*imap*Score: Medical Fact Evaluation Made Easy

Huimin Wang*, **Yutian Zhao***, **Xian Wu**†, **Yefeng Zheng**
Jarvis Lab, Tencent, Shenzhen, China
{hmmmwang,yutianzhao}@tencent.com

Abstract

Automatic evaluation of natural language generation (NLG) tasks has gained extensive research interests, since it can rapidly assess the performance of large language models (LLMs). However, automatic NLG evaluation struggles with medical QA because it fails to focus on the crucial correctness of medical facts throughout the generated text. To address this, this paper introduces a new data structure, *imap*, designed to capture key information in questions and answers, enabling evaluators to focus on essential details. The *imap* comprises three components: Query, Constraint, and Inform, each of which is in the form of term-value pairs to represent medical facts in a structural manner. We then introduce *imap*Score, which compares the corresponding medical term-value pairs in the *imap* to score generated texts. We utilize GPT-4 to extract *imap* from questions, human-annotated answers, and generated responses. To mitigate the diversity in medical terminology for fair term-value pairs comparison, we use a medical knowledge graph to assist GPT-4 in determining matches. To compare *imap*Score with existing NLG metrics, we establish a new benchmark dataset. The experimental results show that *imap*Score consistently outperforms state-of-the-art metrics, demonstrating an average improvement of 79.8% in correlation with human scores. Furthermore, incorporating *imap* into n-gram, embedding, and LLM metrics boosts the base versions, increasing correlation with human scores by averages of 89.9%, 81.7%, and 32.6%, respectively.

1 Introduction

LLMs, assimilated extensive knowledge, have demonstrated impressive potential as medical consultants. They can act as professional experts to aid doctors in diagnosis, treatment decision-making, and offering second opinions on complex cases

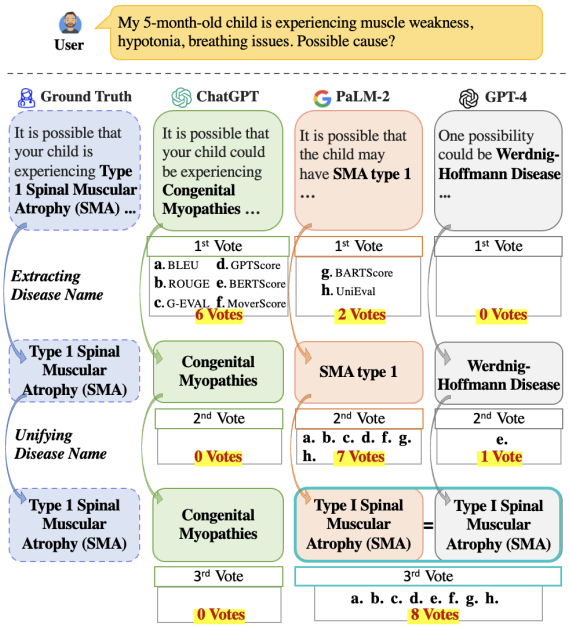


Figure 1: The illustration highlights challenges and potential remedies in automatic medical QA evaluation, comparing responses from ChatGPT, PaLM-2, and GPT-4 to Ground Truth through voting based on metric scores. Initially, ChatGPT erroneously prevails based on votes from the raw response text. In the 2^{nd} round, PaLM-2 takes the lead by focusing on the disease name. However, most of the metrics fail to recognize SMA's alternate name, Werdnig-Hoffmann Disease. GPT-4 correctly identifies this alternate name and should receive equal recognition and votes as PaLM-2. Upon unifying the disease name in the 3^{rd} round, all metrics accurately identified and correctly voted for the disease.

(Thirunavukarasu et al., 2023; Singhal et al., 2022). They also assist patients by providing diagnostic information, treatment options, and medical popular science (Huo et al., 2023). However, the practical application of LLMs in clinical settings, where medical factual correctness is critical, encounters challenges. One significant issue is the tendency of LLMs to generate "hallucinated" content—information that is plausible but incorrect or unfounded in evidence (Maynez et al., 2020; Kad-

*Equal Contribution
†Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 10242–10257

dour et al., 2023). Meanwhile, many LLMs are lauded for their proficiency in medical question-answering (QA), as evidenced by their high scores on standardized single-choice or multiple-choice medical examinations (Tu et al., 2023). However, excelling in such exams does not necessarily equate to being a competent medical professional.

Recently, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have become popular in medical response evaluation (Cai et al., 2023). However, they fall short in effectively assessing content quality or capturing medical language nuances and factual accuracy (Reiter and Belz, 2009). LLM-based metrics (Liu et al., 2023b; Fu et al., 2023) also face challenges due to the diversity and complexity of medical terminology, impacting their ability to verify medical facts accurately. Furthermore, these models' uniform treatment of all tokens in the responses neglects the crucial aspect of medical correctness, affecting the reliability of LLM-based evaluations in healthcare settings.

As shown in Figure 1, an example of automatic evaluation for disease diagnosis QA generated by three LLMs: ChatGPT, PaLM-2 and GPT-4. In this scenario, ChatGPT mistakenly receives the majority of support from six metrics (votes 6) in the first voting on raw generated text, indicating that most metrics are incapable of accurately evaluating disease diagnoses in their entirety. When the focus shifted to disease names in the second round, the majority of metric votes leaned towards the correct answer from PaLM-2. This suggests key information extraction could be a viable approach, mirroring the practice of highlighting and aligning essential details in the generated answer and the label during manual assessment. However, the results still neglect the fact that SMA (the disease mentioned in the example) is also known as Werdnig-Hoffmann Disease, a detail overlooked by most metrics. If we further unify the disease name and conduct the third round of vote, all metrics unanimously select the correct answer (generated by PaLM-2 and GPT-4), underscoring the importance of terminology standardization in evaluation.

These considerations guide a more refined evaluation paradigm for medical QA, involving: 1) identifying core needs and critical evidence, 2) unifying terminology and ensuring alignment, and 3) prioritizing key information and scoring. To extract the key information, we introduce a new data structure, ***imap*, which captures the core requirements and informs from questions and answers,** parsing them into medical term-value pairs based on primary needs, constraints, and informed points. GPT-4 can be readily instructed to generate the *imap* for questions and answers. Additionally, a medical knowledge graph aids in unifying medical expressions. Leveraging the *imap* of the question, reference, and generated response, we proposed a new metric, ***imap*Score, which compares corresponding medical term-value pairs within the *imap* to score generated texts.** This metric can be flexibly applied to prompt LLMs to evaluate text from various perspectives, integrating principles from human scoring to align more closely with expert assessments. We quantitatively evaluate *imap*Score by comparing the correlation between *imap*Score and human score for medical QA. Since there are no public available benchmark for medical QA with human ratings, we created a new benchmark that can also be utilized for other medical QA-related research. We enlisted three medical professionals to rate the generated outputs of two public datasets, HMedQA in Chinese and iCliniq[1] in English from three perspectives: factual accuracy, completeness, and specific. [2] After performing extensive experiments and conducting in-depth analysis from diverse perspectives, we found that *imap*Score surpasses other metrics, significantly improving the correlation with human scores. Furthermore, the integration of *imap* into n-gram, embedding, and LLM metrics notably enhances the base versions. Our contributions are as follows:

- **New Benchmark**: A new benchmark [3] for evaluating the NLG performance in medical QA, especially from the perspective of factual correctness of the LLM-generated content.
- ***imap***: A data structure introduced to parse medical QA into term-value pairs, capturing the core needs and critical evidence from questions and answers to enhance evaluation. Measuring *imap* of the generated content can enhance the performance of n-gram, embedding, and LLM-based metrics.
- ***imap*Score**: A new metric proposed to use the

[1] https://www.icliniq.com/

[2] HMedQA, the latest medical QA dataset in Chinese, is accessible through this link: HMedQA Derived from "Huatuo-Llama-Med-Chinese," we abbreviated it as HMedQA in our paper. The iCliniq dataset, sourced from the iCliniq website, can be found: iCliniq .

[3] https://github.com/HathyHuimin/imapScore

imap for scoring texts, aiming to align closely with expert assessments. Experimental results show that *imap*Score significantly improves correlation with human scores over existing metrics.

2 Related Work

Metrics for Evaluating Generated Response. Various metrics have been proposed to measure the semantic equivalence of the generated texts against the reference. N-gram-based metrics are most commonly used; they rate the quality of generated text by computing scores derived from the lexical overlap between the predictions and the reference. Representatives include BLEU which emphasizes n-gram precision, and ROUGE which focuses on n-gram recall. NIST (Doddington, 2002), a variant of BLEU, emphasizes the importance of informative n-grams, offering a more nuanced assessment. Embedding-based metrics offer better semantic understanding compared to n-gram-based methods that depend on superficial text overlap. BERTScore (Zhang et al., 2019), a prominent embedding-based metric, employs BERT to derive features from sentences, producing similarity scores through word pair inner products. Another noteworthy metric is BARTScore (Yuan et al., 2021), which formulates evaluating generated text as a text generation task from BART (Lewis et al., 2019). MoverScore (Zhao et al., 2019) advances by evaluating semantic similarity, tracking "information units" from reference to generated text. In the realm of fact-based summarization evaluation, QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2021) employ question generation models and question answering models to facilitate a detailed comparison of facts, while AlignScore (Zha et al., 2023) evaluates text alignment through chunk comparison. However, these general task methods may require specific dataset training and lack specificity evaluation. Recently, LLM-based approaches such as GPTScore (Fu et al., 2023) and G-EVAL (Liu et al., 2023b) have been employed for evaluating generated responses. Leveraging LLMs as evaluators presents significant promise in appraising responses across a range of tasks. However, these methods often struggle to measure fine-grained medical facts and are heavily dependent on predefined evaluation protocols and prompts. FActScore (Min et al., 2023) proposes an approach to evaluate factual accuracy by dividing generated text into short sentences. In contrast, our method delves into term-level relationships and explores additional dimensions, including comprehensiveness and specificity, providing a more nuanced and detailed evaluation.

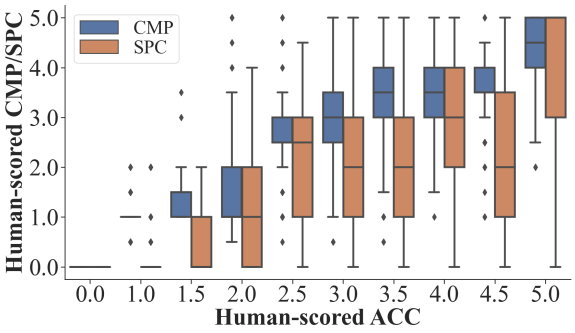


Figure 2: The boxplot displays the distribution of human-scored CMP/SPC at the same ACC in our benchmark. It reveals a substantial variance in CMP/SPC metrics at the same ACC (> 1.0), with only a slight overlap between CMP and SPC distributions, highlighting the necessity of concurrently using all three metrics.

Medical QA Benchmarks. Numerous significant studies have enriched the benchmarks of medical QA. MedQA (Jin et al., 2021) provides a trilingual, open-domain QA dataset derived from medical exams. MedMCQA (Pal et al., 2022) a benchmark primarily based on questions from Indian medical institutions. MultiMedQA (Pal et al., 2022) encompasses a wide range of healthcare topics through multiple-choice questions. MLEC-QA (Li et al., 2021) is the largest Chinese biomedical QA dataset from national exams. CARE-MI (Xiang et al., 2023) is a Chinese benchmark focusing on misinformation in maternity and infant care. CMExam (Liu et al., 2023a) is a dataset with explanations for model reasoning from Chinese medical exams. Recently, Cai et al. (2023) proposed MedBench, a comprehensive Chinese medical benchmark that employs BLEU and ROUGE for self-assessment.

3 Preliminaries

Generated Text Evaluation. We aim to assess the quality of generated responses in terms of factual correctness in medical QA scenarios, where the goal is to generate a response Rs based on a given question Qs. Commonly, one or multiple human references Rf are provided to aid this evaluation. The evaluation metrics are quantified by: $score = f(Rs|Qs, Rf, \alpha)$, where α denotes factual correctness aspects (e.g., accuracy).

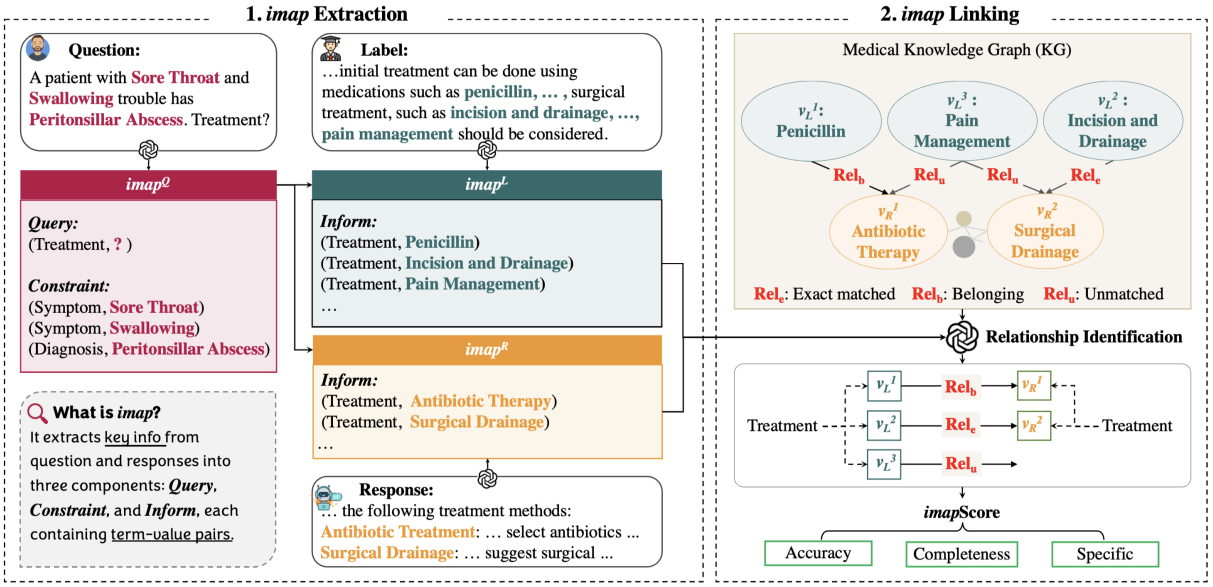


Figure 3: Our framework for *imap*Score calculated based on *imap* extraction and linking. We input the *imap* instruction into the LLMs to generate a question *imap* ($imap^Q$), which is further used to generate label and response *imap* ($imap^L$ and $imap^R$). A medical knowledge graph then aids the LLMs in identifying relationships between term's values of $imap^L$ and $imap^R$. Then the *imap*Score is computed based on these matching relationships.

Gold-standard Factual Correctness Evaluation. Currently, the gold standard for assessing the quality of generated text is human evaluation. This involves using a set of standard criteria to rigorously assess the correctness of medical facts contained within the text. In our benchmark, evaluators are instructed to score the generated responses from three commonly used perspectives independently:

- **Accuracy** (ACC) (Thirunavukarasu et al., 2023) assesses whether the response contains inaccuracies or unfactual content.
- **Completeness** (CMP) (Cai et al., 2023) measures how well the generated text captures the key ideas of the reference.
- **Specific** (SPC) (Fu et al., 2023) determines whether the generated text is generic or specific to the reference.

Figure 2 illustrates the distributions of CMP and SPC annotated scores at the same ACC, emphasizing the importance of using all these metrics. Scores range from 0.0 to 5.0. High ACC but low SPC often results when an answer is correct but too general, a common occurrence in LLMs answering medical questions. For instance, a response like "take medicine" for a specific drug reference is not incorrect, but it lacks detail. Similarly, mentioning only one correct drug out of the referenced three demonstrates ACC but lacks CMP.

The human evaluation process typically begins with identifying the core requirements of the question, followed by searching for and noting the essential elements in the reference answer that address these requirements, with an emphasis on highlighting key points. Subsequently, they assess how comprehensively the key points are covered in both the reference answer and the generated response before assigning a score. Consequently, we propose the *imap*Score to mimic human evaluation logic to improve automatic evaluation in medical QA, as detailed further in Section 4.2.

4 *imap*Score

imap structures questions and answers, encapsulating key information. From this, we propose an effective metric—*imap*Score.

4.1 Interactive MAP (*imap*)

The *imap*, an **i**nteractive ***map*** generalized for single-turn or multi-turn QA, is a data structure designed to distill questions or answers down to a level of medical factual detail that is sufficient for interaction. It consists of three key components: Query (Q), Constraint (C), and Inform (I), which encode the primary needs, their constraints, and the informed points, respectively. Each element within these components is a medical term-value pair. Given that one term can have multiple values, formally, *imap* is

represented as $imap = \{Q, C, I\}$, where $Q = \left[(t^1, v_1^1), (t^1, v_2^1), ..., (t^1, v_{L^1}^1), ..., (t^N, v_{L^N}^N)\right]$. Here N and L^i represent the unique number of medical terms in Q and the number of corresponding values related to t_i, respectively. In this context, "t" and "v" represent the term and the value respectively. The term probes the key intent of the question, while the values communicate the essential response points. These terms and values are not constrained to a fixed set; rather, they are dynamically generated by the LLMs. For instance, consider the following question: "A young woman displays symptoms like erythematous plaques and papules. What's the diagnosis? Which departments can offer treatment?". The Q can be represented as [(diagnosis, ?), (departments, ?)].

Both C and I share the same format as Q, but the specific terms and values can differ. The *imap* is capable of extracting key medical facts from QA interactions for rating. In single-turn QA, the question *imap* simplifies to $\{Q, C\}$, containing only the Query and Constraint, while the answer *imap* simplifies to $\{I\}$, containing only the inform contents. For instance, as illustrated in Figure 3, we derive the *imap* for Question as {Q=[(Treatment, ?)], C=[(Symptom, Sore Throat), ...]}. Similarly, the *imap* for Response is {I=[(Treatment, Antibiotic Therapy), (Treatment, Surgical Drainage)]}.

4.2 The *imap*-Based Metric: *imap*Score

The key insight of *imap*Score is that the values associated with a specific medical term, which are required in the question and present in the response *imap*, should closely align with those in the label *imap*. The extraction of *imap* is efficiently facilitated by LLMs (as elaborated in Section 4.3); however, clarifying the nature of these value relationships and what defines their alignment are essential aspects to consider. We draw upon the framework of relations between entities in medical knowledge graphs to divide the relationships among a label's value v^L and a response's value v^R into four distinct categories: 1) **Exact Match** (Rel_e): v^L and v^R sharing identical meanings but may be phrased through varied expressions, exemplified by "Incision and Drainage" and "Surgical Drainage"; 2) **Belonging** (Rel_b): when v^L is a subset of v^R, exemplified by "Penicillin" and "Antibiotic Therapy"; 3) **Containment** (Rel_c): when v^L includes v^R, exemplified by "Nephritis" and "Glomerulonephritis"; 4) **Unmatched** (Rel_u) relationships that do not fit into the above three categories.

Given the question and its *imap*, we assume that the *imap* for the human-annotated label and the generated response are represented as $imap^L = \left[(\dot{t}^1, \dot{v}_1^1), ..., (\dot{t}^i, \dot{v}_j^i), ...\right]$, and $imap^R = \left[(\bar{t}^1, \bar{v}_1^1), ..., (\bar{t}^i, \bar{v}_j^i), ...\right]$, respectively. Here $\dot{t}^i \in T_L$, $\bar{t}^i \in T_R$, where T_L and T_R are the sets of medical terms in $imap^L$ and $imap^R$, respectively. Similarly, $\dot{v}_j^i \in V_{\dot{t}^i}^L$, $\bar{v}_j^i \in V_{\bar{t}^i}^R$, where $V_{\dot{t}^i}^L$ and $V_{\bar{t}^i}^R$ are the sets of values associated with $\dot{t}^i$ and $\bar{t}^i$, respectively. Then

$$\textbf{\textit{imap}Score} = \sum_{t \in T_L \cap T_R} \left[\frac{\sum_{\dot{v} \in V_t^L} \mathbb{I}_{\mathcal{M}(\dot{v}, V_t^R)}(\dot{v})}{|V_t^L|} + \frac{\sum_{\bar{v} \in V_t^R} \mathbb{I}_{\mathcal{M}(\bar{v}, V_t^L)}(\bar{v})}{|V_t^R|} - \lambda_t \right], \tag{1}$$

where $\mathbb{I}_{\mathcal{M}(\cdot)}(v)$ is an indicator function that maps the value v to one if it satisfies the condition $\mathcal{M}(\cdot)$. $\mathcal{M}(v, V)$ denotes the scenerios where the best match between value v and all values in set V, determined by the priority order: $\text{Rel}_e, \text{Rel}_b, \text{Rel}_c, \text{Rel}_u$, is not Unmatched Rel_u. $|V_t^L|$ and $|V_t^R|$ are the number of values of term t corresponding to $imap^L$ and $imap^R$, respectively. The term $\lambda_t = \sum_{\dot{v} \in V_t^L, \bar{v} \in V_t^R} \frac{\mathbb{I}_{\mathcal{R}(\dot{v}, \bar{v}) = \text{Rel}_c}(\dot{v}, \bar{v})}{\mathbb{I}_{\mathcal{R}(\dot{v}, \bar{v}) \in \{\text{Rel}_e, \text{Rel}_b, \text{Rel}_c\}}(\dot{v}, \bar{v})}$ serves as a penalty term that penalizes the value relationships $\mathcal{R}(\dot{v}, \bar{v})$ where $\dot{v}$ includes $\bar{v}$. This arises when response values are more specific than their labels, leading to potential inaccuracies. For example, when the label is "Nephritis" and the response is specifically "Glomerulonephritis," it could potentially miss other conditions like "IgA Nephropathy", resulting in a misdiagnosis.

4.3 Implementing *imap*Score

The calculation of *imapScore* as outlined in Equation 1 necessitates $imap^L$, $imap^R$, and an understanding of the correspondence between their term values. We introduce a paradigm based on LLMs and a medical knowledge graph (KG) to facilitate the extraction of *imap* and the association of its values. The methodology comprises two main steps, as depicted in Figure 3. Initially, we provide GPT-4 with an instruction to identify the question *imap*. Subsequently, this question *imap* is incorporated into the instruction, prompting GPT-4 to generate both the label *imap* and the response *imap*. In the

Table 1: The *imap* extraction instruction template for questions and answers. Detailed information can be found in Appendix, Section C.

Question	Label or Response
Task Definition: ... Please refer to the following example to extract the *imap* for <Question>.	**Task Definition:** ... Please refer to the following examples to extract the *imap* for <Response>.
Example: [Example Question] [Example *imap*]	**Example:** [...]
<Question>: ... Please output <*imap*>:	<Question>: ... <*imap*>: ... <Response>: ... Please output <*imap*>:

second step, the KG is employed to aid GPT-4 in identifying the relationships between the values of medical terms for both the label *imap* and response *imap*. By following these procedures, we are equipped to compute the *imapScore* in accordance with Equation 1.

Question *imap* Extraction. The *imap* extraction process can be simplified by supplying GPT-4 with a natural language instruction. It should define the *imap*, outline the task, incorporate examples, and pose the question. An illustration of this prompt format is provided in Table 1. More specific prompts are elaborated in Table 5 in the Appendix.

Label or Response *imap* Extraction. To extract the *imap* for a label or response, we update the prompt by adding both the question and its *imap* when instructing GPT-4. As shown in the right column of Table 1, the process involves revising the task description, incorporating the question and its *imap* into the examples, and adding the label/response along with its corresponding question and *imap*. This approach facilitates the extraction of the response *imap*. Further examples are detailed in Table 6 in the Appendix.

***imap* Link: Value Relationship Identification.** The essential step in calculating the *imap*Score is to identify the relationships between each term in V_t^L and V_t^R with respect to term t. We propose two methods to achieve this: Medical KG-based and LLM-based. The KG-based method, as described in (Wang), is executed by 1) mapping each value to a KG entity that shares the same meaning, and 2) identifying whether there exist $\text{Rel}_e, \text{Rel}_b, \text{Rel}_c$ relations between KG-linked entities for values in V_t^L and V_t^R. Specifically, to map a value to KG, we employ a BERT-based encoder to derive word embeddings, generate entity candidates, re-rank them based on BERT similarity scores and the presence of the candidate entity in the value, and finally select the top-ranked entity as the mapped KG entity. To verify the matching relation for values in the two sets, we check for the existence of a "Synonym" relation as Rel_e and a "Belonging" relation as Rel_b or Rel_c for each mapped KG entity in V_t^L and V_t^R. For the LLM-based method, we provide GPT-4 with the label *imap* and response *imap*, and instruct it to generate pairs of term-value for label *imap* and response *imap* corresponding to the four matching relationships ($\text{Rel}_e, \text{Rel}_b, \text{Rel}_c, \text{Rel}_u$). The *imap*Score utilizes the medical KG to aid GPT-4 in determining the relationships between term values.

5 Experimental Settings

5.1 Evaluation Dataset Construction

Unlike previous research that assesses the overall text quality, we focus on evaluating factual correctness across multiple dimensions within the context of medical QA. Given the absence of publicly accessible medical QA datasets with human scores, we explored two available datasets: HMedQA, which is in Chinese, and iCliniq, in English. Both datasets feature single-turn QA interactions between patients and online doctors. Due to the presence of noise, such as incomplete answers and obviously factually false statements in the labels, we eliminate these noisy samples and retain as many as possible, resulting in 2,998 from HMedQA and 2,003 from iCliniq. A comprehensive analysis of these curated datasets is provided in Table 3.

Responses Generation. To thoroughly examine our approach for evaluating generated responses, we utilize LLMs to generate two distinct sets of responses for HMedQA and iCliniq, respectively. For HMedQA, we employed GPT-4 and PaLM-2 to generate responses derived from patient questions. Given that iCliniq already includes responses from ChatGPT, we supplemented it with an additional series of responses given by PaLM-2.

Human Scoring. Despite having questions, labeled answers, and generated responses, we still lack human scores to validate the consistency between our method and manual scoring, thereby proving the effectiveness of *imap*Score. To ensure the professionalism of manual scoring, we invited three medical experts to independently score the

Table 2: The Spearman's (ρ) and Kendall-Tau (τ) correlations of different metrics on HMedQA and iCliniq datasets show performance in accuracy (ACC), completeness (CMP), and specific (SPC). All values are multiplied by 100 for clarity. Darker shades of blue signify better performance. ***+imap*** represents replacing raw question text and answer text with their *imap* for evaluation. Nearly all ***+imap*** instances significantly ($p < 0.05$) outperform their base versions with exceptions marked by *. *imap*Score consistently exceeds all baseline metrics, and **+ KG** (indicating the use of KG to assist value relationship identification in *imap*Score) significantly ($p < 0.05$) exceeds the base *imap*Score, with exceptions marked by ¨. Notably, UniEval results for HMedQA, not supporting Chinese, and original ChatGPT responses for iCliniq were used as provided.

Metrics	HMedQA														iCliniq													
	GPT-4						PaLM-2								ChatGPT						PaLM-2							
	ACC		CMP		SPC		ACC		CMP		SPC		AVE		ACC		CMP		SPC		ACC		CMP		SPC		AVE	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	17	13	14	10	13	9	34	26	28	20	23	17	22	16	12	9	12	8	9	7	18	14	21	15	21	15	16	11
+imap	44	32	41	30	37	28	51	39	41	30	40	29	42	31	27	19	28	20	31	23	26	19	27	20	32	22	28	20
ROUGE-2	19	14	17	12	15	11	37	29	31	23	28	21	24	18	17	12	15	11	14	10	18	13	17	13	20	14	17	12
+imap	47	38	44	34	36	30	53	43	44	34	42	34	44	36	24	18	25	18	25	19	23	17	23	17	27	20	24	18
ROUGE-L	13	10	8	6	10	7	24	18	19	14	13	10	14	11	12	9	13	9	7	5	13	9	17	12	18	13	13	10
+imap	42	32	37	27	34	26	49	38	38	28	34	25	39	29	25	17	26	18	29	21	25	18	25	18	30	22	27	19
BLEU-4	13	10	6	5	9	6	26	19	21	15	15	11	15	11	19	14	18	13	14	10	16	12	17	12	19	13	17	12
+imap	35	26	30	22	29	22	43	33	32	23	27	20	33	24	23	17	24	17	24	18	22	17	23	17	24	17	23	17
BERTScore	18	13	14	10	14	10	24	18	20	14	17	12	18	13	21	15	23	16	19	14	18	13	20	15	22	16	20	15
+imap	37	27	36	26	35	26	41	31	31	23	30	23	35	26	39	28	40	29	39	29	32	24	31	23	36	26	36	26
MoverScore	27	19	22	16	18	13	41	31	35	25	28	21	28	21	18	12	17	12	14	10	20	15	22	16	23	16	19	14
+imap	47	35	42	31	37	28	55	43	44	34	38	29	44	33	38	28	39	29	36	26	34	26	30	22	33	24	35	26
BARTScore	18	13	18	13	16	12	33	24	25	19	20	15	22	16	12	8	12	9	9	7	11	8	11	8	13	9	11	8
+imap	31	23	28	20	28	21	40	30	32	23	28	20	31	23	33	24	32	23	31	22	26	19	22	16	27	20	28	21
UniEval	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15	11	14	10	8	6	4	3	7	5	3	3	8	6
+imap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	30	22	28	20	22	16	13	10	14	10	10	7	20	14
AlignScore	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	6	7	5	7	5	13	10	14	11	7	6	9	7
+imap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25	19	18	13	16	12	30	21	27	19	20	14	23	16
QuestEval	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21	23	31	22	26	19	21	16	23	17	21	15	24	19
+imap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37	27	38	27	34	25	25	19	30	23	31	22	33	24
QAFactEval	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24	18	22	16	21	16	24	18	23	16	23	16	23	17
+imap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25	19	23	18	23	16	27	19	25	18	26	16	25	18
GPTScore	33	25	45	39	27	22	38	31	35	28	25	21	34	28	27	23	26	21	17	16	25	20	24	20	23	15	24	19
+imap	52	31	48	40*	41	36	44	42	45	33	32	42	44	38	37	28	39	29	36	34	27	25	28	25	29	21	34	27
G-EVAL	47	39	49	40	26	22	44	34	41	33	32	25	40	32	32	27	30	24	22	18	23	19	26	21	26	19	26	21
+imap	57	47	51	42*	41	37	54	45	47	38	43	41	49	42	40	33	38	31	42	34	28	23	30	25	34	25	35	28
FActScore	20	17	20	17	27	24	23	18	21	17	29	24	23	20	7	6	6	5	6	4	12	10	17	14	11	9	10	8
+imap	52	41	57	46	42	34	54	40	53	41	39	35	50	40	27	19	17	16	17	15	33	28	30	23	24	19	25	20
***imap*Score**	63	50	62	49	44	41	62	53	60	50	49	45	57	48	56	47	55	43	55	48	37	30	33	27	40	33	46	38
+ KG	**66**	**54**	**65**	**53**	**47**	**44**	**68**	**57**	**64**	**52̈**	**53**	**48**	**60**	**51**	**60**	**50**	**60**	**50**	**61**	**53**	**48**	**40**	**44**	**35**	**40̈**	**33̈**	**52**	**44**

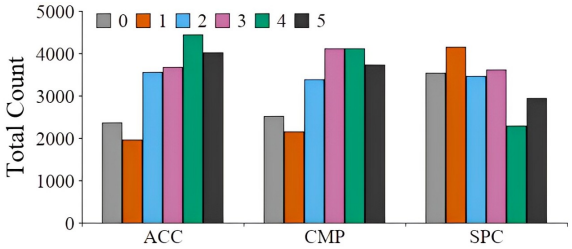


Figure 4: Distribution of human scores (ranging from 0 to 5) across ACC, CMP, and SPC dimensions.

two sets of generated responses of two datasets from three dimensions: Accuracy (ACC), Completeness (CMP), and Specific (SPC). After comprehensive training and a preliminary test, the experts' average ratings were used. As presented in Figure 4, the distribution of overall human scores across the three dimensions is relatively uniform, spanning from 0 to 5.

The percentage agreement among the three experts is calculated pairwise and then averaged. We categorize the scores into the following bins: {0}, {0.5, 1, 1.5}, {2, 2.5}, {3, 3.5}, {4,4.5}, {5}. We consider the ratings of two experts to "agree" if they fall within the same bin. The percentage of agreement for Accuracy, Completeness, and Specificity is 0.87, 0.85, and 0.81, respectively. We also calculated the average difference between the pairwise ratings of the experts on Accuracy, Completeness, and Specificity, which were 0.46, 0.47, and 0.53, respectively. The range of the rating is 5. Given that the raters underwent comprehensive training prior to the rating process, the inter-annotator agreement is quite promising.

5.2 Baselines, Correlations, and Dimension

We consider the metrics of three broad categories: 1) N-gram-based metrics, which include three variations of ROUGE (ROUGE-1, ROUGE-

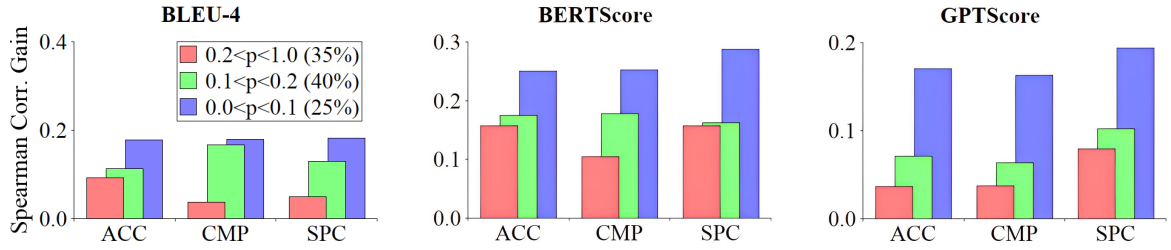


Figure 5: Illustration of the Spearman's correlation enhancement across ACC, CMP, and SPC by integrating *imap* into baselines. We divide the data into three buckets based on *imap* compression ratios (p), encoding the length ratio of *imap* to the original response, distinguished by colors. Legends display the sample distribution by p value range.

2, ROUGE-L) and a variation of BLEU (BLEU-4); 2) Embedding-based metrics, which include BERTScore, MoverScore, BARTScore, UniEval (Zhong et al., 2022), AlignScore (Zha et al., 2023), QuestEval(Scialom et al., 2021), and QAFactEval(Fabbri et al., 2021); 3) LLM-based methods, include GPTScore and G-EVAL. Details on these metrics have been discussed in Section 2. Our evaluation methodology aligns with that of G-EVAL, focusing on analyzing different metrics through sample-level Spearman (Zar, 2005)'s and Kendall-Tau (Kendall, 1938) correlation. The assessment spans three critical dimensions: ACC, CMP, and SPC. Implementation details for baselines and *imap*Score are in Appendix Section A.

6 Experiment Results

Our research aims to explore the following: 1) the efficacy of *imap*Score in medical QA evaluations; 2) the impact of incorporating *imap* on improving current leading approaches.

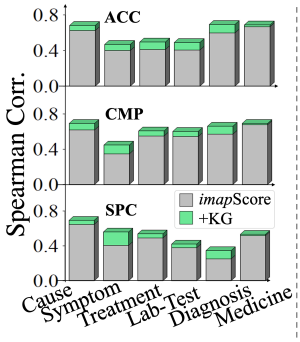
Main Results of *imap*Score. Table 2 displays Spearman's and Kendall-Tau Correlation across metrics for generated responses from two datasets, revealing significant insights: **1) Utilizing *imap* instead of raw text for evaluation significantly elevates baseline metrics**. The enhanced versions ("+*imap*") outperform the originals, with average Spearman's correlations for n-gram-based, embedding-based, and LLM-based metrics rising from 14.9 to 28.3, 15.6 to 28.4, and 28.0 to 37.1, respectively, showing relative improvements of 89.9%, 81.7%, and 32.6%. **2) *imap*Score consistently outperforms all baselines and their +*imap* variants**, surpassing even the previous state-of-the-art, prompt-based G-EVAL and its adding *imap* version, highlighting the importance of identifying *imap*'s value relationships in advancing automatic evaluation methodologies. **3) Integrating KG (+KG) into *imap*Score achieves the best agreement with humans**, resulting in a remarkable 10.1% improvement over *imap*Score alone, which is not surprising since KG helps in refining the unification of value expressions.

Other interesting findings include: 1) LLM-based metrics surpass other baselines, highlighting their effectiveness in medical QA evaluation. 2) On HMedQA, PaLM-2 surpasses GPT-4 through better medical language alignment with labels, but *imap*Score boosts GPT-4's human score alignment, surpassing PaLM-2. 3) Differences in HMedQA and iCliniq correlations hint at dataset variability, yet *imap*Score consistently enhances evaluation performance across both.

6.1 Further Analysis

Further research explores *imap*Score's capability through three questions: 1) How does *imap* integration enhance baseline metrics? 2) What advantages does a KG offer in evaluation? 3) Is *imap*Score applicable in the absence of golden labels?

The Effect of *imap*. The *imap* aims to distill raw text, directing the evaluator's attention towards key info for better ratings. This section delves into the workings of the *imap* mechanism, examining its impact on baseline metrics across *imap* compression ratios ("p"). These ratios represent the length ratio of *imap* to the original response. The outcomes using BLEU-4, BERTScore, and GPTScore metrics are reported in Figure 5, with additional baseline results in Figure 8 in the Appendix. The main results show an inverse correlation between *imap*'s improvement and its compression ratios p; specifically, the lower the *imap* compression ratios, the more significant the absolute improvement gained from integrating *imap*. This outcome is expected, as a smaller p value indicates scarce key information in the raw text, allowing *imap* to more efficiently eliminate irrelevant data when comparing labels and responses for evaluation.



Case 1: Medicine

Label	Response	GPT-4	KG
hereditary hemorrhagic disorder	hereditary abnormality	Rel_e	Rel_b √

Case 2: Treatment

Label	Response	GPT-4	KG
palliative care	painkiller	Rel_u	Rel_b √

Case 3: Lab-Test

Label	Response	GPT-4	KG
ELISA	immunologic test	Rel_e	Rel_b √

Figure 6: The left figure illustrates the Spearman's correlation boost across six medical tasks with KG assisting *imap*Score. The right tables show three cases where KG surpasses GPT-4 in accurately linking medical values.

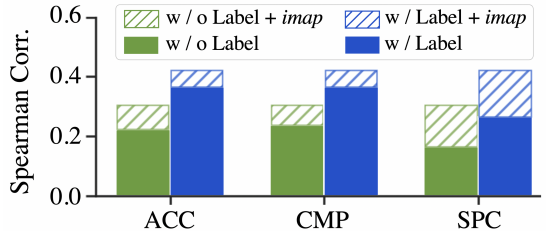


Figure 7: Spearman's correlation comparisons of G-EVAL, both with and without labels, alongside the enhancements from integrating *imap* (+*imap*).

The Effect of KG Integration. While *imap*Score insisted with KG, it achieves the best performance in Table 2, this section aims to dissect KG's benefits into subtasks to illustrate its true performance gains. As depicted in Figure 6, GPT-4 shows better performance in the field of medicine, where the contribution of the KG is less evident. The gains from the KG are relatively average in other tasks. Furthermore, we observe that GPT-4 tends to give "Exact Match" (Rel_e) and "Unmatched" (Rel_u) relationships. In contrast, the KG, with explicit knowledge encoded within, is capable of delivering more accurate relationships, as showcased through the three examples in the right-hand tables in Figure 6.

Adaptability to Unlabeled Scenarios. To explore the efficacy of *imap*Score in the absence of golden labels, we compared G-EVAL and G-EVAL + *imap* in labeled and unlabeled scenarios. Figure 7 reveals that: 1) *imap* is effective even without labels, as shown by the superior performance of the "w/o label +*imap*" compared to using only raw text "w/o label"; 2) having labels benefits performance, as their absence tends to reduce the performance. See Figure 9 in Appendix for further details.

7 Conclusions

This paper presented *imap*, a data structure that parses medical QA into term-value pairs, enhancing evaluation by capturing essential needs and evidence from QAs. Besides, we proposed *imap*Score, a novel metric that uses *imap* for text scoring, aligning closely with expert assessments. Furthermore, we built a new benchmark for evaluating NLG in medical QA, focusing on the medical factual correctness generated by LLMs. Experimental results demonstrated that *imap*Score is in better agreement with humans compared to existing metrics.

Limitations

In this study, we have identified two primary limitations of *imap*Score that warrant further exploration in subsequent research endeavors. The first limitation pertains to the potential insufficiency of the Knowledge Graph (KG) used in our experiments, both in terms of coverage and entity relations. This could introduce bias into the effectiveness of KG's contribution to enhancing *imap*Score. Despite the current "*imap*Score + KG" version demonstrating superior performance in aligning with human scoring, this issue hinders our ability to fully substantiate the positive impact of KG on *imap*Score. To mitigate this, potential solutions include integrating *imap* with a more extensive medical KG that includes a wider array of medical terminologies and considering a greater number of entity relationships for matching.

Additionally, our current benchmark dataset is limited to single-turn QA interactions, which prevents the assessment of *imap*Score's performance in multi-turn QA scenarios, despite *imap*'s inherent design to handle such interactions. To evaluate *imap*'s proficiency in multi-turn settings, a dataset featuring multi-turn medical QAs with human scores is necessary. In our future work, we plan to expand our benchmark to encompass multi-turn QAs, facilitating a more comprehensive evaluation of the *imap*Score's capabilities.

Ethics Statement

Our work adheres to the ACL Ethics Policy. This paper aims to investigate automatic evaluation metrics for question-answering (QA) systems, with the objective of minimizing dependence on manual evaluation and enhancing automatic evaluation methods, thereby simplifying the review process and hastening advancements in medical QA. It is

crucial to emphasize that the proposed metric and benchmark are designed solely for research purposes and are not suitable for direct clinical application due to the potential risks associated with the misuse of medical QA systems.

It is important to note that the introduced benchmark was sourced from two publicly available QA datasets collected from online interactions between doctors and patients, with all patient privacy-related information meticulously eliminated. To ensure data privacy and security, we performed a comprehensive manual review of the dataset, confirming that it contains no identifiable or offensive pieces of information within the experimental dataset. This review was carried out by medical experts serving as human evaluators, who have recommended releasing the dataset for research purposes based on their assessment.

References

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. Medbench: A large-scale Chinese Benchmark for evaluating medical large language models. *arXiv preprint arXiv:2312.12806*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Bright Huo, Giovanni E Cacciamani, Gary S Collins, Tyler McKechnie, Yung Lee, and Gordon Guyatt. 2023. Reporting standards for the use of large language model-linked chatbots for health advice. *Nature Medicine*, pages 1–1.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese multi-choice biomedical question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874.

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.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2023a. Benchmarking Large Language Models on CMExam–A Comprehensive Chinese Medical Exam Dataset. *arXiv preprint arXiv:2306.03030*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. GPTEeval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

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.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical AI. *arXiv preprint arXiv:2307.14334*.

Yuyang Wang. End-to-end entity linking combined with BERT-based Siamese and Interaction Network. In *Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City*, pages 47–53.

Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. 2023. CARE-MI: Chinese benchmark for misinformation evaluation in maternity and infant care. *arXiv preprint arXiv:2307.01458*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jerrold H Zar. 2005. Spearman Rank Correlation. *Encyclopedia of biostatistics*, 7.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

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. *arXiv preprint arXiv:2210.07197*.

A Experiment Settings

Benchmark Dataset Size The comprehensive breakdown of samples in both HMedQA and iCliniq datasets utilized for our benchmark is depicted in Table 3 and Table 4. As shown in Table 4, it is noteworthy that the average length of questions and answers in the iCliniq dataset is significantly longer compared to HMedQA. This disparity may be attributed to cultural differences, as doctors in western countries often tend to express empathy towards patients before offering medical advice.

Settings for Hyperparameters In generating responses from Large Language Models to queries within the HMedQA and iCliniq, we applied the subsequent hyperparameter configurations:

- The "temperature" was configured to 0.9 and "top p" to 0.95 for both GPT-4 and PaLM-2.
- The original responses provided by ChatGPT for the iCliniq dataset were utilized as is.

Configuration of Baseline Models In the deployment of baseline models, we detail the optimum outcomes from trials involving diverse backbone models and hyperparameter adjustments across each dataset:

- BERTScore (Zhang et al., 2019): Employed the "distilbert-base-uncased" as its foundational backbone model.
- MoverScore (Zhao et al., 2019): For the HMedQA dataset evaluation, "bert-base-chinese" was selected as the backbone model; conversely, for iCliniq, the "distilbert-base-uncased" model was maintained.
- BARTScore (Yuan et al., 2021): Utilized "facebook/bart-large-cnn" as its primary backbone model.
- UniEval (Zhong et al., 2022): Adopted "MingZhong/unieval-sum" as the model. Notably, UniEval results for HMedQA were omitted in Table 2 due to its incompatibility with Chinese content.
- G-EVAL (Liu et al., 2023b): Configured the "temperature" to 1.0 and "top p" to 1.0 in alignment with its specified hyperparameters. Results were derived by averaging five samples.

***imap* Extraction Settings** We configured the "temperature" parameter to 0.9 and "top p" to 1 to prompt GPT-4 to extract *imap* and analyze the relationships among values. This experimental procedure was conducted three times to guarantee reliability, selecting outcomes for term-value pairs and value relationships that were consistently replicated more than twice.

Medical Knowledge Graph (KG) The KG utilized in our experiments comprises over three million entities and twelve million relationships. Comprehensive details about the KG are available on our website. The link will be included in the final version of our paper, as it contains sensitive identity information. We plan to make the KG accessible for research purposes.

B Experiment Results

This section reports further insights into Spearman's correlation improvement with *imap*. Firstly, we showcase additional findings on how Spearman's correlation is enhanced through *imap* integration as opposed to *imap* compression ratios. These results are elaborated in terms of ACC, CMP, and SPC metrics, spanning nine baseline evaluations, as depicted in Figure 8.

Furthermore, Figure 9 illustrates the comparative outcomes of applying *imap* in contexts where golden labels are absent within the G-EVAL framework. These results are segmented across various LLMs and datasets for a detailed analysis.

C Example Prompts

We provide the details of the prompts used for *imap* extraction pertaining to questions and answers, as presented in Tables 5 and 6.

Table 3: Data sizes for each medical task in HMedQA and iCliniq.

Dataset	Advice	Treatment	Diagnosis	Cause	Lab-Test	Symptom	Medicine	Department	Others	Sum
HMedQA	125	1,110	769	44	160	121	77	67	1525	2,998
iCliniq	335	291	340	330	158	157	80	162	150	2,003

Table 4: Analysis of Question and Answer Lengths in HMedQA and iCliniq.

Dataset	Type	Advice	Treatment	Diagnosis	Cause	Lab-Test	Symptom	Medicine	Department
HMedQA	Question	42	45	46	39	47	33	38	41
HMedQA	Answer	83	68	62	47	51	51	65	25
iCliniq	Question	295	292	277	214	253	236	208	267
iCliniq	Answer	540	490	540	528	560	531	476	324

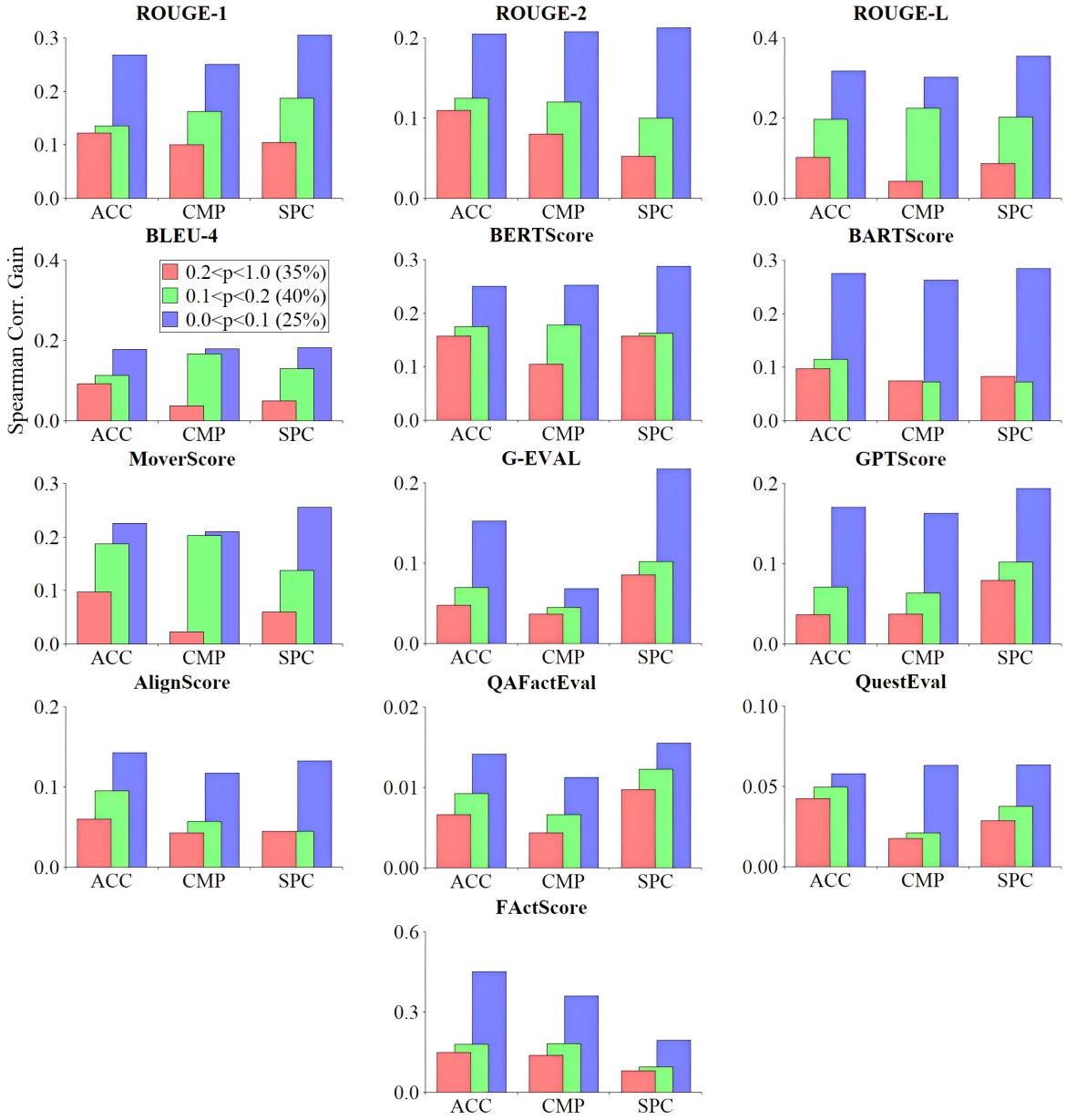


Figure 8: Illustration of the Spearman's correlation enhancement across ACC, CMP, and SPC by integrating *imap* into baseline metrics. We divide the data into three buckets based on *imap* compression ratios (p), encoding the length ratio of *imap* to the original response, distinguished by colors. Legends display the sample distribution by p value range.

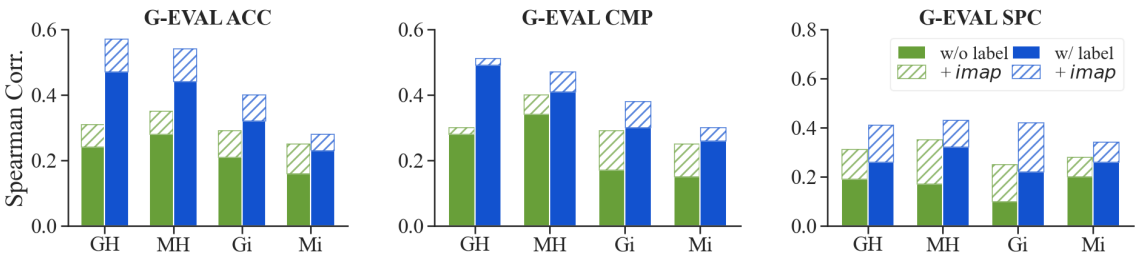


Figure 9: Spearman's correlation of G-EVAL with and without label, and the improvement achieved by adding *imap* on four sets of generated responses. GH, MH, Gi, and Mi represent GPT-4 on HMedQA, PaLM-2 on HMedQA, ChatGPT on iCliniq, and PaLM-2 on iCliniq, respectively.

Prompt:
Task Definition:
imap is a data structure that extracts key info from a question and responses into three components: Query, Constraint, and Inform, each containing term-value pairs.
Please use "Query" and "Constraint" intent to extract the imap for <Question>. You can refer to the following examples first:

Example 1:
<Question>:
Hi doctor, I am a 28-year-old male with redness and pain in the joint area and difficulty urinating. I have a recent history of microbial infection. What disease might I have? How should it be treated?
<imap>:
Query-disease-?
Query-treatment-?
Constraint-gender-male
Constraint-age-28
Constraint-symptom-redness and pain
Constraint-symptom-difficulty urinating
Constraint-symptom-history of microbial infection

Example 2:
<Question>:
Hi doctor, are the treatment plans for gestational hypertension and esophageal cancer the same?
<imap>:
Query-are the treatment plans the same-?
Constraint-disease-gestational hypertension
Constraint-disease-esophageal cancer

<Question>:
Hello doctor, My friend aged 30 had two drops of phenol mistaking for milk. He vomited and had a lot of saltwater. Please advise for any side effects.
<imap>:

Output:
Query-side effects-?
Constraint-age-30
Constraint-substance ingested-phenol
Constraint-action-vomited
Constraint-action-had a lot of saltwater

Table 5: Question *imap* extraction prompt.

Prompt:
Task Definition:
imap is a data structure that extracts key info from a question and responses into three components: Query, Constraint, and Inform, each containing term-value pairs.

You will be given one <Question> along with a <Question imap>. Your task is to extract the “inform” component from <Response> to replay to the “Query” component in <Question imap>. That is you should target “Query-diagnosis-?” and FILL the "[value]" in one or several actions “Inform-diagnosis-[value]”. Please note that the [value] should include ALL the key information. Besides, [Value] should be as SHORT as possible. You can use words or phrases directly extracted from the text to fill in [value]. Just directly give a conclusion without explanation. If there is no relevant information, output Inform-None.

<Question>:
Hi doctor, I have a sore throat on one side of my throat. But, after looking, it looks like a sore next to my throat on the roof of my mouth. What could it be?
<Question imap>:
Query-diagnosis-?
Constraint-symptom-sore throat on one side
Constraint-symptom-sore on the roof of mouth

<Response>: Hi. I have gone through the attachment (attachment removed to protect patient identity). It is oral stomatitis with pharyngitis. It can be due to smoking and alcohol infection. Certain investigations like Hb (hemoglobin), TLC (total leucocyte count), DLC (differential leucocyte count), and ESR (erythrocyte sedimentation rate) can be done. It can be treated by taking oral Vitamin B complex tablets with Chlorhexidine mouthwash. Avoid spicy foods and smoking. For more information consult an ENT otolaryngologist online. Take care.
<imap>:

Output:
Inform-diagnosis-oral stomatitis with pharyngitis

Table 6: Response *imap* extraction prompt.