	0.2087 0.1548 0.1540	0.2095 0.1967 0.1973
BLEU-2 BLEU-3 BLEU-4	0.1824 0.1868 0.1832 0.1967	0.1809 0.1759 0.1825 0.1777	0.1823 0.1873 0.1848 0.1968	0.1849 0.1773 0.1834 0.1786	0.2059 0.1527 0.1537 0.1593	0.2091 0.1963 0.1968 0.1844	0.2087 0.1548 0.1540 0.1602	0.2095 0.1967 0.1973 0.1828
BLEU-2 BLEU-3 BLEU-4 ROUGE-1	0.1824 0.1868 0.1832 0.1967 0.2171	0.1809 0.1759 0.1825 0.1777 0.2026	0.1823 0.1873 0.1848 0.1968 0.2169	0.1849 0.1773 0.1834 0.1786 0.2026	0.2059 0.1527 0.1537 0.1593 0.2294	0.2091 0.1963 0.1968 0.1844 0.2141	0.2087 0.1548 0.1540 0.1602 0.2281	0.2095 0.1967 0.1973 0.1828 0.2135
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911	0.1823 0.1873 0.1848 0.1968 0.2169 0.2043	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042 0.2176	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034	0.1823 0.1873 0.1848 0.1968 0.2169 0.2043 0.2183	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911	0.1823 0.1873 0.1848 0.1968 0.2169 0.2043	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042 0.2176 0.1797 0.2504	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658	0.1823 0.1873 0.1848 0.1968 0.2169 0.2043 0.2183 0.1791 0.2505	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 METEOR BERTScore DEB	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.2183 0.1791 0.2505 0.3224	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834	0.2059 0.1527 0.1537 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR	0.1824 0.1868 0.1832 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212 0.2619	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856 0.2482	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.2183 0.1791 0.2505 0.3224 0.2627	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2834	0.2059 0.1527 0.1537 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760	0.2087 0.1548 0.1540 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469 0.2307	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 METEOR BERTScore DEB	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.2183 0.1791 0.2505 0.3224	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834	0.2059 0.1527 0.1537 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval	0.1824 0.1868 0.1832 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212 0.2619 0.1895 0.2798	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856 0.2482 0.2056 0.2424	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.1791 0.2505 0.3224 0.2627 0.1900 0.2797	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2834\\ 0.2460\\ 0.2037\\ 0.2413\\ \end{array}$	0.2059 0.1527 0.1537 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5)	0.1824 0.1868 0.1868 0.1967 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212 0.2619 0.1895 0.2798 0.3352	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2482 0.2056 0.2482 0.2056 0.2424 0.3289	0.1823 0.1873 0.1873 0.1848 0.1968 0.2169 0.2043 0.2183 0.1791 0.2505 0.3224 0.2627 0.1900 0.2797 0.3483	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2895 0.2792 0.3401	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469 0.2307 0.2608 0.2963 0.3291	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B	0.1824 0.1868 0.1832 0.1967 0.2171 0.2042 0.2176 0.1797 0.2504 0.3212 0.2619 0.1895 0.2798 0.3352 0.3844	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856 0.2482 0.2056 0.2482 0.2056 0.2424 0.3289 0.3805	0.1823 0.1873 0.1873 0.1848 0.1968 0.2169 0.2043 0.2183 0.1791 0.2505 0.3224 0.2627 0.1900 0.2797 0.3483 0.3972	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2975 0.3273 0.3757	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469 0.2307 0.2608 0.2963 0.3291 0.3826	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.2658 0.2452 0.2452 0.2452 0.2424 0.2424 0.3289 0.3805 0.3625	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.1791 0.2505 0.3224 0.2627 0.1900 0.2797 0.3483 0.3972 0.3758	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943 0.3774	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ 0.3291\\ 0.3826\\ 0.3576\\ \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2767 0.2767 0.3398 0.3878 0.3698
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) Qwe0.25-7B G-Eval (GPT-4)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.3842\\ 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2482 0.2056 0.2482 0.2482 0.2056 0.2424 0.3289 0.3805 0.3625 0.4256	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ \hline \end{array}$	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2834\\ 0.2460\\ 0.2037\\ 0.2413\\ \hline 0.3363\\ 0.3943\\ 0.3774\\ 0.4347\\ \hline \end{array}$	0.2059 0.1527 0.1537 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518 0.4264	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878 0.3698 0.3698 0.4537
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.2658 0.2452 0.2452 0.2452 0.2424 0.2424 0.3289 0.3805 0.3625	0.1823 0.1873 0.1873 0.1848 0.2169 0.2043 0.2183 0.1791 0.2505 0.3224 0.2627 0.1900 0.2797 0.3483 0.3972 0.3758	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943 0.3774	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ 0.3291\\ 0.3826\\ 0.3576\\ \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2767 0.2767 0.3398 0.3878 0.3698
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (CPT-3.5) QwC-32B Qwen2.5-7B G-Eval (CPT-4) LLM-Eval (CPT-3.5) LLM-Eval (CPT-4)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline \end{array}$	$\begin{array}{c} 0.1809\\ 0.1759\\ 0.1825\\ 0.1777\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2658\\ \hline 0.2482\\ 0.2056\\ 0.2482\\ 0.2056\\ 0.2424\\ \hline 0.3289\\ 0.3805\\ 0.3625\\ 0.4256\\ 0.3723\\ 0.4621\\ \hline \end{array}$	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline \end{array}$	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2834\\ 0.2460\\ 0.2037\\ 0.2413\\ \hline 0.3363\\ 0.3943\\ 0.3774\\ 0.3697\\ 0.4569\\ \hline \end{array}$	$\begin{array}{c} 0.2059\\ 0.1527\\ 0.1537\\ 0.1593\\ 0.2294\\ 0.1501\\ 0.2367\\ 0.1331\\ 0.2004\\ \hline 0.3480\\ 0.2282\\ 0.2592\\ 0.2975\\ \hline 0.3273\\ 0.3757\\ 0.3518\\ 0.4264\\ 0.3484\\ 0.4660\\ \hline \end{array}$	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3502\\ 0.4689\\ \hline \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878 0.3698 0.3698 0.4537 0.3379 0.4497
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) Curs(w/o LLM)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.3553\\ \hline \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2482 0.2056 0.2482 0.2056 0.2482 0.2056 0.2424 0.3289 0.3805 0.3625 0.4256 0.3723 0.4621 0.3646	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ \hline \end{array}$	0.1849 0.1773 0.1834 0.1786 0.2026 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943 0.3774 0.3363 0.3774 0.3697 0.4569 0.3625	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3273 0.3757 0.3518 0.4264 0.3484 0.4660 0.3473	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528 0.3545	0.2087 0.1548 0.1540 0.1602 0.2281 0.1502 0.2364 0.1341 0.2017 0.3469 0.2307 0.2608 0.2963 0.3291 0.3826 0.3576 0.4302 0.3502 0.4689 0.3497	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) Durs (GPT-3.5 w/o AMR)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.3553\\ 0.4563\\ \hline \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.1561 0.2658 0.2856 0.2482 0.2056 0.2482 0.2056 0.2424 0.3289 0.3805 0.3625 0.3625 0.3723 0.4621 0.3646 0.4592	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ 0.4598\\ \hline \end{array}$	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943 0.3774 0.4347 0.3697 0.4569 0.3625 0.4577	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518 0.4264 0.3484 0.4660 0.3473 0.4583	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528 0.3545 0.4745	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3502\\ 0.4689\\ \hline 0.3497\\ 0.4640\\ \hline \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524 0.4727
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-3.5) LLM-Eval (GPT-4) Ours (GPT-3.5 w/o AMR) Ours (GPT-3.5 w/o SLM)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.4563\\ 0.4755\\ \hline \end{array}$	0.1809 0.1759 0.1825 0.1777 0.2026 0.1911 0.2034 0.2658 0.2482 0.2056 0.2482 0.2056 0.2424 0.3289 0.3805 0.3625 0.4256 0.3723 0.4621 0.3646 0.4592 0.4723	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ 0.4598\\ 0.4788\\ \hline \end{array}$	0.1849 0.1773 0.1834 0.1786 0.2026 0.1911 0.2034 0.1561 0.2639 0.2834 0.2460 0.2037 0.2413 0.3363 0.3943 0.3774 0.4347 0.3697 0.4569 0.3625 0.4577 0.4708	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2975 0.3273 0.3757 0.3518 0.4264 0.3484 0.4660 0.3473 0.4583 0.4583 0.4849	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528 0.3545 0.4745 0.4902	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3502\\ 0.4689\\ \hline 0.3497\\ 0.4640\\ 0.4921\\ \hline \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2767 0.3398 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524 0.4727 0.4887
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) DURS (GPT-3.5 w/o AMR) Ours (GPT-3.5 w/o AMR) Ours (GPT-3.5)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.3212\\ 0.2619\\ 0.3552\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.3553\\ 0.4563\\ 0.4755\\ 0.4865\\ \hline \end{array}$	$\begin{array}{c} 0.1809\\ 0.1759\\ 0.1759\\ 0.1825\\ 0.1777\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2658\\ \hline 0.2452\\ 0.2056\\ 0.2482\\ 0.2056\\ 0.2424\\ \hline 0.3289\\ 0.3805\\ 0.3625\\ 0.4256\\ 0.3723\\ 0.4621\\ \hline 0.3646\\ 0.4592\\ 0.4723\\ 0.4873\\ \hline \end{array}$	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ 0.4598\\ 0.4788\\ 0.4788\\ 0.4892\\ \hline \end{array}$	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2639\\ \hline 0.2834\\ 0.2460\\ 0.2037\\ 0.2413\\ \hline 0.3263\\ 0.3744\\ 0.3363\\ 0.3943\\ 0.3774\\ 0.4347\\ 0.3697\\ \hline 0.4569\\ \hline 0.3625\\ 0.4577\\ 0.4708\\ 0.4862\\ \hline \end{array}$	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518 0.4264 0.3484 0.4660 0.3473 0.4583 0.4849 0.4906	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528 0.3545 0.4745 0.4902 0.4920	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ 0.2963\\ 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4402\\ 0.3497\\ 0.4640\\ 0.4921\\ 0.4981\\ \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2767 0.2767 0.3398 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524 0.4727 0.4887 0.4900
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) Uurs (GPT-3.5 w/o AMR) Ours (GPT-3.5) Ours (GPT-3.5) Ours (GPT-3.5) Ours (GPT-3.5)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.1895\\ 0.2798\\ \hline 0.3352\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.3553\\ 0.4563\\ 0.4755\\ 0.4865\\ 0.5263\\ \hline \end{array}$	$\begin{array}{c} 0.1809\\ 0.1759\\ 0.1759\\ 0.1825\\ 0.1777\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2658\\ \hline 0.2482\\ 0.2056\\ 0.2482\\ 0.2056\\ 0.2482\\ 0.2056\\ 0.2424\\ \hline 0.3289\\ 0.3805\\ 0.3625\\ 0.4226\\ 0.3723\\ 0.4621\\ \hline 0.3646\\ 0.4592\\ 0.4723\\ 0.4673\\ 0.4873\\ 0.5392\\ \hline \end{array}$	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ 0.4598\\ 0.4788\\ 0.4788\\ 0.4892\\ 0.5299\\ \hline \end{array}$	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2834\\ 0.2639\\ \hline 0.2834\\ 0.2639\\ \hline 0.2639\\ \hline 0.2834\\ 0.2639\\ \hline 0.2639\\ \hline 0.363\\ 0.3943\\ 0.3774\\ 0.3363\\ 0.3943\\ 0.3774\\ 0.4347\\ 0.3697\\ 0.4569\\ \hline 0.3625\\ 0.4577\\ 0.4708\\ 0.4862\\ 0.5379\\ \hline \end{array}$	$\begin{array}{c} 0.2059\\ 0.1527\\ 0.1537\\ 0.1593\\ 0.2294\\ 0.1501\\ 0.2367\\ 0.1331\\ 0.2004\\ \hline 0.3480\\ 0.2282\\ 0.2592\\ 0.2975\\ \hline 0.3273\\ 0.3757\\ 0.3518\\ 0.4264\\ 0.3484\\ 0.4660\\ \hline 0.3473\\ 0.4583\\ 0.4849\\ 0.4906\\ 0.5171\\ \hline \end{array}$	$\begin{array}{c} 0.2091\\ 0.1963\\ 0.1968\\ 0.1844\\ 0.2141\\ 0.1693\\ 0.2271\\ 0.1034\\ 0.2230\\ \hline 0.3406\\ 0.2760\\ 0.2895\\ 0.2792\\ \hline 0.3401\\ 0.3829\\ 0.3635\\ 0.4500\\ 0.3421\\ 0.4528\\ \hline 0.3545\\ 0.4745\\ 0.4902\\ 0.4920\\ 0.5522\\ \hline \end{array}$	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1548\\ 0.1540\\ 0.281\\ 0.1502\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ \hline 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ \hline 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3502\\ 0.4689\\ \hline 0.3497\\ 0.4640\\ 0.4921\\ 0.4981\\ 0.5230\\ \hline \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2872 0.2767 0.3398 0.3878 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524 0.4727 0.4887 0.4900 0.5499
BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR Mask-and-fill MDD-Eval G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) Durs (GPT-3.5)	$\begin{array}{c} 0.1824\\ 0.1868\\ 0.1832\\ 0.1967\\ 0.2171\\ 0.2042\\ 0.2176\\ 0.1797\\ 0.2504\\ \hline 0.3212\\ 0.2619\\ 0.3212\\ 0.2619\\ 0.3552\\ 0.3844\\ 0.3615\\ 0.4264\\ 0.3525\\ 0.4283\\ \hline 0.3553\\ 0.4563\\ 0.4755\\ 0.4865\\ \hline \end{array}$	$\begin{array}{c} 0.1809\\ 0.1759\\ 0.1759\\ 0.1825\\ 0.1777\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2658\\ \hline 0.2452\\ 0.2056\\ 0.2482\\ 0.2056\\ 0.2424\\ \hline 0.3289\\ 0.3805\\ 0.3625\\ 0.4256\\ 0.3723\\ 0.4621\\ \hline 0.3646\\ 0.4592\\ 0.4723\\ 0.4873\\ \hline \end{array}$	$\begin{array}{c} 0.1823\\ 0.1873\\ 0.1873\\ 0.1848\\ 0.1968\\ 0.2169\\ 0.2043\\ 0.2183\\ 0.1791\\ 0.2505\\ \hline 0.3224\\ 0.2627\\ 0.1900\\ 0.2797\\ \hline 0.3483\\ 0.3972\\ 0.3758\\ 0.4342\\ 0.3546\\ 0.4315\\ \hline 0.3586\\ 0.4598\\ 0.4788\\ 0.4788\\ 0.4892\\ \hline \end{array}$	$\begin{array}{c} 0.1849\\ 0.1773\\ 0.1834\\ 0.1773\\ 0.1834\\ 0.1786\\ 0.2026\\ 0.1911\\ 0.2034\\ 0.1561\\ 0.2639\\ \hline 0.2639\\ \hline 0.2834\\ 0.2460\\ 0.2037\\ 0.2413\\ \hline 0.3263\\ 0.3744\\ 0.3363\\ 0.3943\\ 0.3774\\ 0.4347\\ 0.3697\\ \hline 0.4569\\ \hline 0.3625\\ 0.4577\\ 0.4708\\ 0.4862\\ \hline \end{array}$	0.2059 0.1527 0.1537 0.1593 0.2294 0.1501 0.2367 0.1331 0.2004 0.3480 0.2282 0.2592 0.2975 0.3273 0.3757 0.3518 0.4264 0.3484 0.4660 0.3473 0.4583 0.4849 0.4906	0.2091 0.1963 0.1968 0.1844 0.2141 0.1693 0.2271 0.1034 0.2230 0.3406 0.2760 0.2895 0.2792 0.3401 0.3829 0.3635 0.4500 0.3421 0.4528 0.3545 0.4745 0.4902 0.4920	$\begin{array}{c} 0.2087\\ 0.1548\\ 0.1540\\ 0.1602\\ 0.2281\\ 0.1502\\ 0.2364\\ 0.1341\\ 0.2017\\ 0.3469\\ 0.2307\\ 0.2608\\ 0.2963\\ 0.2963\\ 0.3291\\ 0.3826\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4302\\ 0.3576\\ 0.4402\\ 0.3497\\ 0.4640\\ 0.4921\\ 0.4981\\ \end{array}$	0.2095 0.1967 0.1973 0.1828 0.2135 0.1685 0.2254 0.1059 0.2221 0.3391 0.2742 0.2767 0.3398 0.3878 0.3698 0.4537 0.3379 0.4497 0.3524 0.4727 0.4887 0.4900

Table 6: Breakdown of Pearson and Spearman correlations with human judgments by evaluation criteria on the DailyDialog++ dataset.

		Standa	ard Set			Advers	arial Set	
	Natu	ralness		erence	Natu	ralness		erence
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ
BLEU-1	0.2014	0.2103	0.2115	0.2204	0.1725	0.1625	0.1807	0.1702
BLEU-2	0.1905	0.1780	0.1996	0.1869	0.1369	0.1581	0.1443	0.1658
BLEU-3	0.1638	0.1897	0.1722	0.1987	0.1340	0.1637	0.1412	0.1715
BLEU-4	0.1954	0.1886	0.2052	0.1976	0.1220	0.1505	0.1286	0.1582
ROUGE-1	0.2078	0.2106	0.2183	0.2214	0.2024	0.2145	0.2127	0.2251
ROUGE-2	0.1967	0.1975	0.2065	0.2073	0.1786	0.2023	0.1879	0.2124
ROUGE-L	0.2051	0.1985	0.2155	0.2085	0.1978	0.2182	0.2077	0.2291
METEOR	0.1949	0.1725	0.2047	0.1812	0.1403	0.1667	0.1476	0.1744
BERTScore	0.2796	0.2656	0.2934	0.2788	0.2196	0.2579	0.2313	0.2708
DEB	0.3562	0.3351	0.3744	0.3518	0.3424	0.3618	0.3601	0.3796
USR	0.3381	0.3370	0.3551	0.3542	0.3591	0.3765	0.3772	0.3954
MDD-Eval	0.3396	0.3328	0.3566	0.3492	0.3640	0.3513	0.3830	0.3689
Mask-and-fill	0.3017	0.3030	0.3168	0.3182	0.3673	0.3525	0.3856	0.3702
G-Eval (GPT-3.5)	0.4773	0.4757	0.5009	0.4988	0.4442	0.4502	0.4662	0.4722
QwQ-32B	0.4905	0.4889	0.5148	0.5126	0.4662	0.4714	0.4894	0.4942
Qwen2.5-7B	0.4675	0.4617	0.4910	0.4844	0.4512	0.4598	0.4735	0.4819
G-Eval (GPT-4)	0.5115	0.5195	0.5368	0.5444	0.5002	0.5391	0.5246	0.5640
LLM-Eval (GPT-3.5)	0.4539	0.4464	0.4757	0.4683	0.4343	0.4506	0.4557	0.4726
LLM-Eval (GPT-4)	0.5193	0.5265	0.5449	0.5522	0.5142	0.5131	0.5396	0.5385
Ours(w/o LLM)	0.3582	0.3696	0.3754	0.3873	0.3858	0.3815	0.4052	0.4007
Ours (GPT-3.5 w/o AMR)	0.4887	0.4878	0.5127	0.5119	0.4892	0.4982	0.5131	0.5229
Ours (GPT-3.5 w/o SLM)	0.4995	0.4946	0.5240	0.5192	0.5074	0.5062	0.5325	0.5313
Ours(GPT-3.5)	0.5385	0.5084	0.5648	0.5335	0.5079	0.5099	0.5329	0.5352
	0.6051	0.5982	0.6347	0.6273	0.6029	0.5862	0.6327	0.6149
Ours (GPT-4 w/o AMR)	0.6051							
Ours (GPT-4 w/o SLM)	0.6116	0.6149	0.6418	0.6450	0.6094	0.6158	0.6397	0.6462
							0.6397 0.6683	0.6462 0.6768
Ours (GPT-4 w/o SLM) Ours (GPT-4)	0.6116 0.6441 Engag	0.6149 0.6448 gingness	0.6418 0.6756 Grour	0.6450 0.6762	0.6094 0.6370 Engag	0.6158 0.6458 gingness	0.6683 Grour	0.6768 ndedness
Ours (GPT-4 w/o SLM)	0.6116 0.6441	0.6149 0.6448	0.6418 0.6756	0.6450 0.6762	0.6094 0.6370	0.6158 0.6458	0.6683	0.6768
Ours (GPT-4 w/o SLM) Ours (GPT-4)	0.6116 0.6441 Engag	0.6149 0.6448 gingness	0.6418 0.6756 Grour	0.6450 0.6762	0.6094 0.6370 Engag	0.6158 0.6458 gingness	0.6683 Grour	0.6768 ndedness
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics	0.6116 0.6441 Engag Pearson's ρ	0.6149 0.6448 gingness Spearman's τ	0.6418 0.6756 Grour Pearson's ρ	0.6450 0.6762 ndedness Spearman's τ	0.6094 0.6370 Engag Pearson's ρ	0.6158 0.6458 gingness Spearman's τ	0.6683 Grour Pearson's ρ	0.6768 idedness Spearman's τ
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1	0.6116 0.6441 Engag Pearson's ρ 0.2052	0.6149 0.6448 gingness Spearman's τ 0.2142	0.6418 0.6756 Grour Pearson's <i>ρ</i> 0.2072	0.6450 0.6762 dedness Spearman's τ 0.2158	0.6094 0.6370 Engag Pearson's ρ 0.1755	0.6158 0.6458 gingness Spearman's τ 0.1657	0.6683 Grour Pearson's <i>ρ</i> 0.1769	0.6768 dedness Spearman's τ 0.1669
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2	0.6116 0.6441 Pearson's <i>ρ</i> 0.2052 0.1940	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813	0.6418 0.6756 Grour Pearson's <i>ρ</i> 0.2072 0.1961	0.6450 0.6762 ndedness Spearman's τ 0.2158 0.1830	0.6094 0.6370 Engag Pearson's <i>ρ</i> 0.1755 0.1397	0.6158 0.6458 gingness Spearman's τ 0.1657 0.1616	0.6683 Grour Pearson's <i>ρ</i> 0.1769 0.1410	0.6768 dedness Spearman's τ 0.1669 0.1621
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3	0.6116 0.6441 Pearson's <i>ρ</i> 0.2052 0.1940 0.1670	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933	0.6418 0.6756 Grour Pearson's ρ 0.2072 0.1961 0.1688	0.6450 0.6762 dedness Spearman's τ 0.2158 0.1830 0.1947	0.6094 0.6370 Pearson's <i>ρ</i> 0.1755 0.1397 0.1367	0.6158 0.6458 gingness Spearman's τ 0.1657 0.1616 0.1673	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4	0.6116 0.6441 Pearson's ρ 0.2052 0.1940 0.1670 0.1992	0.6149 0.6448 Spearman's τ 0.2142 0.1813 0.1933 0.1921	0.6418 0.6756 Grour Pearson's ρ 0.2072 0.1961 0.1688 0.2010	0.6450 0.6762 dedness Spearman's τ 0.2158 0.1830 0.1947 0.1936	0.6094 0.6370 Engag Pearson's ρ 0.1755 0.1397 0.1367 0.1244	0.6158 0.6458 Spearman's τ 0.1657 0.1616 0.1673 0.1537	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L	0.6116 0.6441 Engag Pearson's ρ 0.2052 0.1940 0.1670 0.1992 0.2119 0.2004 0.2091	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \text{Grour}\\ \text{Pearson's }\rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ \end{array}$	0.6450 0.6762 hdedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2168 0.2030 0.2043	0.6094 0.6370 Engag Pearson's ρ 0.1755 0.1397 0.1367 0.1244 0.2067 0.1823 0.2017	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2203 0.2079 0.2244
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ \end{array}$	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024 0.1758	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline \\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline \\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ \end{array}$	0.6450 0.6762 dedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2168 0.2030 0.2043 0.1777	$\begin{array}{c} 0.6094 \\ \textbf{0.6370} \\ \hline \textbf{Engag} \\ \textbf{Pearson's } \rho \\ 0.1755 \\ 0.1397 \\ 0.1367 \\ 0.1244 \\ 0.2067 \\ 0.1823 \\ 0.2017 \\ 0.1430 \end{array}$	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225 0.1697	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2003 0.2079 0.2244 0.1711
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L	0.6116 0.6441 Engag Pearson's ρ 0.2052 0.1940 0.1670 0.1992 0.2119 0.2004 0.2091	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \text{Grour}\\ \text{Pearson's }\rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ \end{array}$	0.6450 0.6762 hdedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2168 0.2030 0.2043	0.6094 0.6370 Engag Pearson's ρ 0.1755 0.1397 0.1367 0.1244 0.2067 0.1823 0.2017	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2203 0.2079 0.2244
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ \end{array}$	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024 0.1758	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline \\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline \\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ \end{array}$	0.6450 0.6762 dedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2168 0.2030 0.2043 0.1777	$\begin{array}{c} 0.6094 \\ \textbf{0.6370} \\ \hline \textbf{Engag} \\ \textbf{Pearson's } \rho \\ 0.1755 \\ 0.1397 \\ 0.1367 \\ 0.1244 \\ 0.2067 \\ 0.1823 \\ 0.2017 \\ 0.1430 \end{array}$	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225 0.1697	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2003 0.2079 0.2244 0.1711
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore	0.6116 0.6441 Engag Pearson's ρ 0.2052 0.1940 0.1670 0.1992 0.2119 0.2004 0.2091 0.1987 0.2850	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024 0.1758 0.2707	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \text{Grour}\\ \text{Pearson's }\rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\end{array}$	0.6450 0.6762 hdedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2168 0.2030 0.2043 0.1777 0.2733	0.6094 0.6370 Engag Pearson's ρ 0.1755 0.1397 0.1367 0.1244 0.2067 0.1823 0.2017 0.1430 0.2240	$\begin{array}{c} 0.6158\\ \hline \textbf{0.6458}\\ \hline \textbf{0.6458}\\ \hline \textbf{Spearman's} \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline \end{array}$	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446 0.2265	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2203 0.2079 0.2244 0.1711 0.2652
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-L METEOR BERTScore DEB	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \\ \rho\\ 0.2052\\ 0.1940\\ 0.1940\\ 0.1992\\ 0.2119\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline \textbf{0.3626} \end{array}$	0.6149 0.6448 gingness Spearman's τ 0.2142 0.1813 0.1933 0.1921 0.2147 0.2014 0.2024 0.1758 0.2707 0.3417	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline \\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline \\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline \\ 0.3678\\ \end{array}$	0.6450 0.6762 dedness Spearman's τ 0.2158 0.1830 0.1947 0.1936 0.2030 0.2043 0.2043 0.1777 0.2733 0.3452	$\begin{array}{c} 0.6094 \\ \textbf{0.6370} \\ \hline \textbf{Engag} \\ \textbf{Pearson's } \rho \\ \hline 0.1755 \\ 0.1397 \\ 0.1367 \\ 0.1244 \\ 0.2067 \\ 0.1823 \\ 0.2017 \\ 0.1430 \\ 0.2240 \\ \hline 0.3489 \end{array}$	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225 0.1697 0.2633 0.3687	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446 0.2265 0.3533	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2003 0.2079 0.2244 0.1711 0.2652 0.3722
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ \end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.2158}\\ 0.1830\\ 0.1947\\ 0.1936\\ 0.2168\\ 0.2030\\ 0.2043\\ 0.2043\\ 0.1777\\ 0.2733\\ \hline \textbf{0.2733}\\ \hline \textbf{0.3452}\\ 0.3478\\ \end{array}$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ \end{array}$	0.6158 0.6458 spearman's τ 0.1657 0.1616 0.1673 0.1537 0.2194 0.2066 0.2225 0.1697 0.2633 0.3687 0.3838	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446 0.2265 0.3533 0.3698	0.6768 dedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2203 0.2079 0.2244 0.1711 0.2652 0.3722 0.3877
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1923\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ 0.3678\\ 0.3488\\ 0.3503\\ \end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.2158}\\ 0.1830\\ 0.1936\\ 0.2158\\ 0.2030\\ 0.2043\\ 0.2043\\ 0.2733\\ \hline \textbf{0.2733}\\ \hline \textbf{0.3452}\\ 0.3478\\ 0.3428\\ \hline \textbf{0.3428}\\ \hline \end{array}$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.2240\\ 0.3489\\ 0.3663\\ 0.3712\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.1657\\ \hline 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5)	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ \hline \textbf{Pearson's}\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.2158}\\ 0.1830\\ 0.1947\\ 0.1936\\ 0.2030\\ 0.2043\\ 0.2030\\ 0.2043\\ 0.2030\\ 0.2043\\ 0.3452\\ 0.3452\\ 0.3478\\ 0.3428\\ 0.3123\\ \hline \textbf{0.3123}\\ \hline \end{array}$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \text{Engag}\\ \text{Pearson's }\rho\\ \hline\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ 0.3594\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ \hline \textbf{0.3533} \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ \end{array}$	0.6768 adedness Spearman's τ 0.1669 0.1621 0.1680 0.1548 0.2203 0.2079 0.2244 0.1711 0.2652 0.3722 0.3877 0.3621 0.3631
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1923\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928 \end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline 0$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \text{Engag}\\ \text{Pearson's }\rho\\ \hline\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.1657\\ \hline 0.1616\\ \hline 0.1673\\ \hline 0.1637\\ \hline 0.2194\\ \hline 0.2066\\ \hline 0.2225\\ \hline 0.1697\\ \hline 0.2633\\ \hline 0.3687\\ \hline 0.3838\\ \hline 0.3581\\ \hline 0.3594\\ \hline 0.4587\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{dedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.2091\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ \end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{0.2158}\\ 0.2158\\ 0.1830\\ 0.1947\\ 0.1936\\ 0.2168\\ 0.2030\\ 0.2043\\ 0.2168\\ 0.2030\\ 0.2043\\ 0.1777\\ 0.2733\\ \hline \textbf{0.3428}\\ 0.3428\\ 0.3428\\ 0.3123\\ \hline \textbf{0.4904}\\ 0.5038\\ \hline \end{array}$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \text{Engag}\\ \text{Pearson's }\rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.1657\\ \hline 0.1657\\ \hline 0.1616\\ \hline 0.1673\\ \hline 0.1537\\ \hline 0.2194\\ \hline 0.2066\\ \hline 0.2225\\ \hline 0.1697\\ \hline 0.2633\\ \hline 0.3687\\ \hline 0.3838\\ \hline 0.3581\\ \hline 0.3594\\ \hline 0.4587\\ \hline 0.4801\\ \hline \end{array}$	0.6683 Grour Pearson's ρ 0.1769 0.1410 0.1380 0.1260 0.2082 0.1839 0.2036 0.1446 0.2265 0.3533 0.3698 0.3760 0.3781 0.4573 0.4809	$\begin{array}{r} \textbf{0.6768} \\ \textbf{dedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwC9-32B Qwen2.5-7B	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1940\\ 0.1970\\ 0.2052\\ 0.2019\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.3073\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{0.6448}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{odedness}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2158\\ 0.1830\\ 0.1947\\ 0.1936\\ 0.2030\\ 0.2043\\ 0.2030\\ 0.2043\\ 0.2030\\ 0.2043\\ 0.2030\\ 0.2043\\ 0.3123\\ \hline 0.3452\\ 0.3452\\ 0.3478\\ 0.3428\\ 0.3123\\ \hline 0.4904\\ 0.5038\\ 0.4761\\ \hline \end{array}$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ \textbf{Pearson's} \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ 0.3594\\ \hline 0.4587\\ 0.4801\\ 0.4683\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ \hline \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen.25-7B G-Eval (GPT-4)	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ \hline \textbf{Pearson's}\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{o.6762}\\ \hline o$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ 0.3594\\ \hline 0.4587\\ 0.4801\\ 0.4683\\ 0.5489\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ \hline \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwCn-25-7B G-Eval (GPT-4) LLM-Eval (GPT-3.5)	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.2091\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{o.6762}\\ \hline o$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ 0.5098\\ 0.4426\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ \hline 0.1616\\ 0.1673\\ \hline 0.1537\\ \hline 0.2194\\ \hline 0.2066\\ \hline 0.2225\\ \hline 0.1697\\ \hline 0.2633\\ \hline 0.3687\\ \hline 0.3687\\ \hline 0.3838\\ \hline 0.3594\\ \hline 0.4587\\ \hline 0.4801\\ \hline 0.4683\\ \hline 0.5489\\ \hline 0.4590\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3760 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.4474 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ 0.4639 \\ \hline \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4)	$\begin{array}{c} 0.6116\\ \hline \textbf{0.6441}\\ \hline \textbf{Engag}\\ \hline \textbf{Pearson's}\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ 0.5284\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ 0.5357\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ 0.5356\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline 0$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ 0.4426\\ 0.5234\\ \end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ 0.3594\\ \hline 0.4587\\ 0.4801\\ 0.4683\\ 0.5489\\ 0.4590\\ 0.5223\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.4474 \\ 0.5304 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ 0.4639 \\ 0.5292 \\ \hline \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) Ours(w/o LLM)	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.2091\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ 0.5284\\ \hline 0.3651\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{gingness}\\ \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ 0.5357\\ \hline 0.3767\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3058\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ 0.5356\\ \hline\\ 0.3685\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.67762}\\ \hline \textbf{0.67776}\\ \hline \textbf{0.677776}\\ \hline \textbf{0.6777776}\\ \hline \textbf{0.6777777776}\\ \hline \textbf{0.677777776}\\ \hline 0.6777777777777777777777777777777777777$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \text{Engag}\\ \text{Pearson's }\rho\\ \hline\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4528\\ 0.5098\\ 0.4426\\ 0.5234\\ \hline\\ 0.3934\\ \end{array}$	$\begin{array}{c} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.657\\ \hline 0.1657\\ \hline 0.1616\\ \hline 0.1673\\ \hline 0.1537\\ \hline 0.2194\\ \hline 0.2066\\ \hline 0.2225\\ \hline 0.1697\\ \hline 0.2633\\ \hline 0.3687\\ \hline 0.3687\\ \hline 0.3687\\ \hline 0.3687\\ \hline 0.3594\\ \hline 0.4587\\ \hline 0.44801\\ \hline 0.4683\\ \hline 0.5489\\ \hline 0.4590\\ \hline 0.5223\\ \hline 0.3890\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3760 \\ 0.3761 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.5146 \\ 0.4474 \\ 0.5304 \\ 0.3973 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \hline \textbf{dedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ 0.4639 \\ 0.5292 \\ \hline 0.3930 \\ \hline \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) Ours (w/o LLM) Ours (GPT-3.5 w/o AMR)	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's ρ\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2091\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ 0.5284\\ \hline 0.3651\\ 0.4978\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{0.6448}\\ \hline \textbf{gingness}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.2147\\ 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline \textbf{0.3417}\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ 0.5357\\ \hline 0.3767\\ 0.4969\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ 0.5356\\ \hline\\ 0.3685\\ 0.5037\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.67762}\\ \hline \textbf{0.6777}\\ \hline \textbf{0.67777}\\ \hline \textbf{0.6777}\\ \hline \textbf{0.6777}\\ \hline \textbf{0.6777}\\ \hline \textbf{0.6777}\\ \hline \textbf$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ \textbf{Pearson's} \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ 0.4426\\ 0.5234\\ \hline\\ 0.3934\\ 0.4984\\ \hline\end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.1657\\ \hline 0.1657\\ \hline 0.1616\\ \hline 0.1673\\ \hline 0.1537\\ \hline 0.2194\\ \hline 0.2066\\ \hline 0.2225\\ \hline 0.1697\\ \hline 0.2633\\ \hline 0.3687\\ \hline 0.3687\\ \hline 0.3838\\ \hline 0.3581\\ \hline 0.3594\\ \hline 0.4587\\ \hline 0.4801\\ \hline 0.4683\\ \hline 0.5489\\ \hline 0.4590\\ \hline 0.5223\\ \hline 0.3890\\ \hline 0.5075\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ \hline \textbf{0.1769} \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ \hline \textbf{0.3533} \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ \hline \textbf{0.4573} \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.4474 \\ 0.5304 \\ \hline \textbf{0.3973} \\ 0.5037 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.2203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.3877 \\ 0.3621 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ 0.4639 \\ 0.5292 \\ \hline 0.3930 \\ 0.5131 \\ \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) MOURS (GPT-3.5 w/o SLM)	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's $\rho\\ \hline 0.2052\\ 0.1940\\ 0.1940\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ 0.5284\\ \hline 0.3651\\ 0.4978\\ 0.5089\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{0.6448}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.34847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ 0.5357\\ \hline 0.3767\\ 0.4969\\ 0.5039\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \\ \rho\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3058\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ 0.5356\\ \hline\\ 0.3685\\ 0.5037\\ 0.5150\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline 0$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ \textbf{Pearson's} \\ \rho\\ \hline\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ 0.4426\\ 0.5234\\ \hline\\ 0.3934\\ 0.4984\\ 0.5170\\ \hline\end{array}$	$\begin{array}{r} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3584\\ \hline 0.3594\\ \hline 0.4587\\ 0.4801\\ 0.4683\\ 0.5489\\ 0.4590\\ 0.5223\\ \hline 0.3890\\ 0.5075\\ 0.5158\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour Pearson's } \rho \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.4474 \\ 0.5304 \\ 0.3973 \\ 0.5037 \\ 0.5226 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1669 \\ 0.1621 \\ 0.1680 \\ 0.1548 \\ 0.203 \\ 0.2079 \\ 0.2244 \\ 0.1711 \\ 0.2652 \\ \hline 0.3722 \\ 0.377 \\ 0.3621 \\ 0.3631 \\ \hline 0.4629 \\ 0.4852 \\ 0.4728 \\ 0.5532 \\ 0.4639 \\ 0.5292 \\ \hline 0.3930 \\ 0.5131 \\ 0.5216 \\ \end{array}$
Ours (GPT-4 w/o SLM) Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwe0.32B G-Eval (GPT-3.5 LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5 W/o SLM) Ours (GPT-3.5 w/o SLM) Ours (GPT-3.5 m/o SLM)	$\begin{array}{c} 0.6116\\ \hline 0.6441\\ \hline Engag\\ Pearson's \\ \rho\\ \hline 0.2052\\ 0.1940\\ 0.1670\\ 0.1992\\ 0.2119\\ 0.2004\\ 0.2091\\ 0.1987\\ 0.2850\\ \hline 0.3626\\ 0.3444\\ 0.3460\\ 0.3073\\ \hline 0.4862\\ 0.4997\\ 0.4762\\ 0.5209\\ 0.4624\\ 0.5284\\ \hline 0.3651\\ 0.4978\\ 0.5089\\ 0.5481\\ \hline \end{array}$	$\begin{array}{r} 0.6149\\ \textbf{0.6448}\\ \hline \textbf{0.6448}\\ \hline \textbf{gingness}\\ \hline \textbf{Spearman's }\tau\\ \hline 0.2142\\ 0.1813\\ 0.1933\\ 0.1921\\ 0.2147\\ 0.2014\\ 0.2024\\ 0.1758\\ 0.2707\\ \hline 0.3417\\ 0.3434\\ 0.3392\\ 0.3087\\ \hline 0.4847\\ 0.4981\\ 0.4704\\ 0.5290\\ 0.4548\\ 0.5357\\ \hline 0.3767\\ 0.4969\\ 0.5039\\ 0.5175\\ \hline \end{array}$	$\begin{array}{c} 0.6418\\ \textbf{0.6756}\\ \hline\\ \textbf{Grour}\\ \textbf{Pearson's} \ \rho\\ \hline\\ 0.2072\\ 0.1961\\ 0.1688\\ 0.2010\\ 0.2139\\ 0.2027\\ 0.2116\\ 0.2005\\ 0.2881\\ \hline\\ 0.3678\\ 0.3488\\ 0.3503\\ 0.3114\\ \hline\\ 0.4928\\ 0.5058\\ 0.4822\\ 0.5271\\ 0.4673\\ 0.5356\\ \hline\\ 0.3685\\ 0.5037\\ 0.5150\\ 0.5553\\ \hline\end{array}$	$\begin{array}{r} 0.6450\\ \textbf{0.6762}\\ \hline \textbf{0.6762}\\ \hline \textbf{o.6762}\\ \hline o$	$\begin{array}{c} 0.6094\\ \textbf{0.6370}\\ \hline\\ \textbf{Engag}\\ Pearson's \\ \rho\\ 0.1755\\ 0.1397\\ 0.1367\\ 0.1244\\ 0.2067\\ 0.1823\\ 0.2017\\ 0.1430\\ 0.2240\\ \hline\\ 0.3489\\ 0.3663\\ 0.3712\\ 0.3744\\ \hline\\ 0.4527\\ 0.4749\\ 0.4598\\ 0.5098\\ 0.4426\\ 0.5234\\ \hline\\ 0.5934\\ 0.4984\\ 0.5170\\ 0.5173\\ \hline\end{array}$	$\begin{array}{c} 0.6158\\ \hline 0.6458\\ \hline 0.6458\\ \hline 0.6458\\ \hline Spearman's \ \tau\\ \hline 0.1657\\ 0.1616\\ 0.1673\\ 0.1537\\ 0.2194\\ 0.2066\\ 0.2225\\ 0.1697\\ 0.2633\\ \hline 0.3687\\ 0.3838\\ 0.3581\\ 0.3594\\ \hline 0.4587\\ 0.4801\\ 0.4683\\ 0.5489\\ 0.4590\\ 0.5223\\ \hline 0.3890\\ 0.5075\\ 0.5158\\ 0.5194\\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6683} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ 0.1769 \\ 0.1410 \\ 0.1380 \\ 0.1260 \\ 0.2082 \\ 0.1839 \\ 0.2036 \\ 0.1446 \\ 0.2265 \\ 0.3533 \\ 0.3698 \\ 0.3760 \\ 0.3781 \\ 0.4573 \\ 0.4809 \\ 0.4645 \\ 0.5146 \\ 0.4474 \\ 0.5304 \\ 0.3973 \\ 0.5037 \\ 0.5226 \\ 0.5234 \\ \end{array}$	$\begin{array}{r} \textbf{0.6768} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1669} \\ \textbf{0.1621} \\ \textbf{0.1680} \\ \textbf{0.1548} \\ \textbf{0.2203} \\ \textbf{0.2079} \\ \textbf{0.2244} \\ \textbf{0.711} \\ \textbf{0.2652} \\ \hline \textbf{0.3722} \\ \textbf{0.3722} \\ \textbf{0.3877} \\ \textbf{0.3621} \\ \textbf{0.3631} \\ \hline \textbf{0.4629} \\ \textbf{0.4852} \\ \textbf{0.4728} \\ \textbf{0.5532} \\ \textbf{0.4728} \\ \textbf{0.5532} \\ \textbf{0.4639} \\ \textbf{0.5292} \\ \hline \textbf{0.3930} \\ \textbf{0.5131} \\ \textbf{0.5216} \\ \textbf{0.5253} \\ \hline \end{array}$

Table 7: Breakdown of Pearson and Spearman correlations with human judgments by evaluation criteria on the PersonaChat dataset.

		Standa	ard Set			Adversa	arial Set	
		ralness		erence		ralness		erence
Metrics	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ	Pearson's ρ	Spearman's τ
BLEU-1	0.2055	0.1938	0.2149	0.2026	0.1408	0.1517	0.1480	0.1588
BLEU-2	0.1682	0.1732	0.1759	0.1812	0.1261	0.1403	0.1330	0.1475
BLEU-3	0.1541	0.1605	0.1613	0.1679	0.1191	0.1293	0.1259	0.1363
BLEU-4	0.1449	0.1468	0.1515	0.1538	0.1289	0.1198	0.1358	0.1258
ROUGE-1	0.2004	0.2095	0.2096	0.2193	0.1710	0.1747	0.1793	0.1829
ROUGE-2	0.1958	0.1982	0.2049	0.2074	0.1789	0.1983	0.1880	0.2075
ROUGE-L	0.2148	0.1965	0.2246	0.2056	0.1863	0.2283	0.1953	0.2388
METEOR	0.1815	0.1539	0.1899	0.1612	0.1483	0.1647	0.1554	0.1724
BERTScore	0.2497	0.2485	0.2613	0.2600	0.2142	0.2501	0.2245	0.2617
DEB	0.3181	0.3230	0.3330	0.3381	0.3341	0.3585	0.3497	0.3752
USR	0.3387	0.3350	0.3545	0.3507	0.3258	0.1667	0.3419	0.1745
MDD-Eval	0.3201	0.3320	0.3353	0.3476	0.3777	0.3475	0.3961	0.3638
Mask-and-fill	0.2929	0.2982	0.3068	0.3123	0.3582	0.3545	0.3754	0.3710
G-Eval (GPT-3.5)	0.4880	0.4645	0.5108	0.4866	0.4663	0.4578	0.4883	0.4798
QwQ-32B	0.4975	0.4724	0.5208	0.4950	0.4773	0.4714	0.5004	0.4936
Qwq-52B Qwen2.5-7B	0.4814	0.4596	0.5041	0.4930	0.4591	0.4569	0.4812	0.4788
G-Eval (GPT-4)	0.5192	0.4939	0.5434	0.5173	0.4880	0.4907	0.5107	0.5137
LLM-Eval (GPT-3.5)	0.4724	0.4687	0.4949	0.4909	0.4408	0.4688	0.4618	0.4909
LLM-Eval (GPT-4)	0.4893	0.4979	0.5123	0.5214	0.5059	0.5137	0.5298	0.5378
Ours(w/o_LLM)	0.3520	0.3516	0.3685	0.3681	0.3529	0.3505	0.3694	0.3669
Ours (GPT-3.5 w/o AMR)	0.4907	0.5003	0.5138	0.5237	0.5001	0.4981	0.5235	0.5216
Ours (GPT-3.5 w/o SLM)	0.5053	0.4981	0.5290	0.5217	0.4994	0.4983	0.5228	0.5218
Ours(GPT-3.5)	0.5080	0.4997	0.5319	0.5233	0.5009	0.4993	0.5245	0.5228
Ours (GPT-4 w/o AMR)	0.6130	0.6122	0.6417	0.6410	0.6055	0.5087	0.6338	0.5326
Ours (GPT-4 w/o SLM)	0.6325	06336	0.6615	0.6629	0.6251	0.6258	0.6546	0.6546
Ours (GPT-4 w/o SLM) Ours (GPT-4)	0.6325 0.6492	0.6336 0.6455	0.6615 0.6791	0.6629 0.6752	0.6251 0.6449	0.6258 0.6524	0.6546 0.6748	0.6546 0.6822
	0.6492	0.6455	0.6791	0.6752	0.6449	0.6524	0.6748	0.6822
	0.6492		0.6791		0.6449		0.6748	
Ours (GPT-4) Metrics	0.6492 Engag Pearson's ρ	0.6455 gingness Spearman's τ	0.6791 Grour Pearson's ρ	0.6752 idedness Spearman's τ	0.6449 Engag Pearson's ρ	0.6524 gingness Spearman's τ	0.6748 Grour Pearson's ρ	0.6822 idedness Spearman's τ
Ours (GPT-4) Metrics BLEU-1	0.6492 Engag Pearson's <i>ρ</i> 0.2091	0.6455 gingness Spearman's τ 0.1971	0.6791 Grour Pearson's <i>ρ</i> 0.2111	0.6752 dedness Spearman's τ 0.1984	0.6449 Engag Pearson's <i>ρ</i> 0.1435	0.6524 gingness Spearman's τ 0.1546	0.6748 Grour Pearson's <i>ρ</i> 0.1452	0.6822 dedness Spearman's τ 0.1561
Ours (GPT-4) Metrics BLEU-1 BLEU-2	0.6492 Engag Pearson's ρ 0.2091 0.1714	0.6455 gingness Spearman's τ 0.1971 0.1765	0.6791 Grour Pearson's <i>ρ</i> 0.2111 0.1731	0.6752 dedness Spearman's τ 0.1984 0.1777	0.6449 Engag Pearson's ρ 0.1435 0.1287	0.6524 gingness Spearman's τ 0.1546 0.1429	0.6748 Grour Pearson's ρ 0.1452 0.1304	0.6822 dedness Spearman's τ 0.1561 0.1449
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215	0.6524 gingness Spearman's τ 0.1546 0.1429 0.1318	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339
Ours (GPT-4) Metrics BLEU-1 BLEU-2	0.6492 Engag Pearson's ρ 0.2091 0.1714	0.6455 gingness Spearman's τ 0.1971 0.1765	0.6791 Grour Pearson's <i>ρ</i> 0.2111 0.1731	0.6752 dedness Spearman's τ 0.1984 0.1777	0.6449 Engag Pearson's ρ 0.1435 0.1287	0.6524 gingness Spearman's τ 0.1546 0.1429	0.6748 Grour Pearson's ρ 0.1452 0.1304	0.6822 dedness Spearman's τ 0.1561 0.1449
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.1647 0.1510 0.2151	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742	0.6524 spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1233 0.1764	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.1510 0.2151 0.2035	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824	0.6524 gingness Spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.1647 0.1510 0.2151	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742	0.6524 spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1233 0.1764	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207	0.6752 hdedness Spearman's τ 0.1984 0.1777 0.1647 0.1510 0.2151 0.2035 0.2023	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829	0.6524 spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020 0.2325	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847 0.1922	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038 0.2344
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.2151 0.2035 0.2023 0.1575	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829 0.1509	0.6524 spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020 0.2325 0.1675	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847 0.1922 0.1527	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038 0.2038 0.2344 0.1693
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.2568 0.3273	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.1510 0.2035 0.2023 0.1575 0.2553 0.3324	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829 0.1509 0.2183 0.3402	0.6524 singness Spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020 0.2325 0.1675 0.2547 0.3649	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038 0.2344 0.1693 0.2575 0.3689
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288 0.3409	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.2568 0.3273 0.3487	0.6752 hdedness Spearman's τ 0.1984 0.1777 0.1647 0.1510 0.2151 0.2035 0.2023 0.1575 0.2553 0.3324 0.3448	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1742 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3320	0.6524 gingness Spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020 0.2325 0.1675 0.2547 0.3649 0.1698	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439 0.3358	0.6822 addeness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038 0.2344 0.1693 0.2575 0.3689 0.1715
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.2568 0.3273	0.6752 dedness Spearman's τ 0.1984 0.1777 0.1647 0.1510 0.2035 0.2023 0.1575 0.2553 0.3324	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829 0.1509 0.2183 0.3402	0.6524 singness Spearman's τ 0.1546 0.1429 0.1318 0.1220 0.1780 0.2020 0.2325 0.1675 0.2547 0.3649	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439	0.6822 dedness Spearman's τ 0.1561 0.1449 0.1339 0.1235 0.1796 0.2038 0.2344 0.1693 0.2575 0.3689
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3248 0.3259 0.2983	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.3288 0.3409 0.3037	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.2111} \\ \textbf{0.1731} \\ \textbf{0.1585} \\ \textbf{0.1585} \\ \textbf{0.2065} \\ \textbf{0.2015} \\ \textbf{0.2005} \\ \textbf{0.2005} \\ \textbf{0.2015} \\ \textbf{0.2207} \\ \textbf{0.1868} \\ \textbf{0.2568} \\ \hline \textbf{0.32568} \\ \textbf{0.3273} \\ \textbf{0.3487} \\ \textbf{0.3296} \\ \textbf{0.3012} \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \textbf{adedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1984 \\ 0.1777 \\ 0.1647 \\ 0.1510 \\ 0.2151 \\ 0.2035 \\ 0.2023 \\ 0.1575 \\ 0.2553 \\ \hline 0.3324 \\ 0.3448 \\ 0.3415 \\ 0.3067 \\ \hline \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3847 0.3649	$\begin{array}{c} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439 0.3358 0.3892 0.3688	$\begin{array}{r} \textbf{0.6822} \\ \textbf{adedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.1693 \\ 0.2575 \\ \hline 0.3689 \\ 0.1715 \\ 0.3576 \\ 0.3646 \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3237 0.3248 0.3259 0.2983 0.4970	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.3288 0.3409 0.3380 0.3037 0.4728	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.22568 0.3273 0.3487 0.3296 0.3012 0.5033	$\begin{array}{r} \textbf{0.6752} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1984 \\ 0.1777 \\ 0.1647 \\ 0.1510 \\ 0.2151 \\ 0.2035 \\ 0.2023 \\ 0.1575 \\ 0.2553 \\ \hline 0.2553 \\ 0.3324 \\ 0.3448 \\ 0.3415 \\ 0.3067 \\ \hline 0.4775 \\ \hline \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3402 0.3847 0.3649 0.4750	$\begin{array}{c} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847 0.1922 0.1527 0.2207 0.3439 0.3358 0.3892 0.3688 0.4800	$\begin{array}{r} \textbf{0.6822} \\ \textbf{dedness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.1693} \\ \textbf{0.2575} \\ \hline \textbf{0.3689} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3248 0.3259 0.2983	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.3288 0.3409 0.3037	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.2111} \\ \textbf{0.1731} \\ \textbf{0.1585} \\ \textbf{0.1585} \\ \textbf{0.2065} \\ \textbf{0.2015} \\ \textbf{0.2005} \\ \textbf{0.2005} \\ \textbf{0.2015} \\ \textbf{0.2207} \\ \textbf{0.1868} \\ \textbf{0.2568} \\ \hline \textbf{0.32568} \\ \textbf{0.3273} \\ \textbf{0.3487} \\ \textbf{0.3296} \\ \textbf{0.3012} \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \textbf{adedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1984 \\ 0.1777 \\ 0.1647 \\ 0.1510 \\ 0.2151 \\ 0.2035 \\ 0.2023 \\ 0.1575 \\ 0.2553 \\ \hline 0.3324 \\ 0.3448 \\ 0.3415 \\ 0.3067 \\ \hline \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3847 0.3649	$\begin{array}{c} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439 0.3358 0.3892 0.3688	$\begin{array}{r} \textbf{0.6822} \\ \textbf{adedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.1693 \\ 0.2575 \\ \hline 0.3689 \\ 0.1715 \\ 0.3576 \\ 0.3646 \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwc9.32B Qwen2.5-7B	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.2983	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288 0.3409 0.3380 0.4728 0.4881	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.2568 0.3273 0.3487 0.3296 0.3012 0.5033 0.5127 0.4962	$\begin{array}{r} \textbf{0.6752} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1984 \\ 0.1777 \\ 0.1647 \\ 0.1510 \\ 0.2151 \\ 0.2035 \\ 0.2023 \\ 0.1575 \\ 0.2035 \\ 0.2023 \\ 0.3057 \\ \hline 0.3324 \\ 0.3448 \\ 0.3415 \\ 0.3067 \\ \hline 0.4775 \\ 0.4852 \\ 0.4731 \\ \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829 0.1509 0.2183 0.3402 0.3320 0.3847 0.3649 0.4750 0.4863 0.4678	$\begin{array}{c} \textbf{0.6524} \\ \hline \textbf{spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4665} \\ \textbf{0.4652} \\ \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847 0.1922 0.1527 0.2207 0.3439 0.3358 0.3688 0.4800 0.4909 0.4727	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.1693} \\ \textbf{0.2575} \\ \hline \textbf{0.3649} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \\ \textbf{0.4846} \\ \textbf{0.4694} \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen2.5-7B G-Eval (GPT-4)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5067 0.4902 0.5286	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.3288 0.3409 0.3380 0.3037 0.4728 0.4809 0.4681 0.5027	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.2111} \\ \textbf{0.1731} \\ \textbf{0.1585} \\ \textbf{0.1585} \\ \textbf{0.2065} \\ \textbf{0.2015} \\ \textbf{0.2005} \\ \textbf{0.2005} \\ \textbf{0.2015} \\ \textbf{0.2207} \\ \textbf{0.2207} \\ \textbf{0.1868} \\ \textbf{0.2568} \\ \textbf{0.3273} \\ \textbf{0.3487} \\ \textbf{0.3296} \\ \textbf{0.3012} \\ \textbf{0.5033} \\ \textbf{0.5127} \\ \textbf{0.4962} \\ \textbf{0.5345} \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.2023} \\ \textbf{0.2023} \\ \textbf{0.2023} \\ \textbf{0.2023} \\ \textbf{0.2053} \\ \textbf{0.2053} \\ \hline \textbf{0.3253} \\ \hline \textbf{0.3324} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3067} \\ \hline \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4731} \\ \textbf{0.5080} \\ \end{array}$	$\begin{array}{c} \textbf{0.6449} \\ \hline \textbf{Engag} \\ Pearson's ρ \\ \hline 0.1435 \\ 0.1287 \\ 0.1215 \\ 0.1312 \\ 0.1742 \\ 0.1824 \\ 0.1899 \\ 0.1509 \\ 0.2183 \\ \hline 0.3402 \\ 0.3320 \\ 0.3847 \\ 0.3649 \\ \hline 0.4750 \\ 0.4863 \\ 0.4678 \\ 0.4970 \\ \hline \end{array}$	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4801} \\ \textbf{0.4652} \\ \textbf{0.4996} \\ \hline \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439 0.3358 0.3892 0.3688 0.4800 0.4909 0.4727 0.5022	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.693} \\ \textbf{0.2575} \\ \hline \textbf{0.3689} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \\ \textbf{0.4846} \\ \textbf{0.4694} \\ \textbf{0.5048} \\ \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwc9.32B Qwen2.5-7B	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.2983	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288 0.3409 0.3380 0.4728 0.4728 0.4881	0.6791 Grour Pearson's ρ 0.2111 0.1731 0.1585 0.1490 0.2065 0.2015 0.2207 0.1868 0.2568 0.3273 0.3487 0.3296 0.3012 0.5033 0.5127 0.4962	$\begin{array}{r} \textbf{0.6752} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1984 \\ 0.1777 \\ 0.1647 \\ 0.1510 \\ 0.2151 \\ 0.2035 \\ 0.2023 \\ 0.1575 \\ 0.2035 \\ 0.2023 \\ 0.3057 \\ \hline 0.3324 \\ 0.3448 \\ 0.3415 \\ 0.3067 \\ \hline 0.4775 \\ 0.4852 \\ 0.4731 \\ \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1829 0.1509 0.2183 0.3402 0.3320 0.3847 0.3649 0.4750 0.4863 0.4678	$\begin{array}{c} \textbf{0.6524} \\ \hline \textbf{spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4665} \\ \textbf{0.4652} \\ \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1847 0.1922 0.1527 0.2207 0.3439 0.3358 0.3688 0.4800 0.4909 0.4727	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.1693} \\ \textbf{0.2575} \\ \hline \textbf{0.3649} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \\ \textbf{0.4846} \\ \textbf{0.4694} \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5067 0.4902 0.5286 0.4811 0.4981	$\begin{array}{r} \textbf{0.6455} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1971} \\ \textbf{0.1765} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.2530} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.4728} \\ \textbf{0.4809} \\ \textbf{0.4681} \\ \textbf{0.5027} \\ \textbf{0.4773} \\ \textbf{0.5069} \\ \end{array}$	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.2111} \\ \textbf{0.1731} \\ \textbf{0.1585} \\ \textbf{0.1585} \\ \textbf{0.1490} \\ \textbf{0.2065} \\ \textbf{0.2015} \\ \textbf{0.2015} \\ \textbf{0.2207} \\ \textbf{0.1868} \\ \textbf{0.2568} \\ \hline \textbf{0.3273} \\ \textbf{0.3273} \\ \textbf{0.3487} \\ \textbf{0.3296} \\ \textbf{0.3012} \\ \hline \textbf{0.5033} \\ \textbf{0.5127} \\ \textbf{0.4962} \\ \textbf{0.5345} \\ \textbf{0.4864} \\ \textbf{0.5036} \\ \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2151} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.2023} \\ \textbf{0.2553} \\ \hline \textbf{0.2553} \\ \hline \textbf{0.3324} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3067} \\ \hline \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4731} \\ \textbf{0.5080} \\ \textbf{0.4823} \\ \textbf{0.5125} \\ \end{array}$	$\begin{array}{c} \textbf{0.6449} \\ \hline \textbf{Engag} \\ Pearson's ρ \\ \hline 0.1435 \\ 0.1287 \\ 0.1215 \\ 0.1215 \\ 0.1312 \\ 0.1742 \\ 0.1824 \\ 0.1829 \\ 0.1509 \\ 0.2183 \\ 0.3402 \\ 0.3320 \\ 0.3320 \\ 0.3847 \\ 0.3649 \\ \hline 0.4750 \\ 0.4863 \\ 0.4678 \\ 0.4970 \\ 0.4489 \\ 0.5147 \\ \end{array}$	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{g} \textbf{ingness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4801} \\ \textbf{0.4652} \\ \textbf{0.4996} \\ \textbf{0.4772} \\ \textbf{0.5226} \\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6748} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.1452} \\ 0.1304 \\ 0.1231 \\ 0.1333 \\ 0.1764 \\ 0.1847 \\ 0.1922 \\ 0.1527 \\ 0.2207 \\ \hline \textbf{0.3439} \\ 0.3358 \\ 0.3892 \\ 0.3688 \\ \hline \textbf{0.4800} \\ 0.4909 \\ 0.4727 \\ 0.5022 \\ 0.4532 \\ 0.5207 \\ \hline \textbf{0.5207} \\ \hline 0.520$	$\begin{array}{r} \textbf{0.6822} \\ \textbf{adedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.693 \\ 0.2575 \\ \hline 0.3689 \\ 0.1715 \\ 0.3576 \\ 0.3646 \\ \hline 0.4710 \\ 0.4846 \\ 0.4694 \\ 0.5048 \\ 0.4826 \\ 0.5287 \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-3.5) LLM-Eval (GPT-4) Ours(w/o LLM)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5286 0.4811 0.3584	$\begin{array}{r} \textbf{0.6455} \\ \hline \textbf{gingness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1971} \\ \textbf{0.1765} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.2130} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.1567} \\ \textbf{0.2530} \\ \hline \textbf{0.2530} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.4728} \\ \textbf{0.4809} \\ \textbf{0.4681} \\ \textbf{0.5027} \\ \textbf{0.4773} \\ \textbf{0.5069} \\ \hline \textbf{0.3580} \\ \end{array}$	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's ρ} \\ \hline \textbf{0.2111} \\ 0.1731 \\ 0.1585 \\ 0.1490 \\ 0.2065 \\ 0.2015 \\ 0.2007 \\ 0.1868 \\ 0.22568 \\ \hline \textbf{0.3273} \\ 0.3273 \\ 0.3487 \\ 0.3296 \\ 0.3012 \\ \hline \textbf{0.5033} \\ 0.5127 \\ 0.4962 \\ 0.5345 \\ 0.4864 \\ 0.5036 \\ \hline \textbf{0.3621} \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.2035} \\ \textbf{0.2035}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3320 0.3847 0.3649 0.4750 0.4863 0.4678 0.4970 0.4489 0.5147 0.3593	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline 0.1546 \\ 0.1429 \\ 0.1318 \\ 0.1220 \\ 0.1780 \\ 0.2020 \\ 0.2325 \\ 0.1675 \\ 0.2547 \\ \hline 0.3649 \\ 0.1698 \\ 0.3539 \\ 0.3609 \\ \hline 0.4665 \\ 0.4801 \\ 0.4652 \\ 0.4996 \\ 0.4772 \\ 0.5226 \\ \hline 0.3569 \\ \hline \end{array}$	0.6748 Grour Pearson's ρ 0.1452 0.1304 0.1231 0.1333 0.1764 0.1922 0.1527 0.2207 0.3439 0.3558 0.3892 0.3688 0.4800 0.4909 0.5022 0.5207 0.3629	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.1693} \\ \textbf{0.2575} \\ \hline \textbf{0.3689} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \\ \textbf{0.4846} \\ \textbf{0.4694} \\ \textbf{0.5048} \\ \textbf{0.4826} \\ \textbf{0.5287} \\ \hline \textbf{0.3605} \\ \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B QweQ-32B QweQ-32B QweQ-32B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-4) Ours (w/o LLM) Ours (GPT-3.5 w/o AMR)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5286 0.4811 0.4981 0.3584 0.4997	0.6455 gingness Spearman's τ 0.1971 0.1765 0.1635 0.1494 0.2130 0.2019 0.2000 0.1567 0.2530 0.3288 0.3409 0.3380 0.3037 0.4728 0.4681 0.5027 0.4773 0.5069 0.3580 0.5094	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's ρ} \\ \hline \textbf{0.2111} \\ 0.1731 \\ 0.1585 \\ 0.1490 \\ 0.2065 \\ 0.2015 \\ 0.2007 \\ 0.1868 \\ 0.2568 \\ \hline \textbf{0.3273} \\ 0.3296 \\ 0.3012 \\ \hline \textbf{0.3012} \\ 0.5033 \\ 0.5127 \\ 0.4962 \\ 0.5345 \\ 0.4864 \\ 0.5036 \\ \hline \textbf{0.3621} \\ 0.5046 \\ \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2151} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.1575} \\ \textbf{0.2023} \\ \textbf{0.1575} \\ \textbf{0.2553} \\ \hline \textbf{0.3324} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3445} \\ \textbf{0.3452} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.47731} \\ \textbf{0.5080} \\ \textbf{0.4823} \\ \textbf{0.5125} \\ \hline \textbf{0.3616} \\ \textbf{0.5145} \\ \end{array}$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3402 0.3404 0.4750 0.4483 0.4678 0.4970 0.5147 0.3593 0.5093	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4801} \\ \textbf{0.4652} \\ \textbf{0.4996} \\ \textbf{0.4772} \\ \textbf{0.5226} \\ \hline \textbf{0.3569} \\ \textbf{0.5074} \\ \end{array}$	$\begin{array}{c} \textbf{0.6748} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ \hline 0.1452 \\ 0.1304 \\ 0.1231 \\ 0.1333 \\ 0.1764 \\ 0.1847 \\ 0.1922 \\ 0.1527 \\ 0.2207 \\ \hline 0.3439 \\ 0.3358 \\ 0.3892 \\ 0.3688 \\ \hline 0.4800 \\ 0.4909 \\ 0.4727 \\ 0.5022 \\ 0.4532 \\ 0.5207 \\ \hline 0.3629 \\ 0.5143 \\ \end{array}$	$\begin{array}{r} \textbf{0.6822} \\ \textbf{idedness} \\ \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.1693 \\ 0.2575 \\ \hline 0.3649 \\ 0.3576 \\ 0.3646 \\ \hline 0.4710 \\ 0.4846 \\ 0.4694 \\ 0.5048 \\ 0.4826 \\ 0.5287 \\ \hline 0.3605 \\ 0.5124 \\ \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-3.5) LLM-Eval (GPT-4) Ours (WO LLM) Ours (GPT-3.5 w/o AMR) Ours (GPT-3.5 w/o SLM)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5067 0.4902 0.5286 0.4811 0.4981 0.3584 0.4997 0.5146	$\begin{array}{r} \textbf{0.6455} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{0.1971} \\ \textbf{0.1765} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.1567} \\ \textbf{0.2530} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.4728} \\ \textbf{0.4809} \\ \textbf{0.4681} \\ \textbf{0.5027} \\ \textbf{0.4773} \\ \textbf{0.5069} \\ \hline \textbf{0.3580} \\ \textbf{0.5094} \\ \textbf{0.5073} \\ \end{array}$	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's ρ} \\ \hline \textbf{0.2111} \\ 0.1731 \\ 0.1585 \\ 0.1490 \\ 0.2065 \\ 0.2015 \\ 0.2007 \\ 0.1868 \\ 0.2568 \\ \hline \textbf{0.3273} \\ 0.3296 \\ 0.3012 \\ \hline \textbf{0.3012} \\ 0.5033 \\ 0.5127 \\ 0.4962 \\ 0.5036 \\ \hline \textbf{0.5036} \\ 0.3621 \\ 0.5046 \\ 0.5200 \\ \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2151} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.1575} \\ \textbf{0.2023} \\ \textbf{0.3244} \\ \textbf{0.3448} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4771} \\ \textbf{0.5080} \\ \textbf{0.4823} \\ \textbf{0.5125} \\ \hline \begin{array}{r} \textbf{0.3616} \\ \textbf{0.5145} \\ \textbf{0.5124} \\ \hline \end{array} \right.$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3847 0.3649 0.4750 0.4863 0.4678 0.4970 0.5147 0.3593 0.5093 0.5089	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{spearman's τ} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4801} \\ \textbf{0.4652} \\ \textbf{0.4996} \\ \textbf{0.4772} \\ \textbf{0.5226} \\ \hline \textbf{0.3569} \\ \textbf{0.5074} \\ \textbf{0.5076} \\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6748} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.1452} \\ 0.1304 \\ 0.1231 \\ 0.1333 \\ 0.1764 \\ 0.1847 \\ 0.1922 \\ 0.1527 \\ 0.2207 \\ \hline \textbf{0.3439} \\ 0.3358 \\ 0.3892 \\ 0.3688 \\ \hline \textbf{0.4800} \\ 0.4909 \\ 0.4727 \\ 0.5022 \\ 0.4532 \\ 0.5207 \\ \hline \textbf{0.3629} \\ 0.5143 \\ 0.5136 \\ \end{array}$	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.1693 \\ 0.2575 \\ \hline 0.3649 \\ 0.715 \\ 0.3576 \\ 0.3646 \\ \hline 0.4710 \\ 0.4846 \\ 0.4694 \\ 0.5048 \\ 0.4826 \\ 0.5287 \\ \hline 0.3605 \\ 0.5124 \\ 0.5127 \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-1 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval Mask-and-fill G-Eval (GPT-3.5) Qwen2.5-7B G-Eval (GPT-4) LLM-Eval (GPT-4) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) MOURS (GPT-3.5) MOURS (GPT-3.5) Ours (GPT-3.5) Ours (GPT-3.5) Ours (GPT-3.5) MOURS (GPT-3.5) Ours (GPT-3.5) MOURS (GPT-3.5) Ours (GPT-3.5) MOURS (GP	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.1993 0.2186 0.1845 0.2542 0.3237 0.3248 0.3259 0.2983 0.4970 0.5067 0.4902 0.5286 0.4811 0.4981 0.3584 0.5146 0.5146	$\begin{array}{r} \textbf{0.6455} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1971} \\ \textbf{0.1765} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.1567} \\ \textbf{0.2530} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.4728} \\ \textbf{0.4809} \\ \textbf{0.4681} \\ \textbf{0.5027} \\ \textbf{0.4773} \\ \textbf{0.5069} \\ \hline \textbf{0.3580} \\ \textbf{0.5094} \\ \textbf{0.5073} \\ \textbf{0.5089} \\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour} \\ Pearson's ρ \\ \hline 0.2111 \\ 0.1731 \\ 0.1585 \\ 0.1585 \\ 0.2065 \\ 0.2015 \\ 0.2007 \\ 0.2065 \\ 0.2015 \\ 0.2207 \\ 0.3296 \\ 0.32568 \\ \hline 0.3273 \\ 0.3487 \\ 0.3296 \\ 0.3012 \\ \hline 0.5033 \\ 0.5127 \\ 0.4962 \\ 0.5345 \\ 0.4962 \\ 0.5345 \\ 0.4864 \\ 0.5036 \\ \hline 0.3621 \\ 0.5046 \\ 0.5200 \\ 0.5226 \\ \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.2035} \\ \textbf{0.3067} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3067} \\ \textbf{0.3448} \\ \textbf{0.3415} \\ \textbf{0.3067} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4731} \\ \textbf{0.5080} \\ \textbf{0.4823} \\ \textbf{0.5125} \\ \textbf{0.3616} \\ \textbf{0.5124} \\ \textbf{0.5142} \\ \textbf{0.5142} \\ \end{array}$	$\begin{array}{c} \textbf{0.6449} \\ \hline \textbf{Engag} \\ Pearson's ρ \\ \hline 0.1435 \\ 0.1287 \\ 0.1215 \\ 0.1215 \\ 0.1312 \\ 0.1742 \\ 0.1824 \\ 0.1899 \\ 0.1509 \\ 0.2183 \\ \hline 0.3402 \\ 0.3320 \\ 0.3847 \\ 0.3649 \\ \hline 0.4750 \\ 0.4863 \\ 0.4678 \\ 0.4970 \\ 0.4489 \\ 0.5147 \\ \hline 0.3593 \\ 0.5093 \\ 0.5089 \\ 0.5102 \\ \end{array}$	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{gingness} \\ \hline \textbf{Spearman's } \tau \\ \hline 0.1546 \\ 0.1429 \\ 0.1318 \\ 0.1220 \\ 0.1780 \\ 0.2020 \\ 0.2325 \\ 0.1675 \\ 0.2547 \\ \hline 0.3649 \\ 0.1698 \\ 0.3539 \\ 0.3609 \\ \hline 0.4665 \\ 0.4801 \\ 0.4652 \\ 0.4996 \\ 0.4772 \\ 0.5226 \\ \hline 0.3569 \\ 0.5074 \\ 0.5076 \\ 0.5087 \\ \hline \end{array}$	$\begin{array}{r} \textbf{0.6748} \\ \hline \textbf{Grour} \\ \textbf{Pearson's } \rho \\ \hline \textbf{0.1452} \\ 0.1304 \\ 0.1231 \\ 0.1333 \\ 0.1764 \\ 0.1922 \\ 0.1527 \\ 0.2207 \\ \hline \textbf{0.3439} \\ 0.3358 \\ 0.3892 \\ 0.3688 \\ \hline \textbf{0.4800} \\ 0.4909 \\ 0.4727 \\ 0.5022 \\ 0.4532 \\ 0.5207 \\ \hline \textbf{0.5022} \\ 0.5207 \\ \hline \textbf{0.3629} \\ 0.5143 \\ 0.5136 \\ 0.5153 \\ \hline \textbf{0.5153} \end{array}$	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline 0.1561 \\ 0.1449 \\ 0.1339 \\ 0.1235 \\ 0.1796 \\ 0.2038 \\ 0.2344 \\ 0.693 \\ 0.2575 \\ \hline 0.3689 \\ 0.1715 \\ 0.3576 \\ 0.3689 \\ 0.1715 \\ 0.3576 \\ 0.3646 \\ \hline 0.4710 \\ 0.4846 \\ 0.4694 \\ 0.4694 \\ 0.5048 \\ 0.4826 \\ 0.5287 \\ \hline 0.3605 \\ 0.5127 \\ 0.5122 \\ \hline 0.5132 \\ \hline \end{array}$
Ours (GPT-4) Metrics BLEU-1 BLEU-2 BLEU-2 BLEU-3 ROUGE-1 ROUGE-2 ROUGE-2 ROUGE-L METEOR BERTScore DEB USR MDD-Eval MBD-Eval Mask-and-fill G-Eval (GPT-3.5) QwQ-32B Qwen2.5-7B G-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) LLM-Eval (GPT-3.5) MR Ours (GPT-3.5 w/o AMR) Ours (GPT-3.5 w/o SLM)	0.6492 Engag Pearson's ρ 0.2091 0.1714 0.1569 0.1474 0.2037 0.2186 0.1845 0.2542 0.3237 0.3448 0.3259 0.2983 0.4970 0.5067 0.4902 0.5286 0.4811 0.4981 0.3584 0.4997 0.5146	$\begin{array}{r} \textbf{0.6455} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{0.1971} \\ \textbf{0.1765} \\ \textbf{0.1635} \\ \textbf{0.1635} \\ \textbf{0.2130} \\ \textbf{0.2019} \\ \textbf{0.2000} \\ \textbf{0.1567} \\ \textbf{0.2530} \\ \hline \textbf{0.3288} \\ \textbf{0.3409} \\ \textbf{0.3380} \\ \textbf{0.3037} \\ \hline \textbf{0.4728} \\ \textbf{0.4809} \\ \textbf{0.4681} \\ \textbf{0.5027} \\ \textbf{0.4773} \\ \textbf{0.5069} \\ \hline \textbf{0.3580} \\ \textbf{0.5094} \\ \textbf{0.5073} \\ \end{array}$	$\begin{array}{c} \textbf{0.6791} \\ \hline \textbf{Grour Pearson's ρ} \\ \hline \textbf{0.2111} \\ 0.1731 \\ 0.1585 \\ 0.1490 \\ 0.2065 \\ 0.2015 \\ 0.2007 \\ 0.1868 \\ 0.2568 \\ \hline \textbf{0.3273} \\ 0.3296 \\ 0.3012 \\ \hline \textbf{0.3012} \\ 0.5033 \\ 0.5127 \\ 0.4962 \\ 0.5036 \\ \hline \textbf{0.5036} \\ 0.3621 \\ 0.5046 \\ 0.5200 \\ \end{array}$	$\begin{array}{r} \textbf{0.6752} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1984} \\ \textbf{0.1777} \\ \textbf{0.1647} \\ \textbf{0.1510} \\ \textbf{0.2151} \\ \textbf{0.2035} \\ \textbf{0.2023} \\ \textbf{0.1575} \\ \textbf{0.2023} \\ \textbf{0.3244} \\ \textbf{0.3448} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.3445} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4775} \\ \textbf{0.4852} \\ \textbf{0.4771} \\ \textbf{0.5080} \\ \textbf{0.4823} \\ \textbf{0.5125} \\ \hline \begin{array}{r} \textbf{0.3616} \\ \textbf{0.5145} \\ \textbf{0.5124} \\ \hline \end{array} \right.$	0.6449 Engag Pearson's ρ 0.1435 0.1287 0.1215 0.1312 0.1742 0.1824 0.1899 0.1509 0.2183 0.3402 0.3847 0.3649 0.4750 0.4863 0.4678 0.4970 0.5147 0.3593 0.5093 0.5089	$\begin{array}{r} \textbf{0.6524} \\ \hline \textbf{spearman's τ} \\ \hline \textbf{Spearman's τ} \\ \hline \textbf{0.1546} \\ \textbf{0.1429} \\ \textbf{0.1318} \\ \textbf{0.1220} \\ \textbf{0.1780} \\ \textbf{0.2020} \\ \textbf{0.2325} \\ \textbf{0.1675} \\ \textbf{0.2547} \\ \hline \textbf{0.3649} \\ \textbf{0.1698} \\ \textbf{0.3539} \\ \textbf{0.3609} \\ \hline \textbf{0.4665} \\ \textbf{0.4801} \\ \textbf{0.4652} \\ \textbf{0.4996} \\ \textbf{0.4772} \\ \textbf{0.5226} \\ \hline \textbf{0.3569} \\ \textbf{0.5074} \\ \textbf{0.5076} \\ \hline \end{array}$	$\begin{array}{c} \textbf{0.6748} \\ \hline \textbf{Grour Pearson's } \rho \\ \hline \textbf{0.1452} \\ 0.1304 \\ 0.1231 \\ 0.1333 \\ 0.1764 \\ 0.1847 \\ 0.1922 \\ 0.1527 \\ 0.2207 \\ \hline \textbf{0.3439} \\ 0.3358 \\ 0.3892 \\ 0.3688 \\ \hline \textbf{0.4800} \\ 0.4909 \\ 0.4727 \\ 0.5022 \\ 0.4532 \\ 0.5207 \\ \hline \textbf{0.3629} \\ 0.5143 \\ 0.5136 \\ \end{array}$	$\begin{array}{r} \textbf{0.6822} \\ \hline \textbf{dedness} \\ \hline \textbf{Spearman's } \tau \\ \hline \textbf{0.1561} \\ \textbf{0.1449} \\ \textbf{0.1339} \\ \textbf{0.1235} \\ \textbf{0.1796} \\ \textbf{0.2038} \\ \textbf{0.2344} \\ \textbf{0.1693} \\ \textbf{0.2575} \\ \hline \textbf{0.3689} \\ \textbf{0.1715} \\ \textbf{0.3576} \\ \textbf{0.3576} \\ \textbf{0.3646} \\ \hline \textbf{0.4710} \\ \textbf{0.4846} \\ \textbf{0.4694} \\ \textbf{0.5048} \\ \textbf{0.4826} \\ \textbf{0.5287} \\ \hline \hline \textbf{0.3605} \\ \textbf{0.5124} \\ \textbf{0.5127} \\ \end{array}$

Table 8: Breakdown of Pearson and Spearman correlations with human judgments by evaluation criteria on the TopicalChat dataset.



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.

Prompt for Dialogue Response 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 coherent is the text of the dialogue response? (on a scale of 1-5, with 1 being the lowest)

Input:

Conversation Context: Would you recommend some places for sightseeing? How about great canyon? Is it worth seeing?

Response: The movie was really good, it was worth watching it.

AMR Graph:

(multi-sentence :snt1 (recommend :ARG0 (you) :polarity (amr-unk) :ARG1 (place :quant (some) :location (sightsee))) :snt2 (canyon :mod (great) :polarity (amr-unk)) :ARG1 (worth) :ARG2 (see) :snt3 (and :mod (worth :ARG1 (watch) :ARG1 (movie) :mod (good :ARG1 (movie)))))

SLM score: 0.32

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

Table 9: Example of prompt template showing how SLM score and AMR graph information are integrated to evaluate dialogue response coherence.