CHIE: Generative MRC Evaluation for in-context QA with Correctness, Helpfulness, Irrelevancy, and Extraneousness Aspects

Wannaphong Phatthiyaphaibun [†] [*] **Surapon Nonesung**[†][*]**,Peerat Limkonchotiwat**[†]**,**
Can Udomcharoenchaikit[†]**, Jitkapat Sawatphol**[†]**,**
Ekapol Chuangsuwanich[§]**, Sarana Nutanong**[†]
[†]School of Information Science and Technology, VISTEC, Thailand
[§]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
wannaphong.p_s21@vistec.ac.th

Abstract

The evaluation of generative models in Machine Reading Comprehension (MRC) presents distinct difficulties, as traditional metrics like BLEU, ROUGE, METEOR, Exact Match, and F1 score often struggle to capture the nuanced and diverse responses. While embedding-based metrics such as BERTScore and BARTScore focus on semantic similarity, they still fail to fully address aspects such as recognizing additional helpful information and rewarding contextual faithfulness. Recent advances in large language model (LLM) based metrics offer more fine-grained evaluations, but challenges such as score clustering remain. This paper introduces a multi-aspect evaluation framework, CHIE, incorporating aspects of **C**orrectness, **H**elpfulness, **I**rrelevance, and **E**xtraneousness. Our approach, which uses binary categorical values rather than continuous rating scales, aligns well with human judgments, indicating its potential as a comprehensive and effective evaluation method.

1 Introduction

Evaluating generative models in machine reading comprehension (MRC) presents distinct challenges, as traditional n-gram-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Exact Match (EM), and F1 score often prove inadequate. These metrics are typically limited in their ability to assess the generalization capabilities of generative models, which are characterized by their production of diverse and nuanced responses.

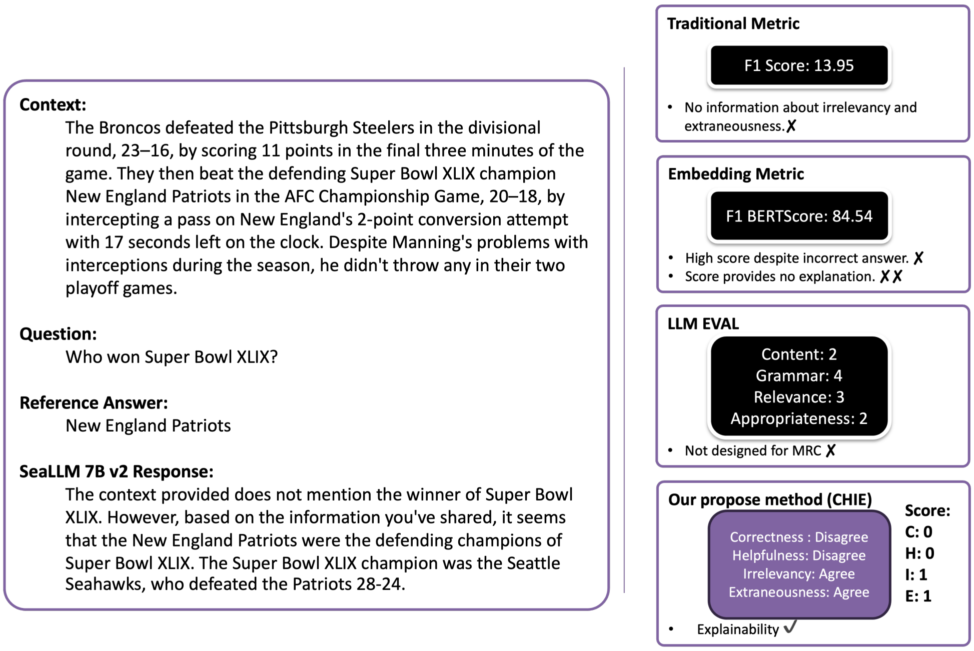
To address the n-gram matching problem, embedding-based metrics, i.e., BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), focus on semantic similarity assessments of the ground truth and prediction. Moreover, various evaluation criteria have been proposed to assess model outputs for generalized evaluation methods. For instance, USR (Mehri and Eskenazi, 2020) evaluates dialogue responses based on fluency, relevance, and knowledge conditioning using a RoBERTa-base model (Liu et al., 2019). Similarly, UniEval (Zhong et al., 2022) employs T5 (Raffel et al., 2020) to assess QA tasks from multiple perspectives, encoding texts as questions and answers and scoring them across various dimensions. However, these methods typically require datasets for fine-tuning, which can limit their applicability.

Recent advances in large language models (LLMs) based metrics, such as a GPT-based metric for translation (Kocmi and Federmann, 2023), summarization (Liu et al., 2023), and dialogue (Lin and Chen, 2023) tasks, offers more fine-grain evaluations using continuous rating scales. Despite these improvements, challenges like score clustering remain. This is because generative answers produced by LLMs require more generalized measurements than extractive ones. Moreover, these evaluation methods have mainly been used for high-resource languages like English. For low-resource languages, current research has not been thoroughly tested, creating a gap in understanding their performance in these languages.

In this paper, we propose a multi-aspect evaluation framework to assess the generalization of in-context learning called CHIE comprising four aspects: **C**orrectness, **H**elpfulness, **I**rrelevance, and **E**xtraneousness. Our work distinguishes itself from existing evaluation metrics, such as F1 and BERTScore, as illustrated in Figure 1. In particular, we introduce a multi-aspect evaluation scheme that delivers a more comprehensive and detailed analysis of a model's ability to present information. Unlike other LLM-based evaluations, such as LLM-EVAL, which provides a single numeric output (e.g., three on a 1-5 scale) that lacks explainability and can be challenging for human interpretation, our approach uses binary categorical values with objectively defined classes. Our method ensures

[*]Equal Contributions.

Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP, pages 154–164
November 16, 2024 ©2024 Association for Computational Linguistics



Context:

The Broncos defeated the Pittsburgh Steelers in the divisional round, 23–16, by scoring 11 points in the final three minutes of the game. They then beat the defending Super Bowl XLIX champion New England Patriots in the AFC Championship Game, 20–18, by intercepting a pass on New England's 2-point conversion attempt with 17 seconds left on the clock. Despite Manning's problems with interceptions during the season, he didn't throw any in their two playoff games.

Question:

Who won Super Bowl XLIX?

Reference Answer:

New England Patriots

SeaLLM 7B v2 Response:

The context provided does not mention the winner of Super Bowl XLIX. However, based on the information you've shared, it seems that the New England Patriots were the defending champions of Super Bowl XLIX. The Super Bowl XLIX champion was the Seattle Seahawks, who defeated the Patriots 28-24.

Figure 1: A comparison between our proposed CHIE framework and different evaluation metrics.

explainability and human-interpretable scoring and is specifically designed for the MRC task.

To demonstrate the generalization of our evaluation method, we evaluate Machine Reading Comprehension (MRC) capabilities in a multilingual environment. In particular, we evaluate six models on three languages compared to two evaluation metrics using XQuAD (Artetxe et al., 2020). Our findings reveal that commonly used metrics, such as F1, EM, and BERTScore, lack generalizability and do not accurately reflect the robustness of the evaluated models. In contrast, our experiments show that CHIE consistently aligns with human judgments, indicating its potential as a more reliable alternative for evaluating model responses. Furthermore, models evaluated using our proposed metric exhibit improved generalization compared to previous methods, suggesting that CHIE is more effective at capturing performance nuances across diverse scenarios. This is particularly significant in complex and ambiguous cases where traditional metrics fall short, underscoring the need for more sophisticated evaluation frameworks.

In summary, our main contributions are as follows:

- We introduce CHIE, a new automatic evaluation framework for machine reading comprehension with large language models, leveraging multi-dimensional aspects within a single prompt.
- We provide experimental evidence demonstrating that our designed binary categorical values align well with human evaluations.
- We show that CHIE can be applied to support MRC evaluations across different languages.

2 Related Work

2.1 N-gram-based Metrics

Metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), Exact Match (EM) and F1 score were primarily designed to rely on the n-gram overlap between model outputs and reference answers. These metrics often fall short when applied to generative models, which produce diverse and contextually nuanced responses. They can overlook subtleties in language use, creativity, and the overall coherence of the generated text. Thus, there is a pressing need to develop and adopt more sophisticated evaluation metrics to capture the multifaceted nature of generative model outputs, ensuring that these models are assessed more accurately and comprehensively.

2.2 Embedding-based Metrics

To enhance the semantic similarity between generated and reference texts, embedding-based metrics utilizing word embeddings have been proposed. BERTScore (Zhang et al., 2020) computes the semantic similarity between the reference and the target text using a pre-trained BERT model, while BARTScore (Yuan et al., 2021) evaluates generated text as a text generation task via a pre-trained BART model. However, embedding-based and n-gram-based methods exhibit inherent limitations due to their reliance on reference texts, restricting their applicability in tasks where a reference is unavailable. Additionally, they may fail to adequately capture important aspects of overall quality, such as fluency, faithfulness, coherence, and adherence to specific instructions.

2.3 Multi-aspect Evaluation

Multiple aspects have been proposed to evaluate various model output dimensions. For instance, summarization tasks require consistency or naturalness assessment, while dialogue tasks must assess fluency and coherence. USR (Mehri and Eskenazi, 2020) proposes evaluating dialogue response generation across multiple aspects, such as fluency, relevance, and knowledge conditioning, using a RoBERTa-base model (Liu et al., 2019). UniEval (Zhong et al., 2022) suggests training a model to evaluate QA tasks from different perspectives using T5. This is achieved by encoding both source and target texts as questions and answers and then computing a score. It can manage different aspects of evaluation by modifying the question format. Unlike these approaches, CHIE employs LLMs as the base model with a single prompt, providing interoperability and eliminating the need for model fine-tuning.

2.4 LLM-based Metrics

As LLMs become increasingly sophisticated, recent studies have developed LLM-based metric approaches for assessing natural language generation (NLG) outputs. Researchers have recognized the limitations of traditional metrics and proposed several novel methods to better evaluate generative models. GPTScore (Fu et al., 2023) outlines a general framework to evaluate different aspects of generated outputs based on posterior probability. However, they are not focused on in-context QA applications. Their score albeit showing high correlation with human judgement, is not easily interpretable, just like how perplexity is harder to understand compared to accuracy. Kocmi and Federmann (2023) propose a GPT-based metric for assessing translation quality. They utilized a continuous rating scale ranging from 0 to 100 or a 1 to 5-star ranking and found that their approach achieves state-of-the-art accuracy, outperforming traditional automatic metrics. However, the comprehensive assessment by task-specific aspects remains insufficiently explored. In a similar vein, Liu et al. (2023) propose G-EVAL, a framework using Large Language Models (LLMs) with chain-of-thought (CoT) reasoning and a form-filling paradigm, feeding task-specific views as prompts in evaluation criteria. However, their study observed that LLMs typically produced integer scores even when explicitly prompted to provide decimal values. This tendency resulted in numerous ties in the evaluation scores. Subsequently, Lin and Chen (2023) introduced LLM-Eval, a comprehensive multi-dimensional automatic evaluation for open-domain conversations with LLMs. This method employs a single prompt alongside a unified evaluation schema encompassing various dimensions of evaluation with a continuous rating scale.

3 Proposed Method

In this section, we discuss an LLM-based evaluator covering multiple aspects of MRC called CHIE. We first describe the desired Features in Section 3.1. Second, we provide evaluation criteria for MRC evaluations in Section 3.2. Last but not least, Section 3.3 explains the components of CHIE-based prompting.

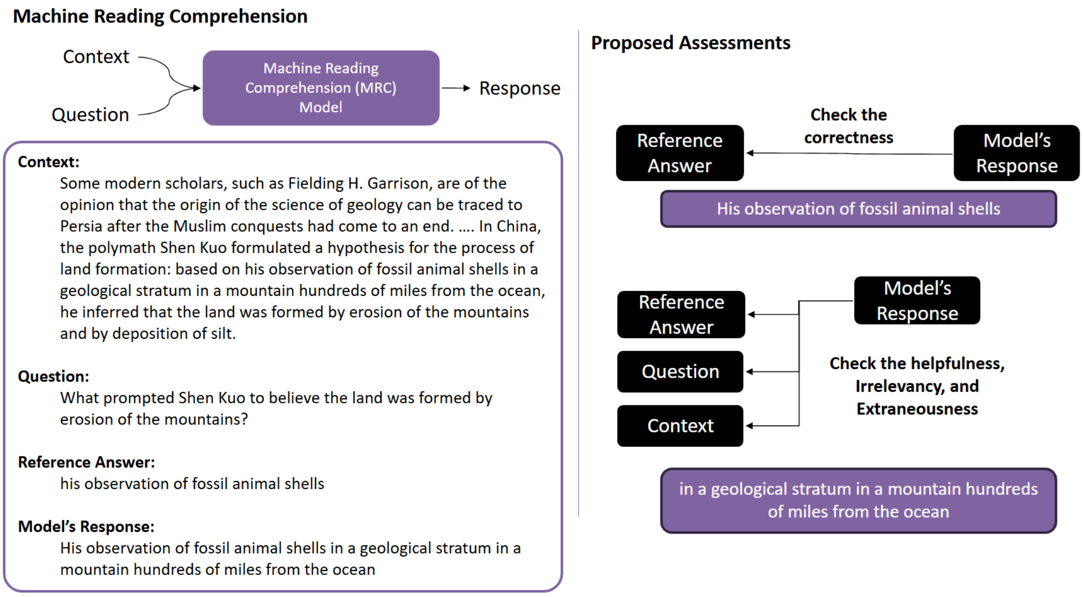
3.1 Desired Features

As shown in Figure 2, we propose an evaluation that goes beyond rewarding the answer to correctness by assessing additional information accompanying the answer as follows.

Reward relevant and helpful information. The method should recognize and reward responses that are *not only* accurate *but also* provide comprehensive and relevant information. This encourages models to generate answers that are both correct and rich in content.

Penalize unconnected information. While helpful additional information is welcome, we want to keep the response concise. The method should penalize additional information that *does not* improve the understanding of the question or answer. This criterion also discourages the model from *"cheating"* by excessively including phrases from the context to increase the chance of obtaining a reward from the previous criterion.

Penalize out-of-context information. The method should penalize the inclusion of out-of-context information, even if factually correct. This criterion aligns with the spirit of reading compre-


Machine Reading Comprehension
Context
Question
Machine Reading Comprehension (MRC) Model
Response
Proposed Assessments
Reference Answer
Check the correctness
Model's Response
His observation of fossil animal shells
Reference Answer
Model's Response
Question
Check the helpfulness, Irrelevancy, and Extraneousness
Context
in a geological stratum in a mountain hundreds of miles from the ocean

Context:
Some modern scholars, such as Fielding H. Garrison, are of the opinion that the origin of the science of geology can be traced to Persia after the Muslim conquests had come to an end. .... In China, the polymath Shen Kuo formulated a hypothesis for the process of land formation: based on his observation of fossil animal shells in a geological stratum in a mountain hundreds of miles from the ocean, he inferred that the land was formed by erosion of the mountains and by deposition of silt.

Question:
What prompted Shen Kuo to believe the land was formed by erosion of the mountains?

Reference Answer:
his observation of fossil animal shells

Model's Response:
His observation of fossil animal shells in a geological stratum in a mountain hundreds of miles from the ocean

Figure 2: An illustration of our proposed CHIE framework: multi-aspects evaluation using a single prompt

hension assessment since we want to evaluate the model's capability to grasp and interpret the question. Furthermore, by encouraging the response to stick to the context, we also mitigate the risks of hallucination.

3.2 Designed Evaluation Criteria For MRC

Our proposed method follows the traditional MRC evaluation where each assessment consists of four components: context, question, reference answer, and response, as shown in Figure 2.

- Correctness: Assess whether the model's response is accurate wrt. the reference answer (↑ higher is better).
- Helpfulness: Determine whether the model's response provides additional relevant details from the context (↑ higher is better).
- Irrelevancy: Check whether the model's response contains irrelevant details from the context (↓ lower is better).
- Extraneousness: Verify whether the model's response includes out-of-context information (↓ lower is better).

3.3 CHIE-based Prompting

Our proposed method, CHIE, is a prompt-based evaluator consisting of three main components:

- **Task Instruction:** This component guides a LLM to do the required task.
- **Evaluation Criteria:** This component uses agree-disagree questions to evaluate four specific aspects of the model.
- **Form-input Structure:** This component provides a template for filling in the necessary information for evaluation.

We concatenate the three components into a single prompt (full prompt shown in Appendix A.2). CHIE can efficiently generate multi-dimensional binary classifications for the responses without requiring multiple prompts. Ratings are then post-processed by assigning "Agree" as 1 and "Disagree" as 0. The large language model is invoked only once, directly providing evaluation scores for each dimension according to the defined schema.

4 Experimental Settings

Data. We focus on Thai, English, and Chinese by leveraging the XQuAD dataset (Artetxe et al., 2020). To ensure feasibility within resource constraints, including limited GPT-4 API access and budget for human evaluators, we employ a subset of the first 100 rows from the Thai XQuAD dataset.
Models. We evaluate openly released LLMs with multilingual capabilities:

- **OpenThaiGPT-7B** (OpenThaiGPT, 2023): A Llama2 and continues pretraining on a Thai corpus with the application of supervised fine-tuning (SFT).
- **SeaLLM-7B V2** (Nguyen et al., 2023): A Mistral-based model that continues pretraining on a Southeast Asia corpus, utilizing both SFT and Direct Preference Optimization (DPO) (Rafailov et al., 2023).
- **WangchanLion-7B** (Phatthiyaphaibun et al.,

2024): A MPT-based model that Sealion continues pretraining on a Southeast Asia corpus and employs SFT.

- **Llama-3-8B Instruct** (Dubey et al., 2024): An instruction model of Llama 3 from Meta that utilizes both SFT and DPO.
- **Llama-3.1-8B Instruct** (Dubey et al., 2024): An instruction model of Llama 3.1 that improved the performance from Llama 3 by expanded multilingual support, an increased context window, enhanced synthetic data generation capabilities, and specialized fine-tuning for tool utilization.
- **Llama-3-8B SEA-LION instruct** (Singapore, 2024): An Llama 3.1 based model that continued pre-training on the Llama 3 architecture, specifically focused on Southeast Asian languages. This model has been fine-tuned with approximately 100,000 English instruction-completion pairs, along with a smaller set of around 50,000 pairs from various ASEAN languages, including Indonesian, Thai, and Vietnamese.

English Prompts vs Native Prompts. We also compared the evaluation performance of English vs Native (i.e., Thai) prompts detailed in Appendix A.3. The results suggest that English prompts yield superior performance. This result conforms with the literature (Lai et al., 2023).

Human Response Collection. The human response annotation phase consists of three steps: training, screening, and deployment. In the training step, candidates were given 15 sample responses with expected assessments to familiarize themselves with the task. Seven candidates participated in this step. In the screening step, candidates were given 10 sample responses that they needed to answer. The training and screening samples were obtained from questions 1 to 100 from the Thai subset in the XQuAD dataset. In the deployment step, we selected candidates who scored more than 80% as our annotators. We obtained five annotators as a result. These five annotators were assigned to assess responses from three models, OpenThaiGPT, SeaLLMs, and WangchanLion, answering 100 Questions in the XQuAD Dataset.

LLM candidates. We select robust and generalized LLMs to be the judge model: GPT-4[1], GPT-4o[2], GPT-3.5 Turbo[3], and Gemini Pro 1.0[4].

5 Experimental Results

In this section, we report experimental results from three studies. Section 5.1 compares our multi-aspect approach, CHIE, with two single-aspect metrics, F1 and BERTScore. Section 5.1 explores the possibility of automating multi-aspect evaluations using an LLM. Section 5.3 provides a component-wise analysis of CHIE through A/B preference evaluation using humans and an LLM.

5.1 Single-Aspect vs Multi-Aspect Evaluations

Table 1 displays a comparison between the two single aspect measures, F1 and BERTScore (BRTSc), and the multi-aspect assessments, CHIE. We can see that the BERTScore and F1 agree with each other in the sense that WangchanLion has the highest F1 and BERTScore, while SeaLLM V2 has the lowest F1 and BERTScore. For the multi-aspect part, we employed five human evaluators and computed the majority vote as the assessment result. Interestingly, the multi-aspect results show a disagreement with BERTScore and F1 in terms of correctness (C). SeaLLM V2 has the highest C score, suggesting the superior capability to produce correct responses with respect to the reference answers. Furthermore, SeaLLM V2 also exhibits the highest helpfulness (H) score, suggesting the capability to add useful information to the main answer while staying within the context.

	Single-Aspect		Multi-Aspect			
Model	F1	BRTSc	C↑	H↑	I↓	E↓
OpenThaiGPT	34.96	75.95	60	38	30	32
SeaLLM V2	14.00	63.10	**80**	**80**	**20**	45
WangchanLion	**50.12**	**81.27**	67	19	23	**5**

Table 1: Comparison between single-aspect evaluation techniques, F1 and BERTScore (BRTSc), and our CHIE multi-aspect summation measurements on three different LLMs.

BRTSc Range	C↑	H↑	I↓	E↓	Avg. Len.
Low	58	**58**	**36**	**45**	30.58
Medium	64	51	27	26	16.87
High	**85**	28	10	11	5.81

Table 2: BERTScore vs. CHIE summation measurements vs. average answer length for different ranges of BERTScore.

Table 2 provides a further analysis of the relation between BERTScore (BERTSc) and each of the CHIE aspects. The table shows three BERTScore (BRTSc) ranges: the lowest, middle, and highest BERTScore terciles of the responses

[1] gpt-4-0613
[2] gpt-4o-2024-05-13
[3] gpt-3.5-turbo-0125
[4] gemini-1.0-pro-002

Assessor	Correctness (C) ↑			Helpfulness (H) ↑			Irrelevancy (I) ↓			Extraneousness (E) ↓			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gemini	97.35	88.89	92.93	85.11	29.20	43.48	**65.38**	23.29	34.34	69.23	32.93	44.63	**89.04**	52.73	67.00
GPT-3.5	91.67	**95.65**	93.62	72.26	**81.75**	**76.71**	63.64	28.77	39.62	44.87	42.68	43.75	75.93	**73.35**	74.62
GPT-4	98.99	94.69	**96.79**	**94.20**	47.45	63.11	51.14	**61.64**	**55.90**	**77.61**	**63.41**	**69.80**	84.83	71.74	**77.74**
GPT-4o	**100.00**	77.29	87.19	94.74	52.55	67.61	29.41	20.55	24.19	74.36	35.37	47.93	84.66	55.31	66.91

Table 3: LLMs-automated evaluation compared to human evaluation. P, R , and F1 denote as precision, recall, and F1 score computed by comparing the evaluation outputs of each LLM compared to human majority responses.

Assessor	Correctness (C) ↑			Helpfulness (H) ↑			Irrelevancy (I) ↓			Extraneousness (E) ↓			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Human 4	97.06	96.12	96.59	**97.98**	46.86	63.40	71.11	37.21	48.85	**89.83**	61.63	**73.10**	**93.37**	64.96	76.61
Human 5	88.94	**97.57**	93.06	87.50	**54.11**	**66.87**	**88.46**	53.49	**66.67**	65.48	**63.95**	64.71	84.49	**70.77**	**77.02**
GPT-4	**98.99**	95.15	**97.03**	94.20	31.40	47.10	62.50	**63.95**	47.10	82.09	**63.95**	71.90	87.91	63.42	71.90

Table 4: Agreement between Human 4, Human 5, and GPT-4 answers using F1 score.

to 100 XQuAD questions from OpenThaiGPT, SeaLLM V2, and WangchanLion, bringing the total of responses to 300. Therefore, each tercile contains exactly 100 responses. We can see that the high BERTScore range is associated with a higher correctness (C) score. This is because, like BERTScore, the correctness aspect (C) assesses whether the model's response conveys the same meaning as the reference answer. We can also see that low BERTScores are associated with higher H, I, and E counts since agreement to these questions involves the inclusion of additional information beyond the reference answer.

These results show that while a high BERTScore indicates semantic faithfulness to the reference answer, a low BERTScore can mean many different things: an incorrect answer, an inclusion of helpful information, a verbose response, or an out-of-context response. In other words, a response can be both correct and helpful but obtain a low BERTScore due to the semantic discrepancy with respect to the reference answer. Furthermore, since we use the XQuAD reference answers for BERTScore similarity determinations, higher BERTScores tend to have shorter answers. In applications demanding contextually rich responses, BERTScore may not be indicative of desired responses. These results highlight the merit of our multi-aspect assessment approach in comparison to single-aspect measures like BERTScore or F1.

5.2 LLMs as Multi-Aspect Evaluators

Let us now explore the possibility of automating the CHIE evaluation using an LLM. We identified four state-of-the-art LLM candidates: Gemini Pro 1.0 (Team, 2024), GPT-3.5 Turbo, GPT-4, and GPT-4o. For consistency, we use the same prompt for all LLMs. Details are given in Appendix A.2.

As shown in Figure 3, this study contains two analyses: LLM-to-LLM and LLM-to-Human comparisons.

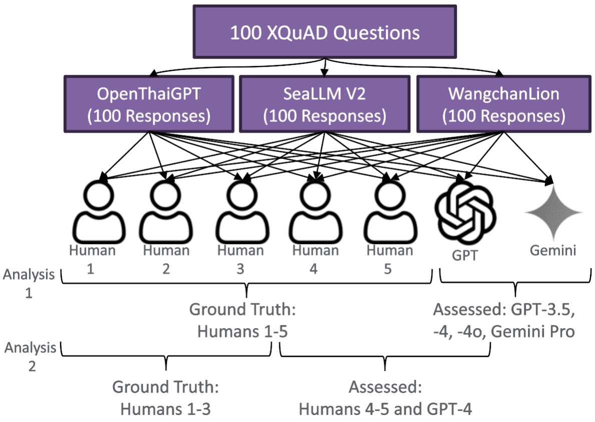

100 XQuAD Questions
OpenThaiGPT (100 Responses)
SeaLLM V2 (100 Responses)
WangchanLion (100 Responses)
Human 1
Human 2
Human 3
Human 4
Human 5
GPT
Gemini
Analysis 1
Ground Truth: Humans 1-5
Assessed: GPT-3.5, -4, -4o, Gemini Pro
Analysis 2
Ground Truth: Humans 1-3
Assessed: Humans 4-5 and GPT-4

Figure 3: Overview of our analyses comparing LLMs and human assessors.

Analysis 1: LLM-to-LLM Comparisons. For ground truths, we use the same voting results from the five human evaluators as explained in Section 5.1. We then compared the assessments from four LLMs, Gemini, GPT-3.5, GPT-4, and GPT-4o. Table 3 shows that GPT-4 outperforms other models in terms of Correctness, Irrelevancy, and Extraneousness. In the aspect of Helpfulness, GPT-3.5 demonstrates superior performance. Overall, GPT-4 provides the highest F1 score among the evaluated models. Consequently, we selected GPT-4 as the LLM evaluator for the rest of the presentation.

Subset	Model	Single-Aspect		Multi-Aspect				Tokens (avg)
		F1	BRTSc	C ↑	H ↑	I ↓	E ↓	
Thai	OpenThaiGPT-7B	34.96	75.95	58	13	28	31	10.35
	SeaLLM-7B V2	14.08	63.10	76	48	32	31	27.81
	WangchanLion-7B	50.12	**81.27**	64	8	28	5	5.50
	Llama-3-8B Instruct	13.03	61.69	88	**68**	9	8	27.76
	Llama-3.1-8B Instruct	41.21	73.02	85	19	12	8	12.67
	Llama-3-8B SEA-LION instruct	**51.22**	78.07	**93**	34	**5**	**0**	12.53
English	OpenThaiGPT-7B	18.12	76.61	42	8	54	52	24.59
	SeaLLM-7B V2	22.86	84.03	**96**	33	6	12	19.98
	WangchanLion-7B	26.40	85.09	68	20	30	22	13.64
	Llama-3-8B Instruct	19.80	83.68	94	**59**	**4**	**6**	21.13
	Llama-3.1-8B Instruct	24.18	84.26	88	41	12	14	18.78
	Llama-3-8B SEA-LION instruct	**42.14**	**87.58**	94	34	5	12	11.58
Chinese	OpenThaiGPT-7B	5.63	54.28	26	12	61	62	147.75
	SeaLLM-7B V2	19.04	58.27	88	39	16	12	24.86
	WangchanLion-7B	**44.55**	**73.28**	52	4	27	21	18.53
	Llama-3-8B Instruct	12.12	53.52	86	**66**	9	7	47.91
	Llama-3.1-8B Instruct	42.00	68.76	**91**	17	**3**	**2**	10.66
	Llama-3-8B SEA-LION instruct	30.95	63.74	88	28	13	8	14.69

Table 5: The result of CHIE evaluation across three different languages (Thai, English, and Chinese) and six LLMs (OpenThaiGPT-7B, SeaLLM-7B V2, WangchanLion-7B , Llama-3-8B Instruct, Llama-3.1-8B Instruct and Llama-3-8B SEA-LION instruct).

Analysis 2: LLM-to-Human Comparisons. We used three human evaluators to compute the ground truths, as shown in Figure 3. The other two evaluators were left out for performance comparison with GPT-4. Table 4 shows that GPT-4's evaluations align closely with human evaluators, achieving an overall F1 score of 71.90. This differs by only 4.71 points from the fourth human evaluator and by 5.12 points from the fifth human evaluator. Thus, given the time and cost of human evaluation, GPT-4 is a viable alternative for assessing the MRC task.

5.3 Human vs LLM Preferences

Due to the extractive nature of the MRC task, we aimed to verify whether humans prefer longer or shorter responses. To investigate this, we conducted a head-to-head comparison by manually creating new XQuAD answers that encapsulate various aspects of our criteria:

- **C vs CH:** Whether humans or GPT-4 prefer answers that contain only the Correctness aspect (C) or those that encompass both Correctness and Helpfulness aspects (CH).
- **C vs CI:** Whether humans or GPT-4 prefer answers that contain only the Correctness aspect (C) or those that include both Correctness and Irrelevancy aspects (CI).
- **CH vs CHI:** Whether humans or GPT-4 prefer answers with Correctness and Helpfulness (CH) or those with Correctness, Helpfulness, and Irrelevancy (CHI).

We instructed five human evaluators to identify their preferred answers in Thai as detailed in Appendix A.1. For comparison, we also used GPT-4 for evaluation following the instructions outlined in Appendix A.1. From Table 6, we found that humans exhibited a strong preference for shorter answers, i.e., preferring C to CH and CI and CH to CHI. For GPT-4, on the other hand, CH was preferred to C. We can also see that although GPT-4 preferred C to CI and CH to CHI, like humans, the score differentials are not as strong. This result conforms with the observation presented by Zheng et al. (2023) that LLMs such as Claude-v1 and GPT-4 tend to prefer longer responses.

Case	Humans			GPT-4		
	A	B	Tie	A	B	Tie
C vs CH	**91**	4	5	15	**83**	2
C vs CI	**99**	1	0	**60**	40	0
CH vs CHI	**98**	1	1	**67**	27	6

Table 6: A/B preference evaluation conducted by humans and GPT-4 as evaluators.

5.4 CHIE on generalizability across languages

After identifying GPT-4 as the most effective evaluation model, we expanded our study to include additional languages from the XQuAD dataset to assess behavior generalization across languages. We added English and Chinese, ensuring that the questions matched the same question IDs. Table 5 presents the results for 100 questions from the XQuAD dataset in Thai, English, and Chinese, evaluated across six diverse models. The experiments reveal the following:

- **F1 and BERTScore (BRTSc) with Correctness (C) and Helpfulness (H):** Higher F1 and BERTScore values are positively correlated with higher Correctness (C) and Helpfulness (H) scores. This means that models with better overall performance, as indicated by F1 and BERTScore, are more likely to generate responses that are accurate and useful.
- **Token length can have both positive and negative effects:** Longer token lengths generally correlate with higher Correctness (C) and Helpfulness (H). This suggests that longer responses tend to be more thorough and accurate. However, as token length increases, there is a risk of higher Irrelevancy (I) and Extraneousness (E). This indicates that overly lengthy responses are more likely to include irrelevant or unnecessary content.
- **Irrelevancy (I) and Extraneousness (E) with F1 and BERTScore:** Lower F1 and BERTScore values are associated with higher Irrelevancy (I) and Extraneousness (E) scores. This means that models with poorer performance tend to produce more irrelevant or extraneous information.

6 Conclusion

We present CHIE, a novel automatic evaluation framework using GPT-4 for assessing MRC model responses. In comparison to single-aspect measures such as BERTScore, CHIE provides a more holistic means of assessing MRC responses by assessing the helpfulness of the answer and screening for irrelevancy and out-of-context information in addition to correctness.

We also explore the possibility of using LLMs as evaluators. The results demonstrate potential for further development for using an LLM in a completely automated evaluation process or as an evaluator to reduce the human evaluation workload.

Limitations

- Although CHIE improves the comprehensiveness in assessing MRC responses, its usefulness heavily relies on the nature of underlying benchmark questions. While XQuAD is an excellent resource for assessing MRC capabilities, due to its extractive nature, its questions do not test the commonsense reasoning capability or the ability to integrate world knowledge into the answer. For future work, we plan to apply CHIE to other benchmarks for richer assessments.
- In terms of preference, results from human evaluation contradict those from GPT-4. As a result, it is still inconclusive whether the helpfulness aspect should be considered as a desired feature or not. One possible explanation lies in the extractive nature of XQuAD questions that can be answered with a short text sequence. Consequently, the inclusion of additional information may not always improve the desirability of responses. For future work, we plan to compose our own benchmark for CHIE.
- CHIE uses an LLM for evaluation, which may introduce bias into the framework and result in a loss of interpretability.

Ethical Statement

The human annotators who participated in this study were fairly compensated according to the applicable labor laws.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models.

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

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*.

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

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia.

OpenThaiGPT. 2023. Openthaigpt 1.0.0-beta. `https://huggingface.co/openthaigpt/openthaigpt-1.0.0-beta-7b-chat-ckpt-hf`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Chompakorn Chaksangchaichot, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Wangchanlion and wangchanx mrc eval.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

A Appendix

A.1 Instruction for Preferred Answers

Thai:

“คำตอบไหนดีกว่ากัน คุณทำหน้าที่ตัดสินและประเมินคุณภาพการตอบ AI ผู้ช่วยสองโมเดลกับผู้ใช้งาน คุณควรเลือกผู้ช่วยที่ปฏิบัติตามคำแนะนำของผู้ใช้และตอบคำถามของผู้ใช้ได้ดีกว่า และเลือกคำตอบโดยพิจารณาปัจจัยต่าง ๆ เช่น ความถูกต้อง ความกระชับ ความเกี่ยวข้อง และการให้ข้อมูลที่มีประโยชน์”

English translation:
“Which answer is better? You shall act as a judge and evaluate the quality of the responses to the user question provided by two AI assistants. You should choose the assistant that follows the user’s instructions and answers the user’s question better. Your evaluation should consider factors such as correctness, conciseness, relevancy, and helpfulness.”

A.2 Evaluation Prompt

> *Please evaluate these answers based on their accuracy and relevance to the provided passage based on the Criteria:*
>
> *Q1. The Answer is Correct concerning the Reference Answer. Do you agree or disagree? Determine if the given answer accurately matches the reference answer provided. The correctness here means the answer must directly correspond to the reference answer, ensuring factual accuracy.*
>
> *Q2. The Answer Includes Relevant, Additional Information from the Context. Do you agree or disagree? Determine if the given answer accurately Assess whether the answer provides extra details that are not only correct but also relevant and enhance the understanding of the topic as per the information given in the context.*
>
> *Q3. The Answer Includes Additional, Irrelevant Information from the Context. Do you agree or disagree? Check if the answer contains extra details that, while related to the context, do not directly pertain to the question asked. This information is not necessary for answering the question and is considered a digression.*
>
> *Q4. The Answer Includes Information Not Found in the Context. Do you agree or disagree? Evaluate if the answer includes any correct information that is not included in the context. This information, even if correct, is extraneous as it goes beyond the provided text and may indicate conjecture or assumption.*
>
> *Passage:* $\{C\}$
>
> *Question:* $\{Q\}$
>
> *Reference Answer:* $\{R\}$
>
> *Prediction Answer:* $\{O\}$

A.3 Thai prompt vs English prompt

Table 7 shows the English prompt is better than the Thai prompt in GPT-4.

Prompt	Correctness (C)			Helpfulness (H)			Irrelevancy (I)			Extraneousness (E)			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
English	98.99	94.69	96.79	94.20	47.45	63.11	51.14	61.64	55.90	77.61	63.41	69.80	84.83	71.74	77.74
Thai	99.49	94.69	97.03	94.44	37.23	53.40	65.08	56.16	60.29	67.27	45.12	54.01	88.08	65.13	74.88

Table 7: Agreement between English prompt and Thai prompt.