npj | health systems

Review

https://doi.org/10.1038/s44401-024-00011-2

Current and future state of evaluation of large language models for medical summarization tasks

Emma Croxford[1], Yanjun Gao[2], Nicholas Pellegrino[3], Karen Wong[3], Graham Wills[4], Elliot First[3], Frank Liao[4,5], Cherodeep Goswami[4], Brian Patterson[4,5] & Majid Afshar[4,6] ✉

Large Language Models have expanded the potential for clinical Natural Language Generation (NLG), presenting new opportunities to manage the vast amounts of medical text. However, their use in such high-stakes environments necessitate robust evaluation workflows. In this review, we investigated the current landscape of evaluation metrics for NLG in healthcare and proposed a future direction to address the resource constraints of expert human evaluation while balancing alignment with human judgments.

The rapid development of Large Language Models (LLMs) has led to significant advancements in the field of Natural Language Generation (NLG). In the medical domain, LLMs have shown promise in reducing documentation-based cognitive burden for healthcare providers, particularly in NLG tasks such as summarization and question answering. Summarizing clinical documentation has emerged as a critical NLG task as the volume of medical text in Electronic Health Records (EHRs) continues to expand[1].

Recent advancements, like the introduction of larger context windows in LLMs (e.g., Google's Gemini 1.5 Pro with a 1 million-token capacity[2]), allow for the processing of extensive textual data, making it possible to summarize entire patient histories in a single input. However, a major challenge in applying LLMs to high-stakes environments like medicine is ensuring the reliable evaluation of their performance. Unlike traditional approaches, generative AI (GenAI) offers greater flexibility by generating natural language narratives that use language dynamically to fulfill tasks. Yet, this flexibility introduces added complexity in assessing the accuracy, reliability, and quality of the generated output where the desired response is not as static.

The evaluation of clinical summarization by LLMs must address the intricacies of complex medical texts and tackle LLM-specific challenges such as relevancy, hallucinations, omissions, and ensuring factual accuracy[3]. Healthcare data can further complicate the LLM-specific challenges because they can contain conflicting or incorrect information. Current metrics, like n-gram overlap and semantic scores, used in summarization tasks are insufficient for the nuanced needs of the medical domain[4]. While these metrics may perform adequately for simple extractive summarization, they fall short when applied to abstractive summarization[5], where complex reasoning and in-depth medical knowledge are required. They are also unable to differentiate the needs of various users or provide evaluations that account for the relevancy of generations.

In the era of GenAI, automation bias further complicates the potential risks posed by LLMs, particularly in clinical settings where the consequences of inaccuracies can be severe. Therefore, efficient and automated evaluation methods are essential. In this review, we examine the current state of LLM evaluation in summarization tasks, highlighting both its applications and limitations in the medical domain. We also propose a future direction to overcome the labor-intensive process of expert human evaluation, which is time-consuming, costly, and requires specialized training.

Search strategy and selection criteria

Comprehensive literature searches were conducted across multiple databases focused on summarization and question-answering tasks with a special focus on clinical applications (Fig. 1). From April 20, 2023 through August 3, 2023, searches were conducted across the Association for Computational Linguistics (ACL) anthology, Medline, and Scopus databases for literature that employed human frameworks or pre-LLM automated metrics for evaluative efforts related to these tasks. This search resulted in 262 abstracts for review. From April 16, 2024 through June 6, 2024, searches were conducted across the Association for Computational Linguistics (ACL) anthology, Association for Computing Machinery (ACM) Digital Library, Web of Science, Institute of Electrical and Electronics Engineers (IEEE) Xplore, and Scopus databases for literature that utilized large language models in evaluative processes related to these tasks. This search resulted in 95 abstracts for review. The free text, filters, and queries by database for each search can be found in Supplementary Tables 1 and 2.

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA. [2]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [3]Epic Systems, Verona, WI, USA. [4]UW Health, Madison, WI, USA. [5]BerbeeWalsh Department of Emergency Medicine, University of Wisconsin, Madison, WI, USA. [6]Department of Medicine, University of Wisconsin, Madison, WI, USA.
✉e-mail: mafshar@medicine.wisc.edu

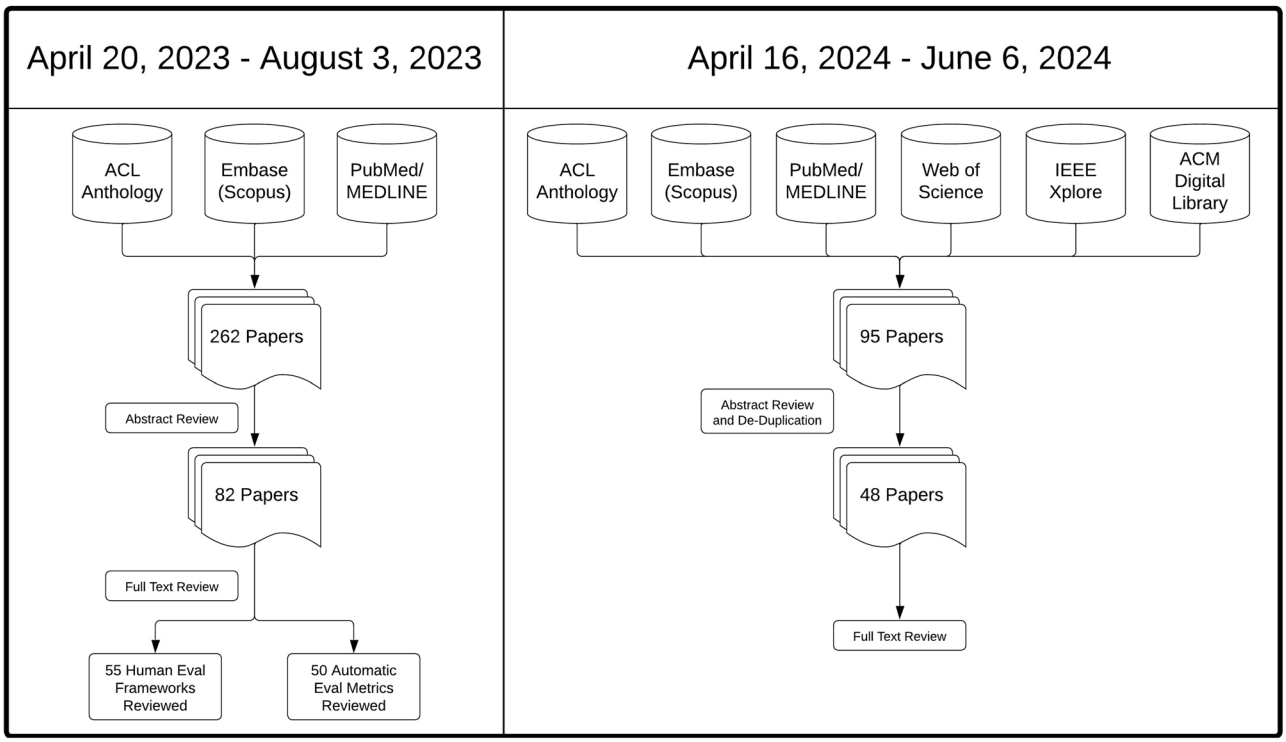

Fig. 1 | Literature search overview. A high level view of the two literature searches conducted for this review and their results.

Materials were selected for inclusion in this review if they (1) introduced novel human evaluation frameworks or automated metrics, (2) centered on a clinically-relevant summarization task, or (3) demonstrated improvements over Recall-Oriented Understudy for Gisting Evaluation (ROUGE). Following abstract reviews with these criteria, there were 82 and 48 papers respectively that underwent full text review for a total of 130. We also choose to include any materials that were referenced by multiple other articles when presenting potential improvement comparisons or fundamental knowledge pertaining to the task of evaluation.

Human evaluations in electronic health record documentation

The current human evaluation frameworks for human-authored clinical notes are largely based on pre-GenAI rubrics that assess clinical documentation quality. These frameworks vary depending on the type of evaluators, content, and the analysis required to generate evaluative scores. Such flexibility allows for tailored evaluation methods, capturing task-specific aspects that ensure quality generation. Expert evaluators, with their field-specific knowledge, play a crucial role in maintaining high standards of assessment.

Some commonly used pre-GenAI rubrics include the SaferDx[6], Physician Documentation Quality Instrument (PDQI-9)[7], and Revised-IDEA[8] rubrics. The SaferDx rubric focuses on identifying diagnostic errors and analyzing missed opportunities in EHR documentation through a 12-question retrospective survey aimed at improving diagnostic decision-making and patient safety. The PDQI-9 evaluates physician note quality across nine criteria questions, ensuring continuous improvement in clinical documentation and patient care. The Revised-IDEA tool offers feedback on clinical reasoning documentation through a 4-item assessment. All three of these rubrics place emphasis on the omission of relevant diagnoses throughout the differential diagnosis process and the relevant objective data, processes, and conclusions associated with those diagnoses. They also require clinical documentation to be free of incorrect, inappropriate, or incomplete information emphasizing the importance of the quality of evidence and reasoning that is present in clinical documentation. Each rubric includes additional questions based on the origin and usage of specific clinical documentation — like the PDQI-9's assessment of organization to ensure a reader is able to understand the clinical course of a patient. Each of the three also uses different assessment styles based on the granularity of the questions and intention behind the assessment. For instance, the Revised-IDEA tool uses a count style assessment for 3 of the 4-items to guarantee the inclusion of a minimum number of objective data points and inclusion of required features for a high-quality diagnostic reasoning documentation. In recent publications, the SaferDx tool has been used as a retrospective analysis of the use of GenAI in clinical practice[9], whereas the PDQI-9 and Revised-IDEA tools have been utilized to compare the quality of clinical documentation that is written by clinicians versus GenAI methods[10–12]. While each of these rubrics was not originally designed to evaluate LLM-generated content, they offer valuable insights into the essential criteria for evaluating text generated in the medical domain.

Human evaluations remain the gold standard for LLM outputs[13]. However, because these rubrics were initially developed for evaluating clinician-generated notes, they may need to be adapted for the specific purpose of evaluating LLM-generated output. Several new and modified evaluation rubrics have emerged to address the unique challenges posed by LLM-generated content, including evaluating the consistency and factual accuracy (i.e., hallucinations) of the generated text. Common themes in these adapted rubrics include safety[14], modality[15,16], and correctness[17,18].

Criteria for human evaluations

In general, the criteria that are used to make up evaluation rubrics for LLM output fall into seven broad criteria: (1) *Hallucination*[4,17–22], (2) *Omission*[14,19], (3) *Revision*[23], (4) *Faithfulness/Confidence*[15,16,23], (5) *Bias/Harm*[14,16,22], (6) *Groundedness*[14,15], and (7) *Fluency*[15,17,20,23]. *Hallucination* encompasses any evaluative questions that intend to capture when information in a generated text does not follow from the source material. Unsupported claims, nonsensical statements, improbable scenarios, and incorrect or contradictory facts would be flagged by questions in this criteria. *Omission*-based questions are used to identify missing information in a generated text. Medical facts, important information, and critical diagnostic decisions can all be considered omitted when not included in generated text, if those items would have been included by a medical professional. When an evaluator is

asked to make revisions or estimate the number of revisions needed for a generated text, the evaluative question would fall under *Revision*. Generated texts are revised until they meet the standards set forth by a researcher, hospital system, or larger government body. *Faithfulness/Confidence* is generally characterized by questions that capture whether a generated text has preserved the content of the source text and presented conclusions that reflect the confidence and specificity present in the source text. Questions about *Bias/Harm* evaluate whether generated text is introducing potential harm to a patient or reflecting bias in the response. Information that is inaccurate, inapplicable, or poorly applied would be captured by questions that fall under this criteria. *Groundedness* refers to evaluative questions that grade the quality of the source-based evidence for a generated text. Any evidence that contains poor reading comprehension, recall of knowledge, reasoning steps, or is antithetical to scientific consensus would result in a poor groundedness score. In addition to the content of a generated text, the *Fluency* of a generated text is also included in evaluations. Coherency, readability, grammatical correctness, and lexical correctness fall under this criteria. In many cases, Fluency is assumed to be adequate in favor of focusing on content-based evaluative criteria.

Analysis of human evaluations

The method of analysis for evaluation rubrics can also vary based upon the setting and task. Evaluative scores can be calculated using binary/Likert categorizations[14,15], counts/proportions of pre-specified instances[22], edit distance[23], or penalty/reward schemes similar to those used for medical exams[24]. *Binary* categorizations answer evaluative questions using True/False or Yes/No response schema. This set-up allows complex evaluations to be broken down into simpler and potentially more objective decisions. A binary categorization places more penalization on smaller errors by pushing responses to be either acceptable or unacceptable. *Likert*-scaled categorizations allow for a higher level of specificity in the score by providing an ordinal scale. These scales can consist of as many levels as necessary, and in many cases there are between 3 and 9 levels including a neutral option for unclear responses. Scales with a higher number of levels introduce more problems with meeting assumptions of a normal distribution into an analysis, along with complexity and disagreement amongst reviewers. *Count/proportion*-based evaluations require an evaluator to identify pre-specified instances of correct or incorrect key phrases related to a particular evaluative criteria. A precision, recall, f-score, or rate can then be computed from an evaluator's annotations to establish a numerical score for a generated text. *Edit distance* evaluations also require an evaluator to make annotations on the generated text that is being evaluated. In these cases, an evaluator makes edits to the generated text until it is satisfactory or no longer contains critical errors. These edits can be corrections on factual errors, inclusion of omissions, or removal of irrelevant items. The evaluative score is the distance from the original generated text and the edited version based upon the number of characters, words, etc. that required editing. The Levenshtein distance[25] is an example of an algorithm used to calculate the distance between the generated text and its edited version. This distance is calculated as the minimum number of substitutions, insertions, and deletions of individual characters required to change the original to the edited version. Finally, one of the more complex ways to compute evaluative scores is to use a *Penalty/Reward* schema. These schema award points for positive outcomes to evaluative questions and penalize negative outcomes. This schema is similar to those seen on national exams which account for positive and negative scores, using the importance and difficulty associated with different questions. For example, the schema used to evaluate LLMs on the Med-HALT dataset is an average of the correct and incorrect answers which are assigned $+1$ and -0.25 points respectively[24]. This evaluation schema provides a high level of specificity for assigning weights representative of the trade-off between false positives and false negatives.

Drawbacks of human evaluations

While human evaluations provide nuanced assessments, they are resource-intensive and heavily reliant on the recruitment of evaluators with clinical domain knowledge. The experience and background of an evaluator can significantly influence how they interpret and evaluate generated text. Additionally, the level of guidance and specificity in evaluative instructions determines how much of the assessment is shaped by the evaluators' personal interpretations and beliefs about the task. Although increasing the number of evaluators could mitigate some of these biases, resources—both time and financial—often limit the scale of human evaluations. These evaluations also require substantial manual effort, and without clear guidelines and training, inter-rater agreement may suffer. Ensuring that human evaluators align with the evaluation rubric's intent requires training, much like annotation guidelines for NLP shared tasks[26–28]. In the clinical domain, medical professionals are typically used as expert evaluators, but their time constraints limit their availability for large-scale evaluations. The difficulty of recruiting more medical professionals, compounded by the time needed for thorough assessments, makes frequent, rapid evaluations impractical.

Another concern is the validity of the evaluation rubric itself. A robust human evaluation framework must possess strong psychometric properties, including construct validity, criterion validity, content validity, and inter-rater reliability, to ensure reproducibility and generalizability. Unfortunately, many frameworks used in clinical evaluations do not provide sufficient details about their creation, making it difficult to assess their validity[15,24]. Often, human evaluation frameworks are developed for specific projects with only one evaluator, and while metrics like inter-rater reliability are crucial to establish validity, they are not always reported[18,23]. Moreover, clinically relevant evaluation rubrics have not been specifically designed to assess LLM-generated summaries. Most existing evaluation rubrics focus on assessing human-authored note quality, and they do not encompass all the elements required to evaluate the unique aspects of LLM-generated outputs[6–8].

Pre-LLM automated evaluations

Automated metrics offer a practical solution to the resource constraints of human evaluations, particularly in fields like Natural Language Processing (NLP), where tasks such as question answering, translation, and summarization have long relied on these methods. Automated evaluations employ algorithms, models, or heuristic techniques to assess the quality of generated text without the need for continuous human intervention, making them far more efficient in terms of time and labor. These metrics, however, depend heavily on the availability of high-quality reference texts, often referred to as "gold standards." The generated text is compared against these gold standard reference texts to evaluate its accuracy and how well it meets the task's requirements. Despite their efficiency, automated metrics may struggle to capture the nuance and contextual understanding required in more complex domains, such as clinical diagnosis, where subtle differences in phrasing or reasoning can have significant implications. Therefore, while automated evaluations are valuable for their scalability, their effectiveness is closely tied to the quality and relevance of the reference texts used in the evaluation.

Categories of automated evaluation

Automated evaluations in the clinical domain can be categorized into five primary types (Fig. 2), each tailored to specific evaluation goals and dependent on the availability of reference and source material for the generated text: (1) *Word/Character-based*, (2) *Embedding-based*, (3) *Learned metrics*, (4) *Probability-based*, (5) and *Pre-Defined Knowledge Base*.

Word/Character-based evaluations rely on comparisons between a reference text and the generated text to compute an evaluative score. These evaluations can be based on character, word, or sub-sequence overlaps depending on the need of the evaluation and the nuance that may be present in the text. Recall Oriented Understudy for Gisting Evaluation (ROUGE)[29] is a prime example of a word/character-based metric. The many variants of ROUGE — N-gram Co-Occurrence (N), Longest Common Sub-sequence (L), Weighted Longest Common Sub-sequence (W), Skip-Bigram Co-Occurrence (S) — represent the level of comparison between the reference and generated texts. ROUGE-L is the current gold standard for automated

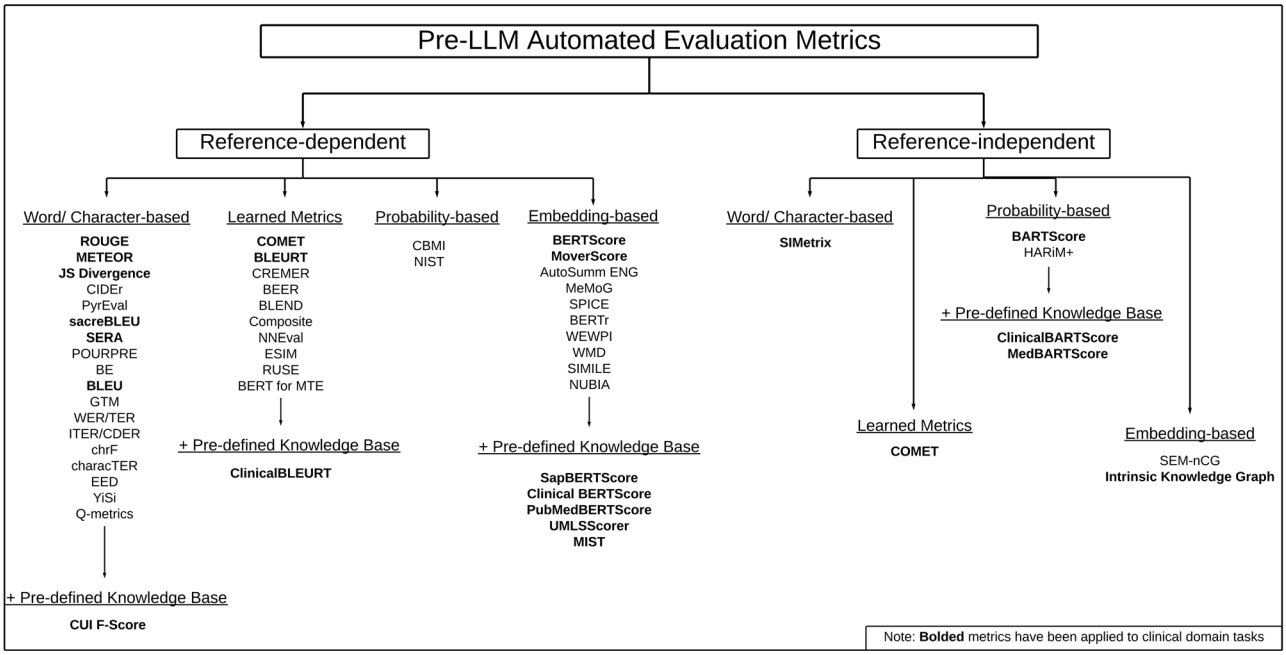

Fig. 2 | Pre-LLM automated evaluation metric taxonomy. A structured organization of pre-LLM automated evaluation metrics categorized by their bases and the need for ground truth references. Those metrics that were built for or have been applied in the clinical domain are in bold. The taxonomy includes Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[29], Metric for Evaluation of Translation with Explicit Ordering (METEOR)[66], Jensen-Shannon (JS) Divergence[67], Consensus-based Image Description Evaluation (CIDEr)[68], PyrEval[69], Standardized Bilingual Evaluation Understudy (sacreBLEU)[70], Summarization Evaluation by Relevance Analysis (SERA)[71], POURPRE[72], Basic Elements (BE)[73], Bilingual Evaluation Understudy (BLEU)[70], General Text Matcher (GTM)[74], Word Error Rate (WER)[75]/ Translation Edit Rate (TER)[76], Improving Translation Edit Rate (ITER)[77]/ CDER (Cover-Disjoint Error Rate)[78], chrF (character n-gram F-score)[79], characTER (Character Level Translation Edit Rate)[80], Extended Edit Distance (EED)[81], YiSi[82], Q-metrics[83], Concept Unique Identifier (CUI) F-Score[37], Crosslingual Optimized Metric for Evaluation of Translation (COMET)[32], Bilingual Evaluation Understudy with Representations from Transformers (BLEURT)[84], Combined Regression Model for Evaluating Responsiveness (CREMER)[85], Better Evaluation as Ranking (BEER)[86], BLEND[87], Composite[88], Neural Network Based Evaluation Metric (NNEval)[88], Enhanced Sequential Inference Model (ESIM)[89], Regressor Using Sentence Embeddings (RUSE)[90], Bidirectional Encoder Representations from Transformers for Machine Translation Evaluation (BERT for MTE)[91], ClinicalBLEURT[19], Conditional Bilingual Mutual Information (CBMI)[92], NIST[93], BERTScore[30], MoverScore[94], AUTOmatic SUMMary Evaluation based on N-gram Graphs (AutoSumm ENG)[95], Merge Model Graph (MeMoG)[95], Semantic Propositional Image Caption Evaluation (SPICE)[96], BERTr[97], Word Embedding-based automatic MT evaluation using Word Position Information (WEWPI)[98], Word Mover-Distance (WMD)[99], SIMILE[100], NeUral Based Interchangeability Assessor (NUBIA)[101], SapBERTScore[102], ClinicalBERTScore[103], PubMedBERTScore[104], UMLSScorer[38], MIST[19], Summary-Input Similarity Metrics (SIMetrix)[67], BARTScore[33], Hallucination Risk Measurement+ (HARiM+)[105], ClinicalBARTScore[33], MedBARTScore[19], Semantic Normalized Cumulative Gain (SEM-nCG)[106], Intrinsic Knowledge Graph[107].

evaluation, especially in summarization, and relies on the longest common subsequence between the reference and generated texts. The evaluative score is computed as the fraction of words in the text that are in the longest common subsequence. Edit distance metrics[25] would also fall under this category as they are based on the number of words or characters that would need to be changed to match the reference and generated texts. Edits can be classified as insertions, deletions, substitutions, or transpositions of the words/characters in the generated text.

Embedding-based evaluations create contextualized or static embeddings for the reference and generated texts for comparison rather than relying on exact matches between words or characters. These embedding-based metrics are able to capture semantic similarities between two texts since the embedding for a word or phrase would be based on the text that surrounds it as well as itself. The BERTScore[30] is a commonly used metric that falls under this category. For this metric, a Bidirectional Encoder Representations from Transformers (BERT) model[31] is used to generate the contextualized embeddings before computing a greedy cosine similarity score based on those embeddings.

Learned metric-based evaluations rely on training a model to compute the evaluations. These metrics can be trained on example evaluation scores or directly on the reference and generated text pairs. Regression and neural network models are the foundation of these metrics providing varying degrees of complexity for the learnable parameters. The Crosslingual Optimized Metric for Evaluation of Translation (COMET)[32] is a metric that would fall under this category as it is a neural model trained for evaluation. It was originally created for evaluation of machine translations, but has since been applied to other generative tasks. COMET uses a neural network with the generated text as input to produce an evaluative score. This metric can be applied to datasets that are reference-less as well as those with reference texts.

Probability evaluations rely on calculating the likelihood of a generated text based on domain knowledge, reference texts, or source material. These metrics equate high-quality generations with those that have a high probability of being coherent or relevant to the reference or source text. They also penalize the inclusion of off-topic or unrelated information. An example is BARTScore[33], which calculates the sum of log probabilities for the generated output based on the reference text. In this case, the log probabilities are computed using the Bidirectional and Auto-Regressive Transformer (BART) model, which assesses how well the generated text aligns with the expected content[34].

Pre-Defined Knowledge Base metrics rely on established databases of domain-specific knowledge to inform the evaluation of generated text. These metrics are particularly valuable in specialized fields like healthcare, where general language models may lack the necessary depth of knowledge. By incorporating domain-specific knowledge bases, such as the National Library of Medicine's Unified Medical Language System (UMLS)[35], these metrics provide more accurate and contextually relevant evaluations. Pre-defined knowledge bases can enhance other evaluation methods, such as contextual embedding, machine learning, or probability-based metrics, by grounding them in the specialized terminology and relationships unique to the domain.

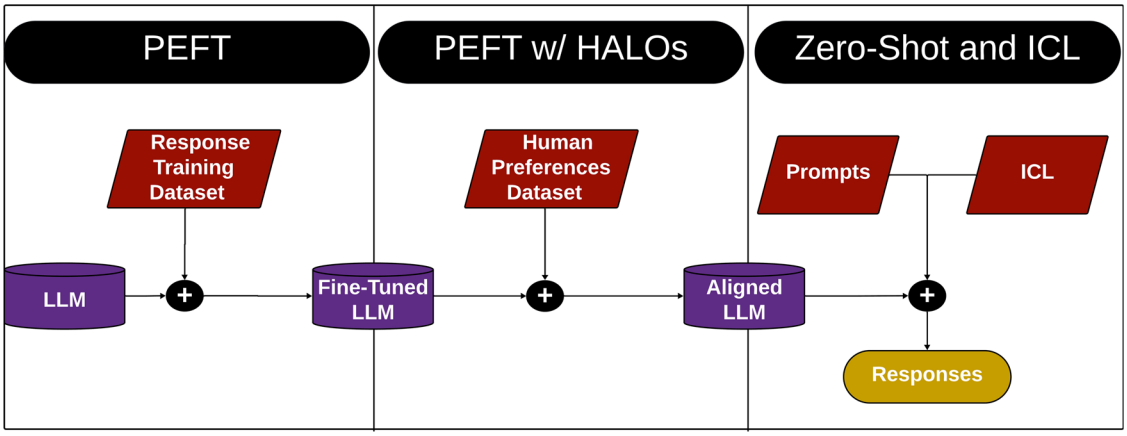

Fig. 3 | Stages of prompt engineering LLMs as judges. The three different aspects of prompt engineering expanded upon in section 5. The three sections - Zero-Shot and In-Context Learning (ICL), Parameter Efficient Fine Tuning (PEFT), and PEFT with Human Aware Loss Function (HALO) - fit together into a larger schema for training and prompting an LLM to serve as an evaluator to complement human expert evaluators.

This combination ensures that evaluations account for both linguistic accuracy and the specialized knowledge required in fields like clinical medicine. BERTScore has a variant that was trained on the UMLS called the SapBERTScore[36]. The score functions similarly to the general domain BERTScore but leverages a BERT model fine-tuned using UMLS data to generate more domain-specific embeddings. Other metrics based on the UMLS include the CUI F-Score[37] and UMLS Scorer[38]. The UMLS Scorer utilizes UMLS-based knowledge graph embeddings to assess the semantic quality of the text[19], providing a more structured approach to evaluating clinical content. Meanwhile, the CUI F-Score represents text using Concept Unique Identifiers (CUIs) from the UMLS, calculating F-scores that reflect how well the generated text aligns with key medical concepts. This enables a more granular evaluation of the relevance and accuracy of medical terminology within the generated content.

Drawbacks of automated metrics

Prior to the advent of LLMs, automated metrics would generate a single score meant to represent the quality of a generated text, regardless of its length or complexity. This single-score approach can make it difficult to pinpoint specific issues in the text, and in the case of LLMs, it is nearly impossible to understand the precise factors contributing to a particular score[13]. While automated metrics offer the benefit of speed, this comes at the cost of relying on surface-level heuristics, such as lexicographic and structural measures, that fail to capture more abstract summarization challenges in medical text. Abstractive summarization introduces unique evaluative challenges because the generated text may not directly correspond to any part of the original documentation. This contrasts with extractive summarization, where generated content is explicitly drawn from the source text, making quality assessments more straightforward. Consequently, automated metrics developed prior to the advent of LLMs are typically optimized for extractive approaches, limiting their ability to fully capture the inferences and new language generated by abstractive summarization. Furthermore, the subjective nature of assessing clinical reasoning and coherence in abstractive summaries presents additional challenges. Automated metrics often fail to account for the alignment of generated content with clinical logic or decision-making pathways, which are critical in the medical domain. This raises the importance of complementing automated metrics with strong human evaluation processes. Specifically, ensuring alignment with subject matter experts and achieving high inter-rater reliability are essential to mitigate subjectivity and provide robust evaluations.

Future directions: LLMs as evaluators to complement human expert evaluators: prompt engineering LLMs as judges

LLMs are versatile tools capable of performing a wide range of tasks, including evaluating the outputs of other LLMs (Fig. 3). This concept, where an LLM acts as a model of a human expert evaluator, has gained traction with the advent of instruction tuning and reinforcement learning with human feedback (RLHF)[39]. These advancements have significantly improved the ability of LLMs to align their outputs with human preferences, as seen in the transition from GPT-3 to GPT-4, which marked a paradigm shift in LLM accuracy and performance[40].

LLMs have the potential to bridge critical gaps in evaluation methodologies for generative clinical text. Human evaluation frameworks, while reliable, demand significant time and effort from expert reviewers, creating a paradoxical bottleneck: LLMs designed to reduce the cognitive burden on clinicians inadvertently introduce additional workload in their evaluation. Automated metrics, as they currently exist, are often insufficient for assessing the abstractive nature of generative outputs in clinical contexts. LLMs, when aligned with expert human preferences, offer an opportunity to augment evaluation processes, reducing the reliance on manual review while maintaining accuracy and relevance to clinical needs.

An effective LLM evaluator would be able to respond to evaluative questions with precision and accuracy comparable to that of human experts, following frameworks like those used in human evaluation rubrics. LLM-based evaluations could provide many of the same advantages as traditional automated metrics, such as speed and consistency, while potentially overcoming the reliance on high-quality reference texts. Moreover, LLMs could evaluate complex tasks by directly engaging with the content, bypassing the need for simplistic heuristics and offering more information into factual accuracy, hallucinations, and omissions.

Although the use of LLMs as evaluators is still emerging in research, early studies have demonstrated their utility as an alternative to human evaluations, offering a scalable solution to the limitations of manual assessment[41]. As the methodology continues to develop, LLM-based evaluations hold promise for addressing the shortcomings of both traditional automated metrics and human evaluations, particularly in complex, context-rich domains such as clinical text generation.

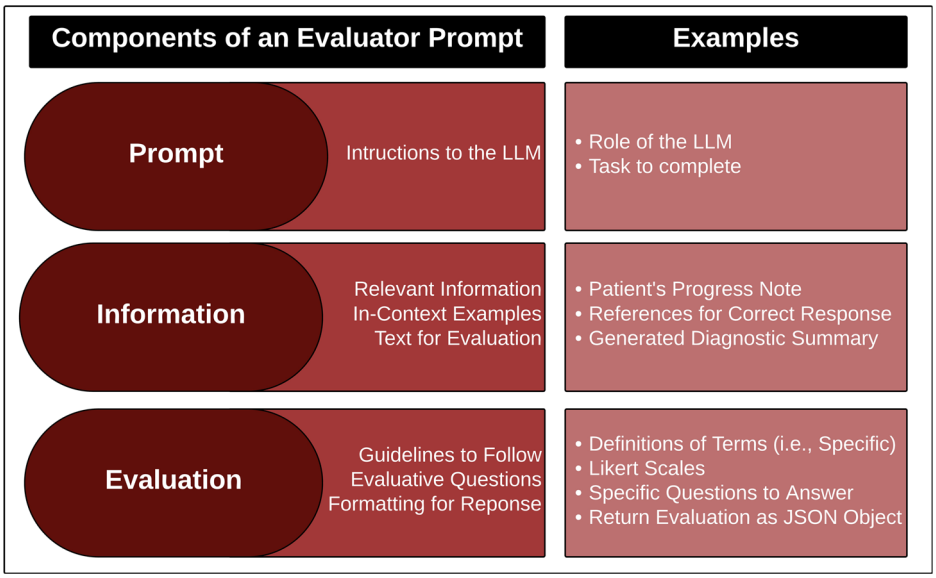

Fig. 4 | Anatomy of an evaluator prompt. An evaluator prompt consists of three sections: prompt, information, and evaluation. All three components are essential for an LLM serving as an evaluator. The Evaluator Prompt needs to instruct the LLM on the task (Prompt), provide the LLM will all the necessary information to make an evaluation (Information), and all the information that defines the guidelines and formatting of the evaluation (Evaluation).

Zero-shot and in-context learning

One method for designing LLMs to perform evaluations is through the use of manually curated prompts (Fig. 4). A prompt consists of the task description and instructions provided to an LLM to guide its responses. Two primary prompting strategies are employed in this context: Zero-Shot and Few-Shot[3]. In Zero-Shot prompting, the LLM is given only the task description without any examples before being asked to perform evaluations. Few-Shot prompting provides the task description alongside a few examples to help guide the LLM in generating output. The number of examples varies based on the LLM's architecture, input window limitations, and the point at which the model performs optimally. Typically, between one and five few-shot examples are used. Prompt engineering, through both Zero-Shot and Few-Shot ("in-context learning") approaches (collectively referred to as "hard prompting"), enables an LLM to perform tasks that it was not explicitly trained to do. However, performance can vary significantly depending on the model's pre-training and its relevance to the new task.

Beyond these manual approaches, a more adaptive strategy involves "soft prompting," also known as machine-learned prompts, which includes techniques like prompt tuning and p-tuning[42]. Soft prompts are learnable parameters added as virtual tokens to a model's input to signal task-specific instructions. Unlike hard prompts, soft prompts are trained and incorporated into the model's input layer, enabling the model to handle a broader range of specialized tasks. Soft prompting has been shown to outperform Few-Shot prompting, especially in large-scale models, as it fine-tunes the model's behavior without altering the core weights.

Through these methods, LLMs can be instructed to serve as evaluators with instructions on the dimensions and scale needed for a thorough evaluation. Promising results for such methods have already been seen in general domain applications like LLM-EVAL[43] and TALEC[44]. LLM-EVAL is a single prompt approach to employing LLM-based evaluators that can consist of multiple dimensions. This approach reported correlation coefficients between human evaluators and LLM evaluators to have an average increase of nearly 30 points over ROUGE-L. TALEC is a GPT-4 based method that incorporates in context learning for establishing evaluation criteria and has shown correlation coefficients of nearly 0.9. Even though applications in the clinical domain are significantly more complex, there have also been positive reports of LLM-based evaluators on clinically relevant tasks. Models like Llama-2, ChatGPT-4o, and Claude-3 have been applied for evaluations on medical question answering and clinical note generation. Brake et al.[45] experimented with model size, quantization, and multiple in context learning varieties on Llama-2 with final reports showing a Cohen Kappa of 0.79 with their human evaluators. Krolik et al.[46] prompted ChatGPT-4o to perform evaluations of medical responses generated for a question answering system on dimensions such as hallucinations, completeness, and coherence.

Parameter efficient fine-tuning

When prompting alone does not achieve the desired performance, fine-tuning the entire LLM may be necessary for optimal task execution. Even though an LLM may be pre-trained on a vast corpus, it can struggle with tasks requiring domain-specific knowledge or handling nuanced inputs. To address these challenges, Supervised fine-tuning (SFT) methods with Parameter Efficient Fine-Tuning (PEFT) using quantization and low rank adaptors can be employed, where the model is trained on a specialized dataset of prompt/response pairs tailored to the task at hand. Fine-tuning every weight in a LLM can require a large amount of time and computational resources. In these instances, quantization and low rank adaptors are added to the fine-tuning process for PEFT. Quantization reduces the time and memory costs of training by using lower precision data types, generally 4-bit and 8-bit, for the LLMs weights[47]. Low rank adaptors (LoRA) freeze the weights of a LLM and decompose them into a smaller number of trainable parameters ultimately also reducing the costs of SFT[48]. PEFT helps refine an LLM by embedding task-specific knowledge, ensuring the model can respond accurately in specialized contexts. The creation of these datasets is critical–performance improvements are directly tied to the quality and relevance of the prompt/response pairs used for fine-tuning. The goal is to adjust the LLM to perform better in specific use cases, such as medical diagnosis or legal reasoning, by narrowing its focus to task-specific behaviors through PEFT.

Training an LLM to serve as an evaluator could require task-specific training, especially in very specialized domains like healthcare, where the evaluation rubrics, scales, or other required definitions are part of the training dataset. PHUDGE[49] and FENCE[50] are examples of PEFT methods applied in general domain tasks for an LLM to serve as the evaluator. PHUDGE is fine-tuned from Phi-3 as a cost-efficient alternative to closed source prompting methods for models like GPT-4. Therefore, performance comparisons were done against human and GPT-4 evaluations both of which had high reported correlations with PHUDGE. FENCE is an example of a framework developed specifically for evaluating factuality. This methodology focuses on using synthetic data to augment public datasets and provide feedback to language generation models. When applied to Llama3-8b-chat, factuality was reported to see more than a 14% increase.

Extensions to traditional PEFT methods have continued to be introduced as research progresses towards specialized domains with specific

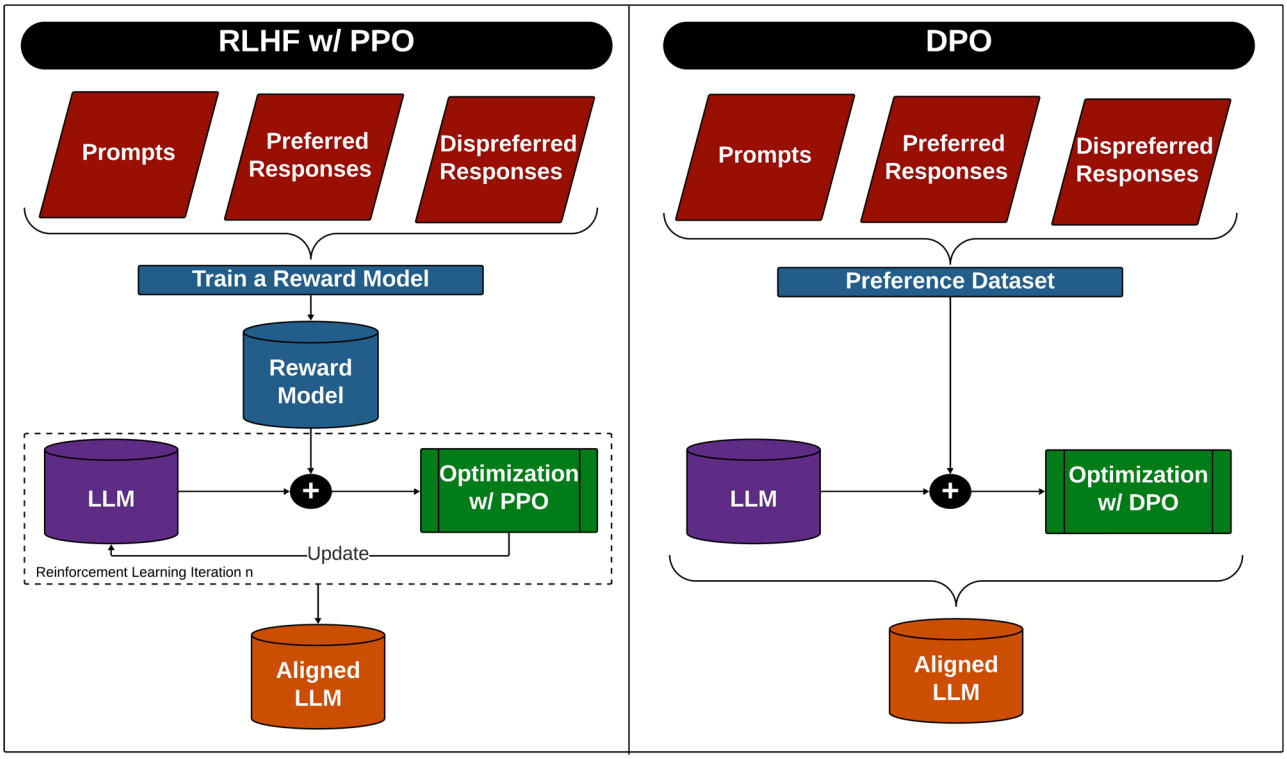

Fig. 5 | Alignment workflow: PPO v. DPO. An overview of the processes for aligning an LLM through Reinforcement Learning Human Feedback (RLHF) with Proximal Policy Optimization (PPO) and Direct Policy Optimization (DPO).

needs. Methodologies such as preference-based learning, probability calibration, text reprocessing, or some combination have emerged to refine the capabilities of LLMs[51]. Preference-based learning is focused on the adaptation of LLMs using human preference data. In this style of training, human preference datasets are curated to train LLMs for specialized evaluations. Probability calibration and text reprocessing are post-processing methodologies employed to refine LLMs through targeted adjustments following analysis of initial outputs. Probability calibration quantifies discrepancies in LLM generations and ground truth texts through mathematical derivations for adjustments. Text reprocessing hinges on integrating various iterations of evaluative outputs to improve accuracy. Because of the nuanced nature of the clinical domain, this review will focus on preference-based learning methods where clinician's preferences are incorporated in training the evaluator. This allows the LLM serving as the evaluator to be guided by feedback for understanding clinical relevancy.

Parameter efficient fine-tuning with human-aware loss function

In certain applications, the focus of fine-tuning is to align the LLM with human values and preferences, especially when the model risks generating biased, incorrect, or harmful content. This alignment, known as Human Alignment training, is driven by high-quality human feedback integrated into the training process. A widely recognized approach in this domain is Reinforcement Learning with Human Feedback (RLHF)[52]. RLHF is applied to update the LLM, guiding it toward outputs that score higher on the reward scale. In the reward model stage, a dataset annotated with human feedback is used to establish the reward, typically scalar in nature, of a particular response. The LLM is then trained to produce responses that will receive higher rewards through a process known as Proximal Policy Optimization (PPO)[53]. This iterative process ensures the model aligns with human expectations, but it can be resource-intensive, requiring significant memory, time, and computational power.

To address these computational challenges, newer paradigms have emerged that streamline Human Alignment training by directly optimizing the LLM-based on human preferences, without the need for a reward model with Direct Preference Optimization (DPO)[54]. DPO reformulates the alignment process into a human-aware loss function (HALO), optimized on a dataset of human preferences where prompts are paired with preferred and dis-preferred responses (Fig. 5). This method is particularly promising for aligning LLMs with human preferences and can be applied to ordinal responses, such as the Likert scales commonly seen in human evaluation rubrics. While PPO improves LLM performance by aligning outputs with human preferences, it is often sample-inefficient and can suffer from reward hacking[55]. DPO, in contrast, directly optimizes model outputs based on human preferences without needing an explicit reward model, making it more sample-efficient and better aligned with human values. DPO simplifies the training process by focusing directly on the desired outcomes, leading to more stable and interpretable alignment. While these methods have been successfully applied in other domains[56–58], their use in the medical field is under-explored. Training data from the human evaluation rubric on a much smaller scale to overcome labor constraints can be incorporated into a loss function designed for human alignment using DPO.

In the last year, many variants of DPO have emerged for alignment training methods that can prevent over-fitting and circumvent DPO's modeling assumptions with modifications to the underlying model and loss function (Fig. 6). Alternative methods such as Joint Preference Optimization (JPO)[59] and Simple Preference Optimization (SimPO)[60] were derived from DPO. These methods introduce regularization terms and modifications to the loss function to prevent premature convergence and ensure more robust alignment over a broader range of inputs. Other alternative methods such as Kahneman-Tversky Optimization (KTO)[61] and Pluralistic Alignment Framework (PAL)[62] use alternatives to the Bradley-Terry preferences model that underlies DPO. The alternative modeling assumptions used in these methods can prevent the breakdown of DPO's alignment in situations without direct preference data and heterogeneous human preferences.

Drawbacks of LLMs as evaluators

LLMs hold promise for automating evaluation, but as with other automated evaluation methods, there are significant challenges to consider. One major

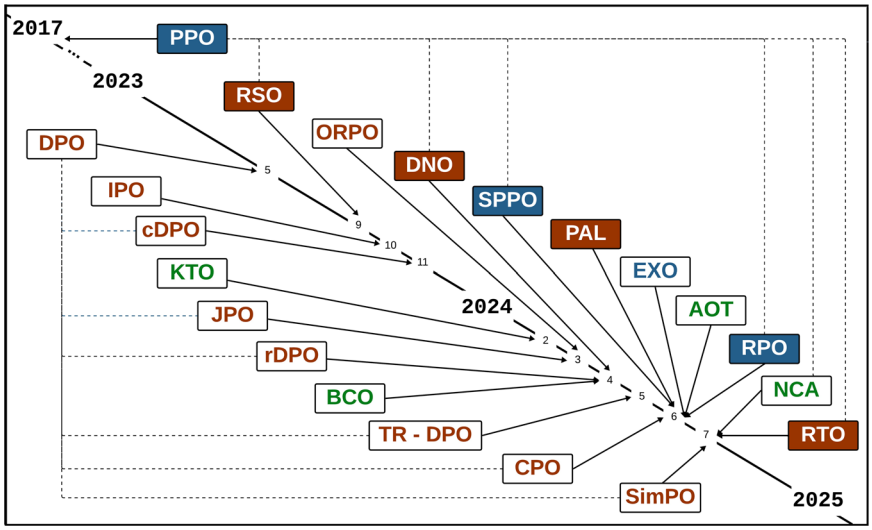

Fig. 6 | Human aware loss functions (HALOs) from PPO to present. The development timeline for HALOs from the advent of Proximal Policy Optimization (PPO) in 2017 through 2024. Each HALO is connected to it's precursor (either DPO or PPO) by a dotted line. If a HALO has an algorithmic basis in reinforcement learning, it is presented as white text on a solid color background. If a HALO has an algorithmic basis that is reinforcement learning free, it is presented as colored text on a white background. Each color, either text or background, corresponds to the data requirements for that HALO. Blue corresponds to HALOs that only use prompt/response pair data. Orange corresponds to HALOs that use response preference pairs in addition to the prompt. Finally, green corresponds to HALOs that use binary judgement data in addition to the prompt/response pair. The figure includes PPO Proximal Policy Pptimization[53], DPO Direct Preference Optimization[54], RSO Statistical Rejection Sampling[108], IPO Identity Preference Optimization[109], cDPO Conservative DPO[110], KTO Kahneman Tversky Optimization[61], JPO Joint Preference Optimization[59], ORPO Odds Ratio Preference Optimization[111], rDPO Robust DPO[112], BCO Binary Classifier Optimization[113], DNO Direct Nash Optimization[62], TR-DPO Trust Tegion DPO[114], CPO Contrastive Preference Optimization[115], SPPO Self-Play Preference Optimization[116], PAL Pluralistic Alignment Framework[62], EXO Efficient Exact Optimization[117], AOT Alignment via Optimal Transport[118], RPO Iterative Reasoning Preference Optimization[119], NCA Noise Contrastive Alignment[120], RTO Reinforced Token Optimization[121], SimPO Simple Preference Optimization[60].

issue is the rapid pace at which LLMs and their associated training strategies have evolved. This rapid development often outpaces the ability to thoroughly validate LLM-based evaluators before they are used in practice. In some cases, new optimization techniques are introduced before their predecessors have undergone peer review, and these advancements may lack sufficient mathematical justification. The speed of LLM evolution can make it difficult to allocate time and resources for proper validation, which can compromise their reliability. The specific method of validation for LLM-based evaluators is another open area of research. In the case of multiple human evaluators, inter-rater reliability metrics have been utilized to identify when different evaluators diverge. LLM-based evaluation output can be compared against that of expert human evaluators, but the standard to which LLM-based evaluators must be held has yet to be determined. In cases of ordinal, count, or other numerical evaluation scoring outputs, validation metrics like root mean squared error are also a possibility. One existing gap towards the reliable validation of LLM-based evaluators is the existence of datasets tailored for this task where the entire evaluative rubric is present and a highly-reliable ground truth exists.

Moreover, despite their advancements, LLMs remain sensitive to the prompts and inputs they receive. As LLMs continue to update and change their internal knowledge representations and as their prompts also change, the output can be highly variable. The exact LLM, or model version, that is used can also add another layer of variability. The same prompts and inputs can produce different results based on the LLM's internal structure and pretraining schema. LLMs have also been noted for egocentric bias which could affect evaluations as more and more LLM-generated text appears in source texts[63]. As a result, the use of LLMs as evaluators must be accompanied by stringent testing and safety checks to mitigate risks. Ensuring fairness in their responses is also critical, particularly in sensitive domains like healthcare, where biased or stigmatizing language could have serious consequences. These challenges highlight the need for continuous evaluation, testing, and refinement to make LLM-based evaluators both reliable and safe for medical evaluations.

Evaluation needs for the clinical domain

The development of reliable evaluation strategies is becoming increasingly important as the pace of innovation in GenAI outstrips the speed at which these technologies are validated. In health systems, the focus on clinical safety must also contend with the time constraints placed on healthcare professionals. While human evaluation rubrics offer a high degree of reliability and accuracy, they are significantly limited by the time commitment required from medical professionals serving as evaluators. Ironically, the technologies being evaluated often aim to reduce the cognitive load on these same professionals, yet they demand further time investment for their performance evaluation.

Automated evaluations, if properly designed for the clinical domain, present a promising alternative to human evaluations. However, traditional non-LLM automated evaluations have thus far fallen short, failing to consistently match the rigor of human evaluation rubrics[5,13]. These metrics frequently overlook hallucinations, fail to assess reasoning quality, and struggle to determine the relevance of generated texts. As LLMs are introduced as potential alternatives for human evaluators, it is critical to consider the unique requirements of the clinical domain. Systematic reviews of LLM-based applications for healthcare reveal that evaluation dimensions like safety, bias, and information quality are of particular importance[64,65]. Since patient safety is at the forefront of many clinical NLG tasks, clinically deployed LLMs must be evaluated for their potential to produce incorrect information or lead to negative patient outcomes. They are also susceptible to adopting biased behavior based on non-objective or non-comprehensive training data. This could infuse LLM generations with stereotypes and biased results that are harmful to patients. The quality of information in an LLM generation is also important in clinical applications. Aspects like factuality, relevancy, usefulness, consistency, and completeness are employed to capture the extent to which clinical text is representative of a patient's clinical course. These factors can be significantly more complex to evaluate using heuristics and require some level of clinical knowledge to judge clinical impact. Evaluation frameworks must incorporate assessments along these dimensions

in addition to those generally associated with high-quality text generations. These considerations require evaluation methodology designed specifically for health system applications that will prioritize such clinically relevant concerns over exact string matching or structural similarities that have been the mainstay of general domain evaluation metrics. A well-designed LLM evaluator—an "LLM-as-a-judge"—could potentially combine the high reliability of human evaluations with the efficiency of automated methods, while avoiding the pitfalls that have limited existing automated metrics. If executed effectively, such LLM-based evaluations could offer the best of both worlds, ensuring clinical safety without sacrificing the quality of assessments.

Data availability

No datasets were generated or analysed during the current study.

Received: 15 November 2024; Accepted: 31 December 2024;
Published online: 03 February 2025

References

1. Patterson, B. W. et al. Call me dr ishmael: trends in electronic health record notes available at emergency department visits and admissions. *JAMIA Open* **7**, ooae039 (2024).
2. Team, G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. http://arxiv.org/abs/2403.05530 (2024).
3. Zhao, W. X. et al. A survey of large language models. http://arxiv.org/abs/2303.18223 (2023).
4. Moramarco, F. et al. Human evaluation and correlation with automatic metrics in consultation note generation. In Muresan, S., Nakov, P. & Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5739–5754 (Association for Computational Linguistics, Dublin, Ireland, 2022). https://aclanthology.org/2022.acl-long.394.
5. Croxford, E. et al. Development of a human evaluation framework and correlation with automated metrics for natural language generation of medical diagnoses 2024.03.20.24304620. https://www.medrxiv.org/content/10.1101/2024.03.20.24304620v2 (2024).
6. Singh, H., Khanna, A., Spitzmueller, C. & Meyer, A. N. Recommendations for using the revised safer dx instrument to help measure and improve diagnostic safety. *Diagnosis* **6**, 315–323 (2019).
7. Stetson, P., Bakken, S., Wrenn, J. & Siegler, E. Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). *Appl. Clin. Inform.* **3**, 164–174 (2012).
8. Schaye, V. et al. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. *J. Gen. Intern. Med.* **37**, 507–512 (2022).
9. Kawamura, R. et al. Incidence of diagnostic errors among unexpectedly hospitalized patients using an automated medical history–taking system with a differential diagnosis generator: retrospective observational study. *JMIR Med. Inform.* **10**, e35225 (2022). Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics publisher: JMIR Publications Inc., Toronto, Canada.
10. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal.* **5**, CAT.23.0404 (2024).
11. Eshel, R. et al. Comparison of clinical note quality between an automated digital intake tool and the standard note in the emergency department. *Am. J. Emerg. Med.* **63**, 79–85 (2023).
12. Cabral, S. et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern. Med.* **184**, 581–583 (2024).
13. Sai, A., Mohankumar, A. & Khapra, M. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surveys* **55**, 1–39 (2023).
14. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
15. Otmakhova, Y., Verspoor, K., Baldwin, T. & Lau, J. H. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5098–5111 (Association for Computational Linguistics, Dublin, Ireland, 2022). https://aclanthology.org/2022.acl-long.350.
16. Adams, G., Zucker, J. & Elhadad, N. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. http://arxiv.org/abs/2303.03948 (2023).
17. Guo, Y., Qiu, W., Wang, Y. & Cohen, T. Automated lay language summarization of biomedical scientific reviews. http://arxiv.org/abs/2012.12573 (2022).
18. Wallace, B. C., Saha, S., Soboczenski, F. & Marshall, I. J. Generating (factual?) narrative summaries of rcts: experiments with neural multi-document summarization. https://arxiv.org/abs/2008.11293v2 (2020).
19. Abacha, A. B., Yim, W.-w., Michalopoulos, G. & Lin, T. An investigation of evaluation metrics for automated medical note generation. http://arxiv.org/abs/2305.17364 (2023).
20. Yadav, S., Gupta, D., Abacha, A. B. & Demner-Fushman, D. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. http://arxiv.org/abs/2107.00176 (2021).
21. Moor, M. et al. Med-flamingo: a multimodal medical few-shot learner. http://arxiv.org/abs/2307.15189 (2023).
22. Dalla Serra, F. et al. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers). Online only: Association for Computational Linguistics; 2022. p. 615–624. Available from: https://aclanthology.org/2022.aacl-main.47.
23. Cai, P. et al. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics*, 6234–6247 (International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022). https://aclanthology.org/2022.coling-1.544.
24. Umapathi, L. K., Pal, A. & Sankarasubbu, M. Med-halt: Medical domain hallucination test for large language models. http://arxiv.org/abs/2307.15343 (2023).
25. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **10**, 707 (1966).
26. Gao, Y. et al. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In Calzolari, N. et al. (eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5484–5493 (European Language Resources Association, Marseille, France, 2022). https://aclanthology.org/2022.lrec-1.587.
27. Goldsack, T., Scarton, C., Shardlow, M. & Lin, C. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K. & Tsujii, J. (eds.) *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 122–131 (Association for Computational Linguistics, Bangkok, Thailand, 2024). https://aclanthology.org/2024.bionlp-1.10.
28. Gupta, D. & Demner-Fushman, D. Overview of the medvidqa 2022 shared task on medical video question-answering. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S. & Tsujii, J. (eds.) *Proceedings of the 21st Workshop on Biomedical Language*

Processing, 264–274 (Association for Computational Linguistics, Dublin, Ireland, 2022). https://aclanthology.org/2022.bionlp-1.25.
29. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004). https://aclanthology.org/W04-1013.
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: evaluating text generation with bert. http://arxiv.org/abs/1904.09675 (2020).
31. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805 (2019).
32. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702 (Association for Computational Linguistics, Online, 2020). https://aclanthology.org/2020.emnlp-main.213.
33. Yuan, W., Neubig, G. & Liu, P. Bartscore: Evaluating generated text as text generation. http://arxiv.org/abs/2106.11520 (2021).
34. Lewis, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880 (Association for Computational Linguistics, Online, 2020). https://aclanthology.org/2020.acl-main.703.
35. Lindberg, D. A., Humphreys, B. L. & McCray, A. T. The unified medical language system. *Yearb. Med. Inf.* **1**, 41–51 (1993).
36. Liu, F., Shareghi, E., Meng, Z., Basaldella, M. & Collier, N. Self-alignment pretraining for biomedical entity representations. In Toutanova, K. et al. (eds.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4228–4238 (Association for Computational Linguistics, Online, 2021). https://aclanthology.org/2021.naacl-main.334.
37. Gao, Y. et al. Summarizing patients problems from hospital progress notes using pre-trained sequence-to-sequence models. http://arxiv.org/abs/2208.08408 (2022).
38. Delbrouck, J. Umls scorer. https://storage.googleapis.com/vilmedic_dataset/packages/medcon/UMLSScorer.zip (2023).
39. Christiano, P. et al. Deep reinforcement learning from human preferences. https://arxiv.org/abs/1706.03741v4 (2017).
40. OpenAI et al. Gpt-4 technical report. http://arxiv.org/abs/2303.08774 (2024).
41. Zheng, L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena. http://arxiv.org/abs/2306.05685 (2023).
42. Lester, B., Al-Rfou, R. & Constant, N. The power of scale for parameter-efficient prompt tuning. In Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-t. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021). https://aclanthology.org/2021.emnlp-main.243.
43. Lin, Y.-T. & Chen, Y.-N. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In Chen, Y.-N. & Rastogi, A. (eds.) *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, 47–58 (Association for Computational Linguistics, Toronto, Canada, 2023). https://aclanthology.org/2023.nlp4convai-1.5.
44. Zhang, K., Yuan, S. & Zhao, H. Talec: teach your llm to evaluate in specific domain with in-house criteria by criteria division and zero-shot plus few-shot. http://arxiv.org/abs/2407.10999 (2024).
45. Brake, N. & Schaaf, T. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency?. http://arxiv.org/abs/2404.06503. ArXiv:2404.06503 (2024).
46. Krolik, J., Mahal, H., Ahmad, F., Trivedi, G. & Saket, B. Towards leveraging large language models for automated medical q&a evaluation. http://arxiv.org/abs/2409.01941 (2024).
47. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. http://arxiv.org/abs/2305.14314 (2023).
48. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. http://arxiv.org/abs/2106.09685 (2021).
49. Deshwal, M. & Chawla, A. Phudge: Phi-3 as scalable judge. http://arxiv.org/abs/2405.08029 (2024).
50. Xie, Y. et al. Improving model factuality with fine-grained critique-based evaluator. http://arxiv.org/abs/2410.18359 (2024).
51. Li, H. et al. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. http://arxiv.org/abs/2412.05579 (2024).
52. Ziegler, D. M. et al. Fine-tuning language models from human preferences. https://arxiv.org/abs/1909.08593v2 (2019).
53. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. http://arxiv.org/abs/1707.06347 (2017).
54. Rafailov, R. et al. Direct preference optimization: Your language model is secretly a reward model. http://arxiv.org/abs/2305.18290 (2023).
55. Wen, J. et al. Language models learn to mislead humans via rlhf. http://arxiv.org/abs/2409.12822 (2024).
56. Cao, X., Xu, W., Zhao, J., Duan, Y. & Yang, X. Research on large language model for coal mine equipment maintenance based on multi-source text. *Appl. Sci.* **14**, 2946 (2024).
57. Iqbal, S. & Mehran, K. Reinforcement learning based optimal energy management of a microgrid. *2022 IEEE Energy Conversion Congress and Exposition (ECCE),* 1–8 (2022).
58. Sun, Z. et al. Improving contextual query rewrite for conversational ai agents through user-preference feedback learning. In Wang, M. & Zitouni, I. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 432–439 (Association for Computational Linguistics, Singapore, 2023). https://aclanthology.org/2023.emnlp-industry.41.
59. Bansal, H. et al. Comparing bad apples to good oranges: aligning large language models via joint preference optimization. http://arxiv.org/abs/2404.00530 (2024).
60. Meng, Y., Xia, M. & Chen, D. Simpo: simple preference optimization with a reference-free reward. http://arxiv.org/abs/2405.14734 (2024).
61. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D. & Kiela, D. Kto: model alignment as prospect theoretic optimization. http://arxiv.org/abs/2402.01306 (2024).
62. Rosset, C. et al. Direct nash optimization: teaching language models to self-improve with general preferences. http://arxiv.org/abs/2404.03715 (2024).
63. Koo, R. et al. Benchmarking cognitive biases in large language models as evaluators. http://arxiv.org/abs/2309.17012 (2024).
64. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit. Med.* **7**, 1–20 (2024).
65. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* e2421700 (2024).
66. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y. & Voss, C. (eds.) *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72 (Association for Computational Linguistics, Ann Arbor, Michigan, 2005). https://aclanthology.org/W05-0909.
67. Louis, A. & Nenkova, A. Automatically assessing machine summary content without a gold standard. *Comput. Linguist.* **39**, 267–300 (2013).

68. Vedantam, R., Zitnick, C. L. & Parikh, D. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575 (IEEE, Boston, MA, USA, 2015). http://ieeexplore.ieee.org/document/7299087/.
69. Gao, Y., Sun, C. & Passonneau, R. J. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL),* 404-418 (Association for Computation Linguistics, Hong Kong, China, 2019).
70. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318 (Association for Computational Linguistics, USA, 2002). https://doi.org/10.3115/1073083.1073135.
71. Cohan, A. & Goharian, N. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 806–813 (European Language Resources Association (ELRA), Portorož, Slovenia, 2016).
72. Lin, J. & Demner-Fushman, D. Automatically evaluating answers to definition questions. In Mooney, R., Brew, C., Chien, L.-F. & Kirchhoff, K. (eds.) *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 931–938 (Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005). https://aclanthology.org/H05-1117.
73. Hovy, E., Lin, C.-Y., Zhou, L. & Fukumoto, J. Automated summarization evaluation with basic elements. In Calzolari, N. et al. (eds.) *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (European Language Resources Association (ELRA), Genoa, Italy, 2006). http://www.lrec-conf.org/proceedings/lrec2006/pdf/438_pdf.pdf.
74. Turian, J. P., Shen, L. & Melamed, I. D. Evaluation of machine translation and its evaluation. In *Proceedings of Machine Translation Summit IX: Papers* (New Orleans, USA, 2003). https://aclanthology.org/2003.mtsummit-papers.51.
75. Su, K.-Y., Wu, M.-W. & Chang, J.-S. A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. https://aclanthology.org/C92-2067 (1992).
76. Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231 (Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 2006). https://aclanthology.org/2006.amta-papers.25.
77. Panja, J. & Naskar, S. K. Iter: Improving translation edit rate through optimizable edit costs. In Bojar, O. et al. (eds.) *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 746–750 (Association for Computational Linguistics, Belgium, Brussels, 2018). https://aclanthology.org/W18-6455.
78. Leusch, G., Ueffing, N. & Ney, H. Cder: Efficient mt evaluation using block movements. In McCarthy, D. & Wintner, S. (eds.) *11th Conference of the European Chapter of the Association for Computational Linguistics*, 241–248 (Association for Computational Linguistics, Trento, Italy, 2006). https://aclanthology.org/E06-1031.
79. Popović, M. chrf: character n-gram f-score for automatic mt evaluation. In Bojar, O. et al. (eds.) *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395 (Association for Computational Linguistics, Lisbon, Portugal, 2015). https://aclanthology.org/W15-3049.
80. Wang, W., Peter, J.-T., Rosendahl, H. & Ney, H. Character: Translation edit rate on character level. In Bojar, O. et al. (eds.) *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 505–510 (Association for Computational Linguistics, Berlin, Germany, 2016). https://aclanthology.org/W16-2342.
81. Stanchev, P., Wang, W. & Ney, H. Eed: Extended edit distance measure for machine translation. In Bojar, O. et al. (eds.) *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 514–520 (Association for Computational Linguistics, Florence, Italy, 2019). https://aclanthology.org/W19-5359.
82. Lo, C.-k. Yisi - a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In Bojar, O. et al. (eds.) *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 507–513 (Association for Computational Linguistics, Florence, Italy, 2019). https://aclanthology.org/W19-5358.
83. Nema, P. & Khapra, M. M. Towards a better metric for evaluating question generation systems. In Riloff, E., Chiang, D., Hockenmaier, J. & Tsujii, J. (eds.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3950–3959 (Association for Computational Linguistics, Brussels, Belgium, 2018). https://aclanthology.org/D18-1429.
84. Sellam, T., Das, D. & Parikh, A. P. Bleurt: Learning robust metrics for text generation. http://arxiv.org/abs/2004.04696 (2020).
85. Lin, Z., Liu, C., Ng, H. T. & Kan, M.-Y. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G. & Park, J. C. (eds.) *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1006–1014 (Association for Computational Linguistics, Jeju Island, Korea, 2012). https://aclanthology.org/P12-1106.
86. Stanojević, M., Simaán, K. Fitting Sentence Level Translation Evaluation with Many Dense Features. In: Moschitti A, Pang B, Daelemans W, editors. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 202–206. Available from: https://aclanthology.org/D14-1025.
87. Ma, Q., Graham, Y., Wang, S. & Liu, Q. Blend: a novel combined mt metric based on direct assessment — casict-dcu submission to wmt17 metrics task. In Bojar, O. et al. (eds.) *Proceedings of the Second Conference on Machine Translation*, 598–603 (Association for Computational Linguistics, Copenhagen, Denmark, 2017). https://aclanthology.org/W17-4768.
88. Sharif, N., White, L., Bennamoun, M. & Ali Shah, S. A. Learning-based composite metrics for improved caption evaluation. In Shwartz, V. et al. (eds.) *Proceedings of ACL 2018, Student Research Workshop*, 14–20 (Association for Computational Linguistics, Melbourne, Australia, 2018). https://aclanthology.org/P18-3003.
89. Chen, Q. et al. Enhanced lstm for natural language inference. In Barzilay, R. & Kan, M.-Y. (eds.) *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1657–1668 (Association for Computational Linguistics, Vancouver, Canada, 2017). https://aclanthology.org/P17-1152.
90. Shimanaka, H., Kajiwara, T. & Komachi, M. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In Bojar, O. et al. (eds.) *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 751–758 (Association for Computational Linguistics, Belgium, Brussels, 2018). https://aclanthology.org/W18-6456.
91. Shimanaka, H., Kajiwara, T. & Komachi, M. Machine translation evaluation with bert regressor. http://arxiv.org/abs/1907.12679 (2019).
92. Zhang, S. et al. Conditional bilingual mutual information based adaptive training for neural machine translation. In Muresan, S.,

Nakov, P. & Villavicencio, A. (eds.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2377–2389 (Association for Computational Linguistics, Dublin, Ireland, 2022). https://aclanthology.org/2022.acl-long.169.
93. Doddington, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* -, 138 (Association for Computational Linguistics, San Diego, California, 2002). http://portal.acm.org/citation.cfm?doid=1289189.1289273.
94. Zhao, W. et al. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In Inui, K., Jiang, J., Ng, V. & Wan, X. (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 563–578 (Association for Computational Linguistics, Hong Kong, China, 2019). https://aclanthology.org/D19-1053.
95. Giannakopoulos, G. & Karkaletsis, V. Autosummeng and memog in evaluating guided summaries. *Theory and Applications of Categories* (2011).
96. Anderson, P., Fernando, B., Johnson, M. & Gould, S. Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N. & Welling, M. (eds.) *Computer Vision – ECCV 2016*, 382–398 (Springer International Publishing, Cham, 2016).
97. Mathur, N., Baldwin, T. & Cohn, T. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In Korhonen, A., Traum, D. & Màrquez, L. (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2799–2808 (Association for Computational Linguistics, Florence, Italy, 2019). https://aclanthology.org/P19-1269.
98. Echizen'ya, H., Araki, K. & Hovy, E. Word embedding-based automatic mt evaluation metric using word position information. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1874–1883 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). https://aclanthology.org/N19-1186.
99. Kusner, M., Sun, Y., Kolkin, N. & Weinberger, K. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, 957–966 (PMLR, 2015). https://proceedings.mlr.press/v37/kusnerb15.html.
100. Wieting, J., Berg-Kirkpatrick, T., Gimpel, K. & Neubig, G. Beyond bleu: Training neural machine translation with semantic similarity. In Korhonen, A., Traum, D. & Màrquez, L. (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4344–4355 (Association for Computational Linguistics, Florence, Italy, 2019). https://aclanthology.org/P19-1427.
101. Kane, H., Kocyigit, M. Y., Abdalla, A., Ajanoh, P. & Coulibali, M. Nubia: Neural based interchangeability assessor for text generation. http://arxiv.org/abs/2004.14667 (2020).
102. Liu, F., Shareghi, E., Meng, Z., Basaldella, M. & Collier, N. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4228–4238 (Association for Computational Linguistics, Online, 2021). https://www.aclweb.org/anthology/2021.naacl-main.334.
103. Alsentzer, E. et al. Publicly available clinical BERT embeddings. *CoRR*abs/1904.03323. http://arxiv.org/abs/1904.03323, 1904.03323 (2019).
104. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. (2020).
105. Son, S. et al. Harim+: evaluating summary quality with hallucination risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 895–924 (Association for Computational Linguistics, Online, 2022).
106. Akter, M., Bansal, N. & Karmaker, S. K. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022*, 1547–1560 (Association for Computational Linguistics, Dublin, Ireland, 2022). https://aclanthology.org/2022.findings-acl.122.
107. Aracena, C., Villena, F., Rojas, M. & Dunstan, J. A knowledge-graph-based intrinsic test for benchmarking medical concept embeddings and pretrained language models (2022).
108. Liu, T. et al. Statistical rejection sampling improves preference optimization. http://arxiv.org/abs/2309.06657 (2024).
109. Azar, M. G. et al. A general theoretical paradigm to understand learning from human preferences. http://arxiv.org/abs/2310.12036 (2023).
110. Mitchell, E. A note on dpo with noisy preferences and relationship to ipo. https://ericmitchell.ai/cdpo.pdf (2023).
111. Hong, J., Lee, N. & Thorne, J. Orpo: monolithic preference optimization without reference model. http://arxiv.org/abs/2403.07691 (2024).
112. Chowdhury, S. R., Kini, A. & Natarajan, N. Provably robust dpo: aligning language models with noisy feedback. http://arxiv.org/abs/2403.00409 (2024).
113. Jung, S., Han, G., Nam, D. W. & On, K.-W. Binary classifier optimization for large language model alignment. http://arxiv.org/abs/2404.04656 (2024).
114. Gorbatovski, A. et al. Learn your reference model for real good alignment. http://arxiv.org/abs/2404.09656 (2024).
115. Xu, H. et al. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. http://arxiv.org/abs/2401.08417 (2024).
116. Wu, Y. et al. Self-play preference optimization for language model alignment. http://arxiv.org/abs/2405.00675 (2024).
117. Ji, H. et al. Towards efficient exact optimization of language model alignment. http://arxiv.org/abs/2402.00856 (2024).
118. Melnyk, I. et al. Distributional preference alignment of llms via optimal transport. http://arxiv.org/abs/2406.05882 (2024).
119. Pang, R. Y. et al. Iterative reasoning preference optimization. http://arxiv.org/abs/2404.19733 (2024).
120. Chen, H. et al. Noise contrastive alignment of language models with explicit rewards. http://arxiv.org/abs/2402.05369 (2024).
121. Zhong, H. et al. Dpo meets ppo: Reinforced token optimization for rlhf. http://arxiv.org/abs/2404.18922 (2024).

Acknowledgements

We thank Anne Glorioso, Leslie Christensen, and Paije Wilson for their assistance with database selection and search query assistance as UW-Madison subject librarians.

Author contributions

E.C.: conceptualization, data curation, methodology, investigation, visualization, writing—original draft, writing—review & editing. Y.G.: conceptualization, data curation, methodology, investigation, supervision, validation, visualization, writing—review & editing. N.P.: conceptualization, investigation, supervision, validation, writing—review & editing. K.W.: conceptualization, investigation, supervision, validation, writing—review & editing. G.W.: conceptualization, investigation, supervision, validation,

writing—review & editing. E.F.: conceptualization, investigation, supervision, validation, writing—review & editing. F.L.: supervision, validation, writing—review & editing. C.G.: supervision, validation, writing—review & editing. B.P.: conceptualization, data curation, methodology, investigation, supervision, validation, writing—review & editing. M.A.: conceptualization, data curation, methodology, funding acquisition, investigation, supervision, validation, writing—review & editing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44401-024-00011-2.

Correspondence and requests for materials should be addressed to Majid Afshar.