

* Các chỉ số đánh giá cho nhiệm vụ tóm tắt của chatbot

1. Các chỉ số tự động

- 1.1. Chỉ số trùng lặp từ vựng (Lexical Overlap)
- 1.2. Chỉ số tương đồng ngữ nghĩa (Semantic Similarity)
- 1.3. Chỉ số trung thực & chính xác về sự thật (Factuality)

2. Chỉ số không cần bản tham chiếu (Reference-free / Unsupervised)



- 1. BARTScore
- 2. QAEval
- 3. UniEval

3. Đánh giá của con người (Human Evaluation)

- Mạch lạc, Liên quan, Tính dễ đọc và lưu loát, Trung thực, Tính hữu ích.

* 1. Các chỉ số tự động

1.1. Chỉ số trùng lặp từ vựng (Lexical Overlap)

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation):
- ROUGE-N: đo n-gram overlap (thường dùng ROUGE-1, ROUGE-2).
- ROUGE-L: dựa trên chuỗi con chung dài nhất (Longest Common Subsequence).
-  Ưu điểm: phổ biến, dễ hiểu.
-  Hạn chế: không nhận ra cách diễn đạt lại cùng nghĩa.

1.2. Chỉ số tương đồng ngữ nghĩa (Semantic Similarity)

- BERTScore:
 - Tính cosine similarity giữa embedding ngữ cảnh của token (BERT/transformer).
 - Đưa ra precision, recall, F-score.
- MoverScore:
 - Đo "khoảng cách Earth Mover" giữa embedding của hai văn bản.
 - nắm được sự dịch chuyển ý nghĩa toàn cục.
- BLEURT:
 - Mô hình deep learning được fine-tune theo đánh giá con người.
 - Xuất ra điểm phản ánh mức “giống đánh giá con người”.

* 1. Các chỉ số tự động

1.3. Chỉ số trung thực & chính xác về sự thật (Factuality)

- Đảm bảo tóm tắt không bịa đặt thông tin.
- QAGS (Question Answering and Generation for Summarization):
- Sinh câu hỏi từ tóm tắt → kiểm tra văn bản nguồn có trả lời được không.
- Nếu không trả lời được → tóm tắt có thể sai.
- FactCC:
- Mô hình phân loại (classification) để phát hiện inconsistency giữa nguồn và tóm tắt.
- SummaC:
- Dùng mô hình suy diễn (NLI – Natural Language Inference) để chấm điểm khả năng suy ra tóm tắt từ nguồn.



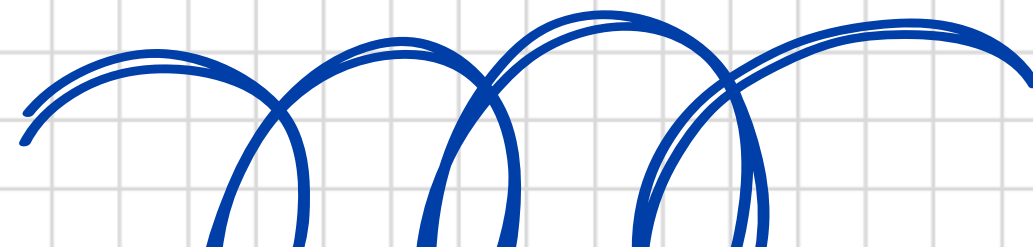
2. Chỉ số không cần bản tham chiếu (Reference-free / Unsupervised)

- 1. BARTScore
 - Cách hoạt động:
 - Dùng mô hình sinh BART đã pretrained.
 - Tính log-likelihood (xác suất sinh ra text) để chấm điểm.
 - Có 2 hướng:
 - source → summary: đo xem bản tóm tắt có hợp lý khi sinh từ văn bản nguồn không.
 - summary → source: đo xem văn bản nguồn có khớp ngược lại với tóm tắt không.
 - Ý nghĩa: Điểm số càng cao → tóm tắt càng “phù hợp” với nguồn theo xác suất mô hình.
 - Ưu điểm: Không cần bản tham chiếu.
 - Hạn chế: Phụ thuộc vào chất lượng mô hình nền (BART).
- 2. QAEval
 - Cách hoạt động:
 - Dựa trên ý tưởng của QAGS (Question Answering & Generation for Summarization).
 - Sinh ra các câu hỏi từ bản tóm tắt.
 - Dùng văn bản nguồn để trả lời → nếu trả lời được → nghĩa là tóm tắt trung thực.
 - Điểm khác với QAGS:
 - Có thể chạy không cần reference summary.
 - Chỉ cần nguồn gốc (source text) là đủ.
 - Ý nghĩa: Nếu tóm tắt chứa thông tin không có trong nguồn → hệ thống QA sẽ không trả lời được.
 - Ưu điểm: Kiểm tra factuality trực tiếp.
 - Hạn chế: Tốn tài nguyên (vì cần QA model).



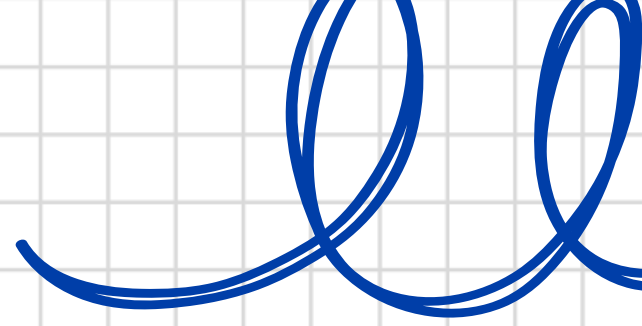
2. Chỉ số không cần bản tham chiếu (Reference-free Unsupervised)

- 3. UniEval
 - Cách hoạt động:
 - Là một mô hình pretrained đa nhiệm cho nhiều loại evaluation.
 - Được huấn luyện để chấm điểm các tiêu chí như:
 - Relevance (liên quan)
 - Consistency (nhất quán, trung thực với nguồn)
 - Readability (dễ đọc, trôi chảy)
 - Không cần gold summary.
 - Ý nghĩa: Cho điểm “giống người” hơn, vì được fine-tune cho nhiều chiều chất lượng.
 - Ưu điểm: Bao phủ nhiều tiêu chí trong một mô hình.
 - Hạn chế: Cần mô hình pretrained riêng, không phổ biến bằng ROUGE/BERTScore.

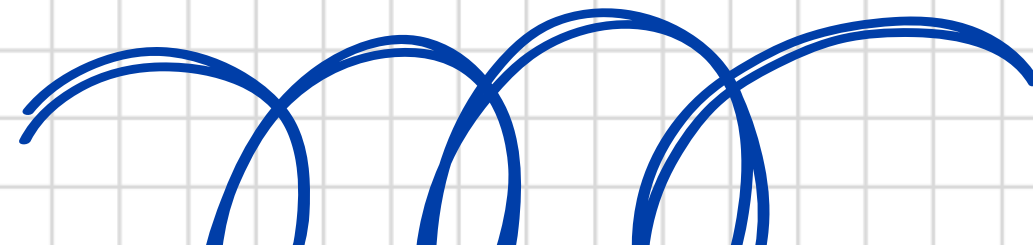




3. Đánh giá của con người (Human Evaluation)



- Mạch lạc (Coherence): Cấu trúc logic, ý nối tiếp mạch lạc.
- Liên quan (Relevance): Bao phủ đủ ý chính, không bỏ sót chi tiết quan trọng.
- Dễ đọc & lưu loát (Fluency): Ngữ pháp chuẩn, tự nhiên.
- Trung thực (Faithfulness): Không thêm/bịa thông tin sai.
- Hữu ích (Usefulness): Có giá trị thực cho người dùng.



Nhóm	Phương pháp	Cách đo	Ưu điểm	Hạn chế	Use case
Lexical Overlap	ROUGE	So trùng n-gram, LCS (chuỗi con dài nhất)	Dễ tính, phổ biến	Không nhận ra paraphrase	So sánh mô hình khi có gold summary
	BLEU	Precision của n-gram	Phạt nội dung thừa	Không tốt cho tóm tắt ngắn	Ban đầu dùng cho dịch, ít dùng tóm tắt
	METEOR	Trùng từ + stemming + đồng nghĩa	Gần human hơn BLEU	Vẫn dựa nhiều vào từ	Dịch & tóm tắt có paraphrase
Semantic Similarity	BERTScore	Cosine sim. giữa embedding từ (BERT)	Hiểu paraphrase	Cần mô hình pretrained	Đo ý nghĩa gần đúng
	MoverScore	Earth Mover Distance giữa embedding	Nhận ra dịch chuyển ngữ nghĩa lớn	Tính toán nặng	So sánh tóm tắt dài, nhiều paraphrase
	BLEURT	Mô hình fine-tune theo đánh giá con người	Tương quan cao với human	Cần mô hình đã huấn luyện	Đánh giá gần giống human judgment
	UniEval	Pretrained evaluator đa nhiệm (fluency, relevance,	Đa tiêu chí, không cần reference	Phụ thuộc model có sẵn	Đánh giá chatbot sinh tóm tắt tự do

Nhóm	Phương pháp	Cách đo	Ưu điểm	Hạn chế	Use case
Factuality	FactCC	Classifier check inconsistency	Nhanh, tự động	Không phát hiện mọi sai fact	Tóm tắt tin tức, pháp lý
	QAGS	Sinh câu hỏi từ summary → check source	Kiểm tra fact chính xác	Sinh QA tốn tài nguyên	Đảm bảo không “hallucination”
	QuestEval	QA-based nhưng cải tiến hơn QAGS	Không cần gold summary	Phức tạp hơn ROUGE	Khi không có bản tham chiếu
	SummaC	NLI-based: check entailment giữa source-summary	Nhận diện mâu thuẫn	Phụ thuộc NLI model	Phát hiện lỗi logic, fact
	DAE	Đánh giá từng câu trong summary có augmented không	Chi tiết, theo câu	Cần alignment	Tóm tắt tài liệu dài
	FEQA	Tương tự QAGS nhưng tối ưu hóa QA	Tốt hơn cho câu dài	Khó với câu phức tạp	Fact-check tài liệu khoa học
	MNLI-doc	Dùng MultiNLI để check entailment document-level	Phát hiện mâu thuẫn đa câu	Khó tính toàn văn bản	Văn bản dài, đa chiều
Reference-free	BARTScore	Xác suất sinh (source→summary, ngược lại)	Không cần gold summary	Dựa vào 1 model duy nhất	Khi chỉ có văn bản nguồn
	QAEval	Sinh QA từ summary, check bằng source	Đảm bảo fact	Tốn chi phí QA	Fact-check khi không có reference
	UniEval	Pretrained evaluator đa nhiệm (fluency, relevance,	Đa tiêu chí, không cần reference	Phụ thuộc model có sẵn	Đánh giá chatbot sinh tóm tắt tự do

* nghiên cứu tiêu biểu

ChatGPT as a Factual Inconsistency Evaluator for Text Summarization

Zheheng Luo, Qianqian Xie*, Sophia Ananiadou

Department of Computer Science, The University of Manchester
{zheheng.luo, qianqian.xie, sophia.ananiadou}@manchester.ac.uk

Human-like Summarization Evaluation with ChatGPT

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, Xiaojun Wan

Wangxuan Institute of Computer Technology, Peking University
{gaomingqi, xjyin, wanxiaojun}@pku.edu.cn
{ruanjie, sunrenliang}@stu.pku.edu.cn
yangshiping@bupt.edu.cn



ngiên cứu tiêu biểu

Human-like Summarization Evaluation with ChatGPT

- Bài báo này cho thấy ChatGPT có khả năng thực hiện đánh giá tóm tắt văn bản một cách linh hoạt theo nhiều phương pháp giống con người, và trong nhiều trường hợp, hiệu suất của nó vượt trội hơn các thước đo tự động phổ biến

Metric Name	consistency			relevance			fluency			coherence		
	sample	system	dataset	sample	system	dataset	sample	system	dataset	sample	system	dataset
ROUGE-1	0.153	0.744	0.137	0.326	0.744	0.302	0.113	0.730	0.080	0.167	0.506	0.184
ROUGE-2	0.179	0.779	0.129	0.290	0.621	0.245	0.156	0.690	0.062	0.184	0.335	0.145
ROUGE-L	0.111	0.112	0.109	0.311	0.362	0.284	0.103	0.306	0.079	0.128	0.138	0.141
BERTScore	0.105	-0.077	0.118	0.312	0.324	0.362	0.189	0.246	0.150	0.284	0.477	0.317
MoverScore	0.151	0.679	0.150	0.318	0.724	0.294	0.126	0.687	0.119	0.159	0.474	0.178
BARTScore_s_h	0.299	0.800	0.269	0.264	0.524	0.363	0.243	0.614	0.187	0.322	0.477	0.335
BARTScore_h_r	0.097	0.606	0.101	0.178	0.147	0.246	0.002	0.261	0.000	0.017	-0.115	0.064
BARTScore_r_h	-0.075	-0.556	-0.090	-0.081	-0.112	-0.136	0.013	-0.212	0.019	0.044	0.165	-0.010
BARTScore_cnn_s_h	0.367	0.435	0.334	0.356	0.765	0.394	0.349	0.746	0.285	0.448	0.700	0.408
BARTScore_cnn_h_r	0.171	0.771	0.106	0.320	0.456	0.244	0.111	0.561	0.066	0.153	0.174	0.130
BARTScore_cnn_r_h	0.001	-0.079	-0.004	0.146	0.312	0.221	0.107	0.297	0.145	0.228	0.506	0.236
ChatGPT	0.435	0.833	0.425	0.433	0.901	0.445	0.419	0.889	0.410	0.561	0.832	0.557

Table 1: Spearman's ρ of sample level, system level, and dataset level on SummEval.

Metric Name	Accuracy
ROUGE-1	0.5869
ROUGE-2_f	0.4997
ROUGE-L_f	0.5647
BARTScore	0.5674
MoverScore	0.5864
BARTScore_s_h	0.5858
BARTScore_h_r	0.6151
BARTScore_r_h	0.5317
BARTScore_cnn_s_h	0.5880
BARTScore_cnn_h_r	0.5934
BARTScore_cnn_r_h	0.5089
ChatGPT	0.6178

Table 3: Accuracy of pairwise comparison on TLDR.

Metric Name	Accuracy
DAE	0.6304
FactCC	0.5362
ChatGPT	0.6436

Table 4: Accuracy of the binary determination of SCUs on REALSumm.

	QAGS_CNN	QAGS_XSUM
DAE	0.8459	0.6360
FactCC	0.7731	0.4937
ChatGPT	0.8488	0.7573

Table 5: Accuracy of binary factuality evaluation on QAGS.

* nghiên cứu tiêu biểu

ChatGPT as a Factual Inconsistency Evaluator for Text Summarization

- Bài báo này tập trung chuyên sâu vào khả năng của ChatGPT trong việc phát hiện sự mâu thuẫn về mặt thông tin (factual inconsistency) và kết luận rằng nó có tiềm năng lớn nhưng cũng bộc lộ những điểm yếu cần khắc phục

Methods	SUMMAC Benchmark Datasets					
	CoGenSum	XsumFaith	Polytope	FactCC	SummEval	FRANK
NER Overlap	53.0	63.3	52.0	55.0	56.8	60.9
MNLI-doc	57.6	57.5	61.0	61.3	66.6	63.6
FactCC-CLS	63.1	57.6	61.0	75.9	60.1	59.4
DAE	63.4	50.8	62.8	75.9	70.3	61.7
FEQA	61.0	56.0	57.8	53.6	53.8	69.9
QuestEval	62.6	62.1	70.3	66.6	72.5	82.1
SummaC _{ZS}	70.4	58.4	62.0	83.8	78.7	79.0
SummaC _{Conv}	64.7	66.4	62.7	89.5	81.7	81.6
ChatGPT _{ZS}	63.3	64.7	56.9	74.7	76.5	80.9
ChatGPT _{ZS-COT}	74.3	63.1	61.4	79.5	83.3	82.6

Table 2: Balanced accuracy results of inconsistency detect models on the test set of SummaC. Results of baselines are referenced from the paper (Laban et al., 2022).

* nghiên cứu tiêu biểu

ChatGPT as a Factual Inconsistency Evaluator for Text Summarization

Model	Ranking Acc.
FactCC	70.0
MNLI-doc	78.3
Rule-based dependency	74.8
DAE	83.6
Human	83.9
ChatGPT	85.2

Table 3: Performance of models on the summary ranking task. Results of baselines are reported in [Goyal and Durrett \(2020\)](#).

Metrics	FRANK		FRANK(CNN/DM)		FRANK(XSum)		SummEval	
	Pear. ρ	Spear. r	Pear. ρ	Spear. r	Pear. ρ	Spear. r	Pear. ρ	Spear. r
FEQA	0.00	0.01	-0.01	-0.01	0.02	0.07	-	-
QAGS	0.06	0.08	0.13	0.09	-0.02	0.01	-	-
DAE	0.16	0.14	0.25	0.24	0.04	0.28	0.20	0.27
FactCC	0.20	0.30	0.36	0.33	0.07	0.25	0.32	0.34
ChatGPT	0.70	0.69	0.50	0.46	0.34	0.27	0.49	0.35

Table 4: Pearson correlation, and spearman rank correlation coefficients between human judgements and evaluation scores of different methods.