Combining Multiple Metrics for Evaluating Retrieval-Augmented Conversations

Jason Ingyu Choi **Marcus D. Collins** **Eugene Agichtein**
Oleg Rokhlenko and **Shervin Malmasi**
Amazon.com, Inc. Seattle, WA, USA
{chojson,collmr,eugeneag,olegro,malmasi}@amazon.com

Abstract

Conversational AI is a subtype of Human-Computer Interaction that has gained wide adoption. These systems are typically powered by Large Language Models (LLMs) that use Retrieval Augmented Generation (RAG) to infuse external knowledge, which is effective against issues like hallucination. However, automatically evaluating retrieval augmented conversations with minimal human effort remains challenging, particularly in online settings. We address this challenge by proposing a lexical metric, and a novel method for combining it with other metrics, including semantic models. Our approach involves: (1) Conversational Information Utility (CIU), a new automated metric inspired by prior user studies on web search evaluation, to compute information overlap between conversation context and grounded information in an unsupervised, purely lexical way; and (2) a generalized reward model through Mixture-of-Experts (MoE-CIU) that dynamically ensembles CIU with other metrics, including learned ones, into a single reward. Evaluation against human ratings on two public datasets (Topical Chat and Persona Chat) shows that CIU improves correlation against human judgments by 2.0% and 0.9% respectively compared to the second best metric. When MoE is applied to combine lexical and learned semantic metrics, correlations further improve by 9.9% and 5.0%, suggesting that unified reward models are a promising approach.

1 Introduction

Conversational AI is a specific type of Human-Computer Interaction that has been widely studied in recent years (Ouyang et al., 2022; Team et al., 2023), leading to the development of multi-purpose chat assistants (*e.g.* ChatGPT, Claude) based on Large Language Models (LLMs). However, as more customers interact with such assistants, addressing limitations like hallucination, factual consistency, prompt brittleness and controllability has gained more attention (Kaddour et al., 2023). One widely-adopted solution is Retrieval Augmented Generation (RAG), which allows choosing a context document ($d_{context}$) to ground LLM responses, and increase truthfulness with respect to the source document (Lewis et al., 2020).

Our work focuses on the task of automatically assessing the quality of retrieval-augmented responses in knowledge-grounded conversations. By examining both the context and the response, we estimate the degree to which the retrieved document was used in generation, in order to identify uninformative or inconsistent responses. Our approach is designed for real-time use, where using a large model may be infeasible. Compared to offline tasks, online evaluation (e.g., live monitoring of defects) requires efficient solutions. Recent work utilizes LLMs, either through prompt engineering or fine-tuning, to automatically predict evaluation metrics and reduce dependency from human annotators (Thapa et al., 2023; Chan et al., 2023). Despite demonstrated potentials, a large number of parameters, high latency, and potential legal issues significantly limits deploying LLM-based solutions for live traffic monitoring. As an alternative, we propose an approach that combines much simpler and scalable metrics to predict user ratings, or potentially other business metrics. Our approach can also support offline evaluations, and is relevant to recent trends in Reinforcement Learning from Human Feedback (RLHF), which aligns LLM responses toward human preferences (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2024).

Early attempts on automatic dialog evaluation relied on existing metrics (*e.g.* BLEU) from machine translation literature to evaluate assistant conversations against 'gold' conversations (Papineni et al., 2002). However, defining the full space of 'gold' conversations is infeasible due to the non-deterministic nature of dialogs and many existing works simply penalize any response that slightly de-

viates from 'gold' (Chen et al., 2019). On the other hand, learned, semantic conversational quality metrics trained on labeled data tend to show higher correlations against human judgements than exact word-overlap metrics, because word embedding-based approaches can compute overlap in a 'soft' way that accounts for lexical variation, *e.g.* , (Sellam et al., 2020; Lowe et al., 2017a). However, such 'soft' approaches also suffer from different issues, such as over-fitting, performance degradation on longer inputs, or learning similar representations for antonyms.

To address these gaps, we ask: *is it feasible to unify multiple independent metrics into a single reward?* To answer this question, we investigated two research questions:

Q1 What is the most effective and robust standalone metric (whether lexical or learned) that aligns with human ratings from open-source knowledge-grounded conversations?

Q2 Given a set of independent metrics, how much improvement does a unified reward model gain compared to the best standalone metric?

For **Q1**, previous studies on 'exact' word overlap metrics showed they correlate poorly with human preferences, such as question answering accuracy (Chen et al., 2019) and response appropriateness (Lowe et al., 2017b). To address this, we compare our newly proposed lexical metric, **Conversational Information Utility (CIU)**, which is inspired from user-centric studies on web search evaluation (Azzopardi et al., 2018; Moffat et al., 2013) against existing metrics. A key insight is that for the user to gain useful information, they must ask a series of questions, or make statements, that cause the conversational system (or human partner) to respond with information overlapping with $d_{context}$. Our main novelty is how CIU quantifies information overlap to reward relevancy, information novelty and conciseness while penalizing repetitive information and high user effort. Experiments validate that CIU improves correlation against human ratings by 2.0% and 0.9% against the second best metric on Topical Chat (Gopalakrishnan et al., 2019) and Persona Chat (Zhang et al., 2018) datasets for predicting Overall Ratings.

For **Q2**, we experiment with different ensemble learning strategies to (1) validate whether previously identified strong metrics are considered as strong predictors (metrics); (2) demonstrate the superiority of an unified reward compared to any standalone metric. Experimental results in feature selection ratio show that CIU is selected 76.4% across 17 different feature selection approaches, which justifies our findings on **Q1**. When Mixture-of-Experts (MoE) (Masoudnia and Ebrahimpour, 2014) was applied, the resulting MoE-CIU model further improved correlation with human ratings by 9.9% and 5.0% on Topical Chat and Persona Chat compared to the best standalone metric. In summary, our contributions are:

- A simple and effective lexical metric for estimating Conversational Information Utility (CIU) within information-seeking retrieval augmented conversations
- A generalized, domain-agnostic model MoE-CIU that utilizes Mixture of Experts to dynamically adjust metric weights of different modalities into an unified reward signal

2 Related Work

Web Search Evaluation and Utility For search engine evaluation, evaluation measures evolved from precision- and recall-based to utility- and cost-based with more emphasis on interactions between users and search results (Moffat et al., 2013). This is because simply measuring how well search engine ranks relevant documents does not always translate to increased user satisfaction. To model interactions, additional information such as likelihood of user continuing or stopping after at a given rank or estimated effort to read each document (Zhang et al., 2017; Sakai and Dou, 2013) is considered when defining a utility (Wicaksono and Moffat, 2020). Overall, web-search utility is an aggregated metric that combines precision and recall of ranked documents with user interaction signals derived from search logs.

However, the main challenge is on applying these intuitions to multi-turn conversations. In conversational settings, many existing word-overlap and learned metrics (Papineni et al., 2002; Tao et al., 2017; Zhang et al., 2019) still rely on word overlap or semantic similarity to evaluate responses while neglecting potential user interactions. An ideal utility should holistically consider word-level precision, semantic relatedness, novelty of information, repetition, conversational history and user effort to evaluate conversations.

Learned Metrics One popular approach is to utilize pretrained contextual embeddings from Transformer models to compute a similarity score between two texts. For example, BERTScore (Zhang et al., 2019) computes a token-level similarity matrix and re-weights the scores based on IDF scores to boost signals from more novel matches. ADEM (Lowe et al., 2017b) uses a hierarchical RNN encoder to predict human-annotated ratings on Twitter data. While ADEM requires human judgments, RUBER-BERT (Ghazarian et al., 2019; Tao et al., 2017) uses an unsupervised negative sampling strategy to train a model that measures information relatedness between query and response. USR (Mehri and Eskenazi, 2020) is another transformer-based model that is shown effective for evaluating model generated responses. To evaluate USR, the authors sampled a small number of conversations from Topical Chat and Persona Chat datasets to annotate several useful ratings (e.g. overall rating of responses), which we adopt for our evaluations.

Mixture-of-Experts Models When there are multiple representations of the same input, *e.g.* complementary representations computed by different expert modules, the Mixture-of-Experts (MoE) approach (Masoudnia and Ebrahimpour, 2014) can take these independent knowledge sources and conditionally combine them into a joint representation (Shazeer et al., 2017). It does so by training a gating mechanism that dynamically assigns weights to the experts, depending on the input (Jain et al., 2019). MoE has been shown to be effective in various settings such as combining Support Vector Machines (Collobert et al., 2002), hierarchical networks (Yao et al., 2009) and Named Entity Recognition (Meng et al., 2021). For our use case, each expert is a representation of different metrics. By training a gating network to dynamically weight experts per instance, we expect MoE to improve over heuristic-based feature combination strategies, such as sum or mean of different metrics (Ghazarian et al., 2019).

3 Proposed Metrics and Models

We define our task and usefulness ratings, followed by details on CIU and MoE-CIU approaches.

3.1 Usefulness Rating Prediction Task

Given a conversation history (C) and a specific utterance at turn i (utt_i), our task is to predict how much useful information (Rosset et al., 2020) is present at utt_i, with respect to the retrieved knowledge ($d_{context}$). Usefulness ratings (1.0 - 5.0) measures whether the response helps towards fulfilling the information needs (*i.e.* learning new information, or asking questions about products in online shopping). To be useful, utterances should meet the information needs of users and drive the conversation forward to elicit more interaction, while staying relevant to C and $d_{context}$.

3.2 CIU: Conversational Information Utility

Utility is defined as the fulfillment a user receives after search, and attempts to model how users aim to gain optimal overall satisfaction (Machmouchi et al., 2017). We hypothesize that successful information-seeking conversations deliver useful and factually correct information from $d_{context}$. CIU is specifically designed to rank responses with respect to salient information overlap in an unsupervised, lexical way. To measure information overlap, we utilize Rapid Automatic Keyword Extraction (RAKE) (Rose et al., 2010), an existing algorithm that extracts and ranks keywords based on word co-occurrences. Highly ranked phrases from $d_{context}$ are then matched against tokens from each turn, combined with multiple token-level discounting criteria, which are discussed in next sections.

We define information overlap as the sum of all token-level match $score$ between utt_i and each sentence d_{sent} from $d_{context}$. In case $d_{context}$ is long and contains many paragraphs irrelevant to utt_i, we limit d_{sent} to only those from the most relevant paragraph if such annotations are available. However for other datasets without annotations, we apply RAKE to $d_{context}$ to extract highly relevant phrases. For simplicity, we use the same notation d_{sent} for extracted phrases.

CIU is a normalized, discounted information overlap. For each turn (i), it is calculated as:

$$CIU_i = [\sum_{token}^{d_{sent}} \frac{score(token, utt_i) \cdot \gamma}{freq(token)}] - E_i. \quad (1)$$

CIU accepts any scoring function $score$ which outputs a relevance score between each token in d_{sent} and utt_i. Here, we use a binary function that outputs 1 if each token from d_{sent} appears in utt_i and 0 otherwise. Although binary scoring seems rudimentary, preliminary experiments showed that embedding-based token similarity scores were noisy and did not generalize well

across diverse samples. Another reason to favor binary scores is that CIU is meant to be purely lexical and efficient; embedding-based similarity can significantly slow down predictions as token-level comparisons are expensive. Eq. 1 also has several discounting terms: γ is position-based, $freq$ is word frequency-based, and E_i is an effort discounting term that subtracts time required to read utt_i. Next, we describe the discounting terms.

Position-based discounting Each $score$ is discounted based on which position token appears in utt_i. For example, we boost weights of $score$ that appears earlier in utt_i. This was inspired from earlier work (Sakai and Dou, 2013), which claims the value of relevant information decays based on how much user effort is required to process information. This is particularly effective for longer utt_i since users prefer useful information to appear earlier than later. We adopt the linear discounting proposed from same earlier work:

$$\gamma = max(0, 1 - pos(token)/|utt_i|), \quad (2)$$

where $pos(token)$ is the token index, and $|utt_i|$ is the number of tokens in utt_i.

Frequency-based discounting Without frequency discounting, all tokens are treated equally regardless of how frequent or novel they are. Prior work (Qi et al., 2020) computes informativeness as how many unseen tokens from information units overlap with an answer, measured with a unigram precision function. However, this assumes all repetitive information is irrelevant. Ideally, our utility function should assign smaller weights to frequently observed tokens and higher weights to novel tokens. The simplest way of achieving this is to divide each token $score$ by token's term frequency (TF), which is measured and updated throughout the conversation.

User effort and cost Several methods of evaluating search engines have considered the trade-off between user effort (E) and relevance gain (Zhang et al., 2017; Azzopardi et al., 2018). In conversational settings, we hypothesize that turn-level effort (E_i) can be approximated by computing the total time a user has spent each turn to read a response. To understand how E_i influences user satisfaction, we analyzed turn-level human annotations from Topical Chat and Persona Chat corpus. According to Figure 1, we first observed that users are less likely to rate longer utterances as useful than shorter utterances. To quantify this relationship, we computed Spearman correlation between usefulness ratings and character length. There is a statistically significant negative correlation of -0.203 ($p < 0.001$), justifying the need for a length-based effort discounting.

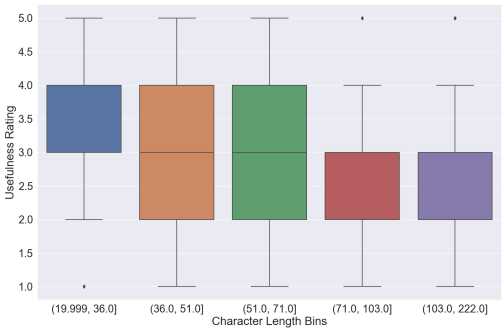

Figure 1: A box plot showing the distribution of utterance length, divided into five equal-sized bins, over human-annotated usefulness ratings from Topical Chat and Persona Chat.

The simplest way of penalizing longer utterances is assuming a constant cost (reading speed) per character, C_{char}.

$$E_i = C_{char} * |utt_i| \quad (3)$$

Then, E required per turn is subtracted from CIU_i to favor shorter turns with identical information as shown in Equation 1. Note that this value can be tuned for different datasets if human annotations are available. Otherwise, we propose to reuse our value 0.005, which was tuned against open-source dialogs with human annotations.

3.3 MoE-CIU: Mixture-of-Experts with CIU

Next, we focus on methodologies to model an unified reward from independent metrics. Perhaps the simplest way of leveraging multiple metrics is computing the average or sum. However, previous work (Ghazarian et al., 2019) highlighted that simple arithmetic operations on raw metric values degrades performance since each metric captures orthogonal measures of relatedness or utility in different scales. Hence, we propose an MoE approach to dynamically normalize and combine the metrics with weights that vary over the input.

MoE requires vectors as inputs rather than scalar scores. Instead of using raw metric scores, we categorize metric scores to one of N bins based on

each score distribution. For bin size, we found that $N = 5$ gives the best performance. Then, we create trainable embedding layers of dimensionality (M, N, D) where M is the number of metrics to combine, N is the number of bins, and D is a hyperparameter that defines the embedding size. For our experiments, we used $(6, 5, 16)$.

Two different feature combination strategies were explored. First, we trained a baseline MoE-Concat that uses a concatenated vector of label representations to predict scores. The second model MoE-CIU uses the MoE gating mechanism (Meng et al., 2021) to learn input-dependent weights for each metric to compute a final score. Both approaches can be trained as a binary classifier (with cross-entropy loss), or as a regressor (with mean-squared error loss) depending on use cases.

4 Public Datasets with Human Ratings

We used two publicly available dialog corpora to evaluate our approach.

Topical Chat Topical Chat (Gopalakrishnan et al., 2019), contains 11k conversations that are grounded to web articles ($d_{context}$). For human ratings, we use a subset of more recent and reliable ratings obtained on the official test split (Mehri and Eskenazi, 2020), which contains 360 samples. From multiple available ratings, we selected two ratings that are most relevant for our scope:

- Overall Rating (0 - 5): What is the overall impression of this utterance based on understandability, naturalness, coherency, interestingness, and relevancy of used knowledge?
- Uses Knowledge (0 - 1): Given the fact that the response is conditioned on ($d_{context}$), how well does the response use that fact?

Persona Chat Persona chat (Zhang et al., 2018) is another popular dataset that contains knowledge-grounded conversations on different personas ($d_{context}$). Similar to Topical Chat, Mehri and Eskenazi (2020) released a more recent human annotations on 300 test samples. For consistency, we will choose the exact same type of ratings for our evaluation.

5 Experimental Settings

We first present an overview of our selected baseline metrics and evaluation criteria. For **Q1**, each metric performance is (1) evaluated independently. Then for **Q2**, we experiment with two settings for metric unification: (2) static combination where optimal features (metrics) are pre-selected from existing feature selection algorithms; (3) dynamic combination that utilizes MoE to automatically learn and weight all input features.

5.1 Standalone Metrics Evaluation

We compare CIU performances against standard word-overlap and learned baseline metrics (1) Random Baseline (2) BLEU; (3) METEOR; (4) ROUGE-L; (5) RUBER-BERT; (6) BERTScore (Banerjee and Lavie, 2005; Lin, 2004). Random baseline is added to highlight the relative difficulty of different tasks.

To evaluate the effectiveness of individual metrics, we compute Spearman rank correlation between our metric predictions and two different types of human ratings: (1) 'Overall' ratings; (2) 'Uses Knowledge' ratings, as discussed in Section (§4). Spearman correlation was chosen over Pearson because Spearman is more suited for benchmarking monotonic relationship while Pearson only models linear relationships.

Ablation Analysis on CIU parameters To evaluate individual contributions of different discounting terms within CIU, we include an ablation analysis that systematically removes each discounting terms on Table 1 and Table 3. Effort terms were tuned on the Topical Chat training corpus, and used $C_{char} = 0.005$ for other experiments.

5.2 Unified Reward Evaluation

Static Combination with Feature Selection We experiment with existing feature selection strategies to first identify strong predictors, and second train a model to ensemble strong estimators for predicting human ratings. The evaluation criteria we adopt is the feature selection ratio, which computes how many times each metric is identified as a top-k predictor against others. We experimented with following feature selection[1] strategies:

- Univariate feature selection
- Feature selection using random forest
- Recursive feature elimination
- Forward & backward feature selection
- No feature selection, uses all features

[1] https://scikit-learn.org/stable/modules/feature_selection.html

Since training ensemble models requires labels, we reserved 50 random samples each from both datasets for testing and remainders for training. Support Vector Regressor (SVR) was chosen because this model achieved the strongest performance on multiple experiments over other choices (e.g., gradient-boosted decision tree). The optimal hyperparameters were identified using Grid Search. Since we have only 50 test samples, experiments were repeated 15 times using different sampling seeds and performances were averaged to reduce variance. For consistency, we also report Spearman correlation against human ratings.

Dynamic Combination with Mixture of Experts We compare the static feature selection models to MoE-Concat and MoE-CIU (discussed in Section 3.3), which does not require any feature selection in theory because Mixture of Experts are designed to automatically learn and combine different metric representations. Hence by default, these models take all features as inputs. Performances are also averaged over 15 different sampling seeds.

6 Main Results

We present the results on each public dataset, followed by ablation study and error analysis.

6.1 Topical Chat Results

Standalone Metric Performance Table 1 lists the correlation of different metrics in a standalone setting. Overall, the best CIU configuration that uses all proposed discounting terms achieved the highest correlation for predicting Overall Ratings, which answers **Q1**. It is impressive that CIU was able to outperform learned metrics without any training data. However for predicting Uses Knowledge, a learned metric (RUBER-BERT) outperformed CIU by 3.6%. All of correlation coefficients reported in Table 1, including the difference between CIU and the second best lexical metric (METEOR) are statistically significant ($p < 0.001$).

Unified Reward Performance Table 2 shows the results from different feature selection strategies. The best combination strategy for predicting Overall Ratings was to simply use all metrics. This achieved 0.432 Spearman correlation, a +1.7% improvement over the best standalone metric, CIU. All correlations reported in Table 2 are statistically significant ($p = 0.001$). For predicting Uses Knowledge, the best 'Recursive-5' model excluded ROUGE as the weakest feature, achieving 0.781 correlation. Generally, there is a clear trend that correlation improves with more features. This is a strong evidence showing that leveraging multiple metrics is more effective than any single metric alone.

Metric	Overall Ratings	Uses Knowledge
Random Guessing	0.016	0.023
BLEU	0.298	0.631
METEOR	0.352	0.716
ROUGE-L	0.339	0.688
RUBER-BERT	0.385	**0.778**
BERTScore	0.395	0.717
CIU - $freq$	0.411	0.728
CIU - pos	0.412	0.729
CIU	**0.415**	0.742

Table 1: Spearman correlation between metric predictions and human ratings on Topical Chat. Ablation study is indicated with minus sign where $freq$ stands for frequency and pos for position.

For all 17 different feature selection strategies we note that CIU, RUBER-BERT and BERTScore were almost always selected. They were also the top metrics on Topical Chat (Table 1). For backward selection (which outperform forward selection), we see that 'Backward-1' first picks up RUBER-BERT as the most useful feature, followed by CIU and BERTScore. Although CIU was best in predicting Overall Ratings, other feature selection strategies did not always prioritize CIU on first iterations. Nonetheless, these results demonstrate that the majority of feature selection strategies consider CIU and RUBER-BERT as one of the strongest features, which justifies our findings on **Q1**.

Having validating the effectiveness of combining multiple metrics, we trained the MoE-Concat and MoE-CIU models on the same data splits. To answer **Q2**, MoE-CIU achieved 0.514 correlation for Overall Ratings (+8.2% improvement), and 0.799 correlation for Uses Knowledge (+1.8% improvement) against the best static combination approach, both of which are statistically significant. Accordingly, we claim that MoE-based approaches are superior to traditional feature selection strategies as the MoE gating mechanism can dynamically adjust the weights of different metrics while feature selection is binary and static (features are either used or not, and have a fixed weight).

Metric	Overall Ratings	Uses Knowledge	BLEU	METEOR	ROUGE-L	CIU	RUBER-BERT	BERTScore
Univariate	0.427	0.773	✓	✓	✓	✓	-	✓
Random Forest	0.405	0.721	✓	✓	-	✓	✓	✓
Recursive-5	0.385	0.734	-	✓	-	-	-	-
Recursive-4	0.392	0.733	-	✓	-	-	-	✓
Recursive-3	0.396	0.767	-	✓	-	-	✓	✓
Recursive-2	0.411	0.781	-	✓	-	✓	✓	✓
Recursive-1	0.422	**0.781**	✓	✓	-	✓	✓	✓
Forward-1	0.361	0.729	-	-	✓	-	-	-
Forward-2	0.351	0.742	-	-	✓	✓	-	-
Forward-3	0.392	0.747	-	-	✓	✓	✓	-
Forward-4	0.418	0.757	-	-	✓	✓	✓	✓
Forward-5	0.417	0.771	-	✓	✓	✓	✓	✓
Backward-1	0.389	0.735	-	-	-	-	✓	-
Backward-2	0.386	0.736	-	-	-	✓	✓	-
Backward-3	0.403	0.752	-	-	-	✓	✓	✓
Backward-4	0.413	0.761	-	-	✓	✓	✓	✓
Backward-5	0.425	0.771	✓	-	✓	✓	✓	✓
All	**0.432**	0.778	✓	✓	✓	✓	✓	✓
Selection Ratio	-	-	5 (29.4%)	9 (52.9%)	9 (52.9%)	**13 (76.4%)**	**13 (76.4%)**	**12 (70.5%)**
MoE-Concat	0.507	0.788	✓	✓	✓	✓	✓	✓
MoE-CIU	**0.514**	**0.799**	✓	✓	✓	✓	✓	✓

Table 2: Spearman correlation between model prediction with feature selection and human ratings on Topical Chat. Selection ratio indicates how many times each feature was selected by different feature selection algorithms.

6.2 Persona Chat Results

Standalone Metric Performance Table 3 lists individual metric performance on Persona Chat. For Overall Ratings, CIU again showed the strongest correlation of 0.481 and for Uses Knowledge, RUBER-BERT achieved 0.688. Although the top metrics are identical to Table 1, the remaining metrics not only performed worse, but also fluctuated. For Topical Chat, we observed that BLEU was the least effective in predicting both Overall Ratings and Uses Knowledge. However in Persona Chat, BLEU outperforms METEOR and is comparable to CIU. BERTScore also has poor generalization as correlation dropped by 15.4% and 29.2% compared to Topical Chat. All of the correlations reported in Table 2 are statistically significant ($p < 0.001$).

Metric	Overall Ratings	Uses Knowledge
Random Guessing	0.011	0.017
BLEU	0.472	0.515
METEOR	0.223	0.379
ROUGE-L	0.202	0.387
RUBER-BERT	0.435	**0.688**
BERTScore	0.241	0.486
CIU - $freq$	0.461	0.667
CIU - pos	0.461	0.669
CIU	**0.481**	0.685

Table 3: Spearman correlation between metric predictions and human ratings on Persona Chat. Ablation study is indicated with minus sign where $freq$ stands for frequency and pos for position.

These findings show existing metrics have high variance across tasks. This is true for both lexical and learned metrics as BLEU, METEOR, ROUGE and BERTScore all suffered from significant performance drops. We believe that it is difficult to determine which metric works best ahead of time; nonetheless, CIU is consistently strong and reliable across both domains.

Unified Reward Performance According to Table 4 on predicting Overall Ratings, univariate feature selection combining five metrics excluding ROUGE performed best and achieved 0.529 correlation, a +4.8% improvement compared to best standalone metric CIU. Similarly for predicting Uses Knowledge, SVR using all features achieved the strongest correlation of 0.718, a +3.0% improvement over RUBER-BERT in Table 3. All correlations reported in Table 4 are statistically significant ($p < 0.001$). Overall, it is clear that benefits of combining different metrics generalize to different domains.

For feature selection ratios, we observed BLEU and RUBER-BERT were each selected 82.3% from 17 different feature selection strategies. While CIU was one of the most selected features in Topical Chat, CIU is the third best in Persona Chat with 76.4% selection ratio. Although BLEU were selected the most in Persona Chat, these performances do not carry over to Topical Chat since BLEU was only selected 5 times according to Table 2. Across both datasets, RUBER-BERT was selected most with 75.0% and CIU was second with 72.2%. All of these findings validate that CIU is the strongest and most reliable lexical metric in evaluating retrieval augmented conversations without any training.

Lastly, our proposed MoE-CIU outperformed the strongest feature selection baseline by 1.5% on predicting Uses Knowledge, but only a tiny increase on Overall Ratings. We suspect that with more

Metric	Overall Ratings	Uses Knowledge	BLEU	METEOR	ROUGE-L	CIU	RUBER-BERT	BERTScore
Univariate	**0.529**	0.683	✓	✓	-	✓	✓	✓
Random Forest	0.492	0.636	✓	-	✓	✓	✓	✓
Recursive-5	0.436	0.675	-	-	-	✓	-	-
Recursive-4	0.487	0.675	✓	-	-	✓	-	-
Recursive-3	0.483	0.666	✓	-	-	✓	✓	-
Recursive-2	0.505	0.711	✓	-	✓	✓	✓	-
Recursive-1	0.514	0.713	✓	-	✓	✓	✓	✓
Forward-1	0.381	0.674	-	-	-	✓	-	-
Forward-2	0.455	0.678	-	-	-	✓	✓	-
Forward-3	0.511	0.658	✓	-	-	✓	✓	-
Forward-4	0.504	0.701	✓	✓	-	✓	-	✓
Forward-5	0.507	0.702	✓	-	✓	✓	✓	✓
Backward-1	0.377	0.659	-	-	-	-	✓	-
Backward-2	0.446	0.654	✓	-	-	-	✓	-
Backward-3	0.501	0.645	✓	✓	-	-	✓	-
Backward-4	0.518	0.699	✓	✓	-	-	✓	✓
Backward-5	0.521	0.699	✓	✓	✓	-	✓	✓
All	0.519	**0.718**	✓	✓	✓	✓	✓	✓
Selection Ratio	-	-	**14 (82.3%)**	6 (35.2%)	6 (35.2%)	**13 (76.4%)**	**14 (82.3%)**	8 (47.0%)
MoE-Concat	0.521	0.721	✓	✓	✓	✓	✓	✓
MoE-CIU	**0.531**	**0.733**	✓	✓	✓	✓	✓	✓

Table 4: Spearman correlation between model predictions with feature selection and human ratings on Persona Chat. Selection ratio indicates how many times each feature was selected by different feature selection algorithms.

training data, the benefits will become more visible. Nonetheless, since it is difficult to expect which feature selection works best in advance, learning dynamic metric weights through MoE seems extremely useful.

Case Study To illustrate benefits of MoE-CIU, we selected one example in Table 5 where we show the context, the response being assessed, different metric scores, and the gold human rating.

$d_{context}$: Until 1805 in the us, the runner up in a presidential election automatically became the vice president.

Response: Yeah i wonder what the president of zimbabwe looks like?

CIU: 0.32 | RUBER-BERT: 0.00 | BERTScore: 0.36 | MoE-CIU: 3.33 | Rating: 3.66

Table 5: An example showing how MoE-CIU handles potentially conflicting signals from individual metrics.

Here, although RUBER-BERT was very confident in classifying this example as NOT useful, MoE-CIU still predicted 3.33 ratings, which is much closer to gold ratings given other useful signals from CIU and BERTScore. A more comprehensive insights and additional examples are included in Appendix A.

7 Conclusion

We introduced CIU, a novel utility metric for assessing the quality of retrieval augmented conversations. Based on our experiments on two popular retrieval augmented (a.k.a. knowledge-grounded) conversation corpus, we conclude that CIU was the best metric among other lexical baselines. Although RUBER-BERT surpassed CIU performance on Persona Chat, considering the complexity of RUBER-BERT (e.g., training and inference), CIU is still an easy-to-use metric that can achieve similar results with no training, which answers **Q1**.

For **Q2**, we demonstrated the potentials of unifying multiple independent metrics into a single reward signal without any LLM dependency. This was achieved through our MoE-CIU model, and experiments confirm its effectiveness over any standalone metric. Insights from this study suggest promising directions for applying MoE-CIU as a proxy for an unified reward signal to optimize.

Limitations

Although our work proposes an approach to model unified rewards, reward optimization approaches to update dialog policy (e.g., RLHF) are left for future work. MoE-CIU also requires a small number of human annotations, thus our approach will require manual labeling. Our proposed discounting functions were only validated on English corpora, thus it is unclear how well CIU can generalize to multilingual setting. Unfortunately, experiments involving different languages and cultures are beyond the scope of this paper. Lastly, instead of exhaustive comparisons, we only selected the most widely used metrics to keep our experiments simple.

Acknowledgements

We would like to express our gratitude to Zhiyu Chen and the anonymous reviewers for their insightful feedback on our work.

References

Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 605–614.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. 2002. A parallel mixture of svms for very large scale problems. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *CoRR*, abs/1904.10635.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Abhinav Jain, Vishwanath P. Singh, and Shakti P. Rath. 2019. A multi-accent acoustic model using mixture of experts for speech recognition. In *INTERSPEECH*, pages 779–783. ISCA.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017b. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Widad Machmouchi, Ahmed Hassan Awadallah, Imed Zitouni, and Georg Buscher. 2017. Beyond success rate: Utility as a search quality metric for online experiments. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293.

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

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 659–668.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 25–40.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170.

Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 473–482.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. *CoRR*, abs/1701.03079.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, user models, and satisfaction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 654–662.

Bangpeng Yao, Dirk Walther, Diane Beck, and Li Feifei. 2009. Hierarchical mixture of classification experts uncovers interactions between brain regions. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–434.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

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

Appendix

A Additional Case Studies of MoE-CIU

To illustrate how MoE-CIU benefits from ensemble learning, we highlight three examples in Table 6 where we show the context, the response being assessed, and the scores from the various models, and the gold human rating.

$d_{context}$: Until 1805 in the us, the runner up in a presidential election automatically became the vice president. Response: Yeah i wonder what the president of zimbabwe looks like? CIU: 0.32 RUBER-BERT: 0.00 BERTScore: 0.36 MoE-CIU: 3.33 Rating: 3.66
$d_{context}$: In september of 2010, the united nations appointed official ambassador to extraterrestrials in they case they would ever make contact with earth. Response: I'm not sure. I wonder if the un has an ambassador to aliens. CIU: 0.36 RUBER-BERT: 0.99 BERTScore: 0.31 MoE-CIU: 2.66 Rating: 2.67
$d_{context}$: Burger king had a promotion where if you unfriend 10 people on facebook you get a free whopper. Response: Did you know burger king had a promotion where they gave you a burger for every 10 people you unfriended? I could have had 100 burgers! CIU: 0.91 RUBER-BERT: 0.99 BERTScore: 0.62 MoE-CIU: 4.53 Rating: 5.00

Table 6: Example predictions from MoE-CIU highlighting how different metrics help recover false-positives from individual metrics.

One obvious issue we noticed was that it is difficult to calibrate the score distribution of learned models because on all three examples, RUBER-BERT is extremely confident about its decision. However, MoE architecture is particularly helpful since it can combine different metric outputs into a more continuous score distribution that is better aligned with human ratings.

In the first example, RUBER-BERT fails because the utterance and reference do not contain strong semantic relationship. Instead, they are loosely connected with an important keyword ('president'). RUBER-BERT was very confident in classifying this pair as NOT useful although human usefulness rating is 3.66. CIU and BERTScore successfully capture the overlap and assign a reasonable score compared to RUBER-BERT. The resulting model is capable of correcting RUBER-BERT's prediction to 3.33, which is only 0.33 off to human usefulness ratings. Without MoE, RUBER-BERT alone will predict this pair with 0.0 rating.

In the second example, RUBER-BERT strongly believes that the utterance and reference are semantically related. Although both inputs talk about alien ambassadors, the utterance does not use the information correctly. The reference clearly states United Nations appointed alien ambassadors but the utterance still questions the fact. RUBER-BERT is very confident that this example is highly related. However, CIU and BERTScore are able to regularize these effects if trained under MoE-CIU. The final score correctly predicted usefulness ratings with only 0.01 difference.

In the last example, it is clear that the input is highly relevant to $d_{context}$. Since individual metrics provide strong signals, MoE-CIU also predicted a very high rating of 4.53, which is close to 5.0.