

Large Language Model Applications

Evaluation

I'm in school and my phone's dying. Any solid recommendations?

If you're looking for a phone that handles everyday life smoothly, there are quite a few directions that can work well depending on what you enjoy.

A recent iPhone, even if it's not the newest release, is a dependable choice. iPhones receive updates for many years, stay responsive for common tasks, and offer a consistent, tightly integrated experience across apps and accessories. Setup is simple and the interface is easy to navigate.

A current Samsung Galaxy S phone is another strong all-around option. You get a bright display, versatile cameras, and durable hardware. Samsung provides regular Android and security updates, and there are built-in customization features if you like adjusting layouts or workflows.

An international or refurbished Android flagship from an online marketplace can appeal if you enjoy exploring. These devices often include fast processors, generous RAM, and region-specific firmware builds discussed in enthusiast forums. Sideload utilities or trying alternate app stores can make the experience feel more personalized.

You might also consider a midrange Android device from a smaller manufacturer that emphasizes battery life or distinctive hardware design. Some brands focus on minimalist software, oversized batteries, or niche features that larger brands skip.

Finally, a privacy- or enthusiast-oriented Android phone that allows deeper system customization can be interesting if you enjoy adjusting permissions or experimenting with alternative builds.

A

If you're evaluating phones in terms of security architecture, update guarantees, and long-term platform stability, a smaller set of devices offers clearly defined lifecycle models.

A current-generation iPhone provides one of the most predictable support horizons in the industry. Apple's vertically integrated hardware-software stack, secure enclave subsystem, and strict code-signing framework create a tightly controlled security boundary. Major OS releases and security patches are deployed consistently across supported devices, reducing fragmentation risk.

A current Samsung Galaxy S-series device represents a mature flagship Android implementation with an explicitly documented multi-year OS and security update policy. Samsung's hardware stack—high-refresh AMOLED displays, advanced ISP processing, and modern modem integration—is competitive at the top tier. Enterprise support features and formalized patch commitments contribute to structured lifecycle management.

A recent Google Pixel functions as Google's reference Android platform. Pixels typically receive Android releases and monthly security patches with minimal latency, reflecting direct alignment with Google's upstream roadmap. Hardware-backed protections such as verified boot and dedicated security modules are integrated into the broader system design.

Together, these devices emphasize formal vendor commitments, transparent update cadences, and well-defined security boundaries over time.

B

poll: if someone tells you that they
work on evaluation, what does it mean?

Evaluation

- Evaluation refers to the measurement of different properties in a system's outputs.

poll: what sort of properties are
we interested in measuring?

AI as a Tool

- AI can be a **tool** to accomplish tasks.
- People use AI to accomplish a broad set of request types in a wide set of domains,
 - travel booking
 - navigation
 - education
 - ...
- **User-focused evaluation:** how well does the system help the user accomplish the task the user is interested in?



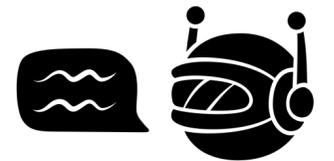
start
task



interact



finish
task



receive
feedback



User-focused evaluation

Types of measures

- **Nominal/categorical:** Categorize data into distinguish groups, without any order or quantitative meaning assigned to categories.
- **Ordinal:** Data are ranked in an order, but the differences between ranks are not equal or meaningful.
- **Interval:** Values are ordered, and the differences between them are meaningful and consistent, but there is no true zero point.
- **Ratio:** Values have order, equal intervals, and a true zero that indicates absence of the quantity.

User-focused evaluation

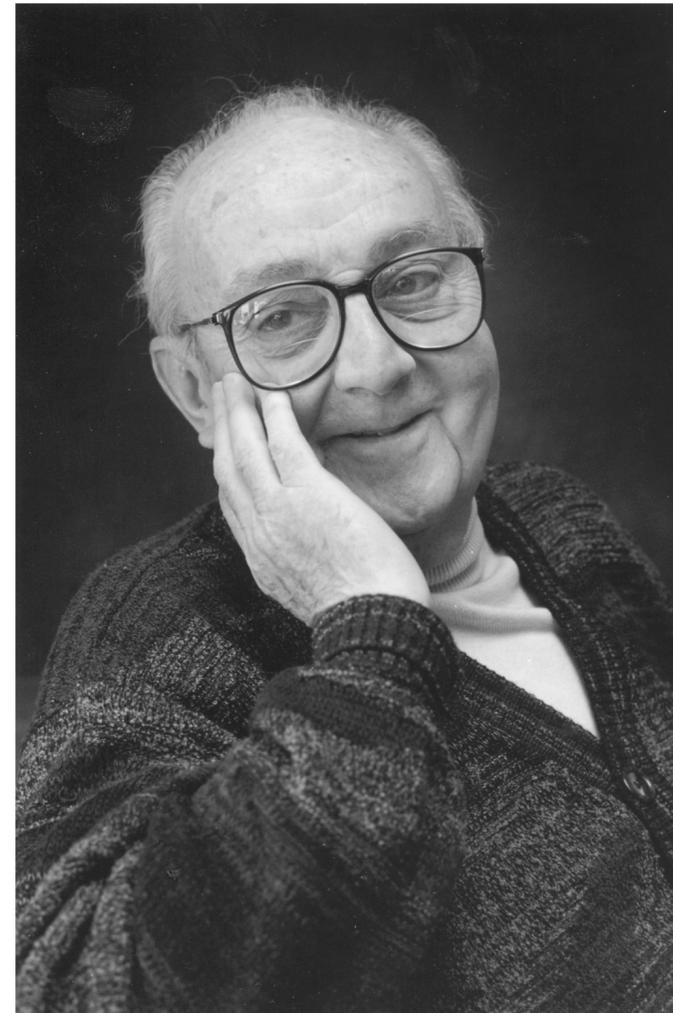
Measurement

- We are interested in the quantification of task assistance.
 - allows us to conduct quantitative experiments (e.g., "is the mean task assistance of model A better than that of model B?")
 - allows us to directly optimize for task assistance
- Gold standard measure often comes from extrinsic evaluation (2 Feb lecture) in a real environment.
 - observing to see if the user completes
 - post-task survey
- Gold standard measure is expensive
 - data collection cost
 - time

User-focused evaluation

Measurement

- Can address cost of measurement by approximating, simulating, or modeling the real data
 - do not need to constantly collect data
 - can often simulate or model faster than surveys
- A good evaluation is one that approximates the gold standard well.
- Beyond approximation, there are other properties that we look for in evaluations...



"all models are wrong but some are useful."

—George Box

User-focused evaluation

Validity

- “Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1989b).”
- **face validity:** visual inspection of the metric; "do the concepts assumed make sense?"
 - advantage: demonstrates interpretability of the metric
 - disadvantage: considered subjective
- **criterion validity:** consistency with a gold standard evaluation; "how correlated are the ratings with the gold standard?"
 - advantage: more objective
 - disadvantage: need to collect a gold standard

User-focused evaluation

Validity

- **predictive validity:** consistency with some future outcome; "how well does a recommended lecture predict future learning?"
 - advantage: reflects extrinsic goal
 - disadvantage: need to wait to observe future outcome; more difficult to attribute
- **convergent validity:** consistency with other evaluations of the same property; "how correlated are the ratings with established metrics of the same property?"
 - advantage: measures whether a new evaluation is consistent with accepted norms
 - disadvantage: too much consistency = no need information
- **discriminate validity:** consistency with other evaluations of a different property; "how different are the ratings with established metrics of a different property?"
 - advantage: measures how much new information there is in an evaluation
 - disadvantage: without other validities, can be random

User-focused evaluation

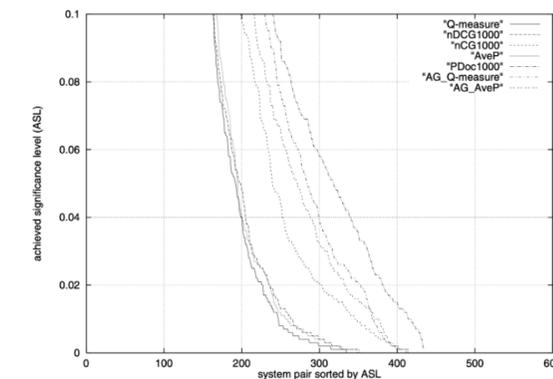
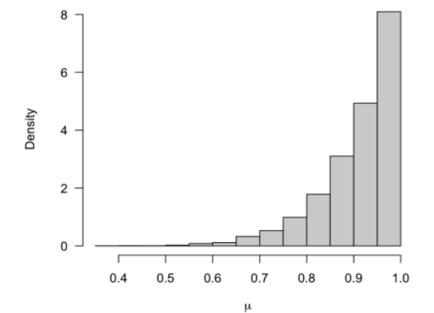
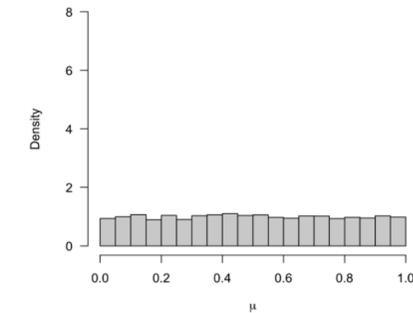
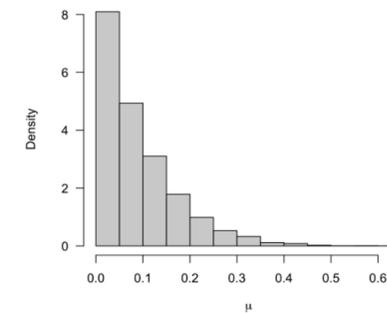
Reliability

- “The reliability of a metric is the extent to which the result is subject to random measurement error and thus (in)consistent across repeated measures, such as different (sub-)datasets within a benchmark or different raters scoring the model’s output in human evaluation (Zhang et al. 2023).”
- **robustness to labeling error:** for metrics that use labeled data, how does the value change as labeled data is corrupted or removed?
 - advantage: measures the sensitivity to different sources of error in human labeling
 - disadvantage: need a good noise model
- **inter-rater reliability:** when have multiple raters, consistency between raters (e.g., Krippendorff's alpha).
 - advantage: well-established methods
 - disadvantage: may depend on the rater pool

User-focused evaluation

Sensitivity

- **ties:** how many ties does an evaluation detect between a set of models?
 - advantage: we are primarily interested in comparing systems, not necessarily outputs
 - disadvantage: there may be many ties but still be able to distinguish systems; depends on systems
- **discriminative power:** how many statistically significant differences does an evaluation detect between a set of models?
 - advantage: we are primarily interested in significant differences when comparing systems, not just ties
 - disadvantage: more coarse; there may be many granular inconsistencies



User-focused evaluation

Measurement

- Can address cost of measurement by approximating, simulating, or modeling the real data
 - do not need to constantly collect data
 - can often simulate or model faster than surveys
- Extrinsic labels from human raters (e.g., crowdsourcing)
 - instead of gathering data from someone in some real context, simulate the context and ask people in a lab setting.
 - increases throughput (especially of rare contexts)
- Intrinsic labels from human raters
 - instead of gathering data from someone in some simulated context, gather data that can be used with a model to predict task improvement
 - can be faster; can generalize (if model is good)

poll: what are some metrics that use intrinsic data?

User-focused evaluation

ROUGE

- Given a gold standard or target reference string for an input, we can compute the similarity between the two as a measure.
- ROUGE measures the overlap in tokens between the reference and the system output.
- Used widely in text summarization.
- How can we game this metric?

$$\mu(y, \tilde{y}, n) = \frac{|\mathcal{G}_n(y) \cap \mathcal{G}_n(\tilde{y})|}{|\mathcal{G}_n(y)|}$$

$\mathcal{G}_n(y)$ n -grams in gold standard answer

$\mathcal{G}_n(\tilde{y})$ n -grams in system answer

User-focused evaluation

ROUGE

Method	DUC 2001 100 WORDS SINGLE DOC						DUC 2002 100 WORDS SINGLE DOC					
	1 REF			3 REFS			1 REF			2 REFS		
	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP	CASE	STEM	STOP
R-1	0.76	0.76	0.84	0.80	0.78	0.84	0.98	0.98	0.99	0.98	0.98	0.99
R-2	0.84	0.84	0.83	0.87	0.87	0.86	0.99	0.99	0.99	0.99	0.99	0.99
R-3	0.82	0.83	0.80	0.86	0.86	0.85	0.99	0.99	0.99	0.99	0.99	0.99
R-4	0.81	0.81	0.77	0.84	0.84	0.83	0.99	0.99	0.98	0.99	0.99	0.99
R-5	0.79	0.79	0.75	0.83	0.83	0.81	0.99	0.99	0.98	0.99	0.99	0.98
R-6	0.76	0.77	0.71	0.81	0.81	0.79	0.98	0.99	0.97	0.99	0.99	0.98
R-7	0.73	0.74	0.65	0.79	0.80	0.76	0.98	0.98	0.97	0.99	0.99	0.97
R-8	0.69	0.71	0.61	0.78	0.78	0.72	0.98	0.98	0.96	0.99	0.99	0.97
R-9	0.65	0.67	0.59	0.76	0.76	0.69	0.97	0.97	0.95	0.98	0.98	0.96
R-L	0.83	0.83	0.83	0.86	0.86	0.86	0.99	0.99	0.99	0.99	0.99	0.99
R-S*	0.74	0.74	0.80	0.78	0.77	0.82	0.98	0.98	0.98	0.98	0.97	0.98
R-S4	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-S9	0.84	0.85	0.84	0.87	0.88	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-SU*	0.74	0.74	0.81	0.78	0.77	0.83	0.98	0.98	0.98	0.98	0.98	0.98
R-SU4	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-SU9	0.84	0.84	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99
R-W-1.2	0.85	0.85	0.85	0.87	0.87	0.87	0.99	0.99	0.99	0.99	0.99	0.99

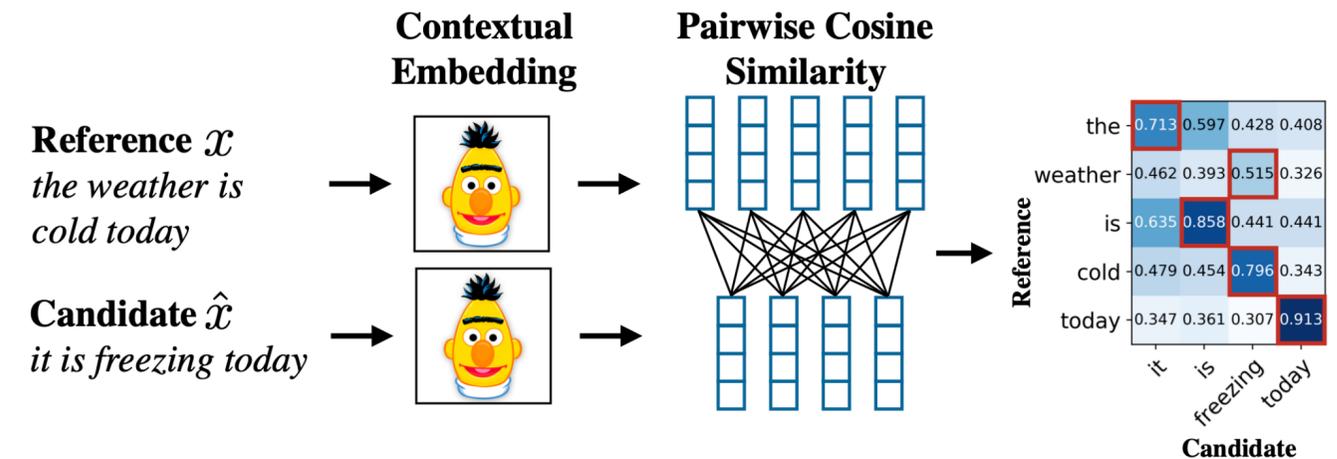
Table 1: Pearson's correlations of 17 ROUGE measure scores vs. human judgments for the DUC 2001 and 2002 100 words single document summarization tasks

poll: what are some limitations of
ROUGE?

Embedding-based metrics

BERTScore

- can we leverage advances in NLP to address lack of non-lexical similarity in metrics?
 - assume we have access to a model that provides word similarity.
- **BERTScore** uses embeddings to generalize classic evaluation metrics.
- Given a length n target sequence, construct an $k \times n$ matrix of embedded words; do the same for the system output.
- Compare the sets of embeddings.



$$\mu_P(y, \tilde{y}) = \frac{1}{|\tilde{y}|} \sum_{\tilde{y}_i \in \tilde{y}} \max_{y_i \in y} \phi_i^\top \tilde{\phi}_i$$

$$\mu_R(y, \tilde{y}) = \frac{1}{|y|} \sum_{y_i \in y} \max_{\tilde{y}_i \in \tilde{y}} \phi_i^\top \tilde{\phi}_i$$

ϕ_i Bert embedding of y_i

Embedding-based metrics

BERTScore

Metric	en↔cs (5k/5k)	en↔de (78k/ 20k)	en↔et (57k/32k)	en↔fi (16k/10k)	en↔ru (10k/22k)	en↔tr (9k/1k)	en↔zh (33k/29k)
BLEU	.233/.389	.415/.620	.285/.414	.154/.355	.228/.330	.145/.261	.178/.311
ITER	.198/.333	.396/.610	.235/.392	.128/.311	.139/.291	-.029/.236	.144/ –
RUSE	.347/ –	.498/ –	.368/ –	.273/ –	.311/ –	.259/ –	.218/ –
YiSi-1	.319/.496	.488/.691	.351/.546	.231/.504	.300/.407	.234/.418	.211/.323
P_{BERT}	.387/.541	.541/.715	.389/.549	.283/.486	.345/.414	.280/.328	.248/.337
R_{BERT}	.388/. 570	.546/. 728	.391/. 594	.304/.565	.343/.420	.290/. 411	.255/. 367
F_{BERT}	.404/.562	.550/.728	.397/.586	.296/.546	.353/.423	.292/.399	.264/.364
F_{BERT} (idf)	.408/.553	.550/.721	.395/.585	.293/.537	.346/. 425	.296/.406	.260/.366

Table 4: Kendall correlations with segment-level human judgments on WMT18. For each language pair, the left number is the to-English correlation, and the right is the from-English. We bold correlations of metrics not significantly outperformed by any other metric under bootstrap sampling for that language pair and direction. The numbers in parenthesis are the number of candidate-reference sentence pairs for each language pair and direction.

Embedding-based metrics

BARTScore

- can leverage pre-trained models to generate token-level scores for sequences
- flexible formalism that allows multiple metric definitions
- can be extended to fine-tune or prompt pre-trained model

general form

$$\mu(y, x) = \sum_{i=1}^{|y|} \omega_{y_i} \log_2 p_{\theta}(y_i | y_{<i}, x)$$

θ pretrained model

precision $\mu(y, \tilde{y}) = \sum_{i=1}^{|\tilde{y}|} \omega_{\tilde{y}_i} \log_2 p_{\theta}(\tilde{y}_i | \tilde{y}_{<i}, y)$

recall $\mu(y, \tilde{y}) = \sum_{i=1}^{|y|} \omega_{y_i} \log_2 p_{\theta}(y_i | y_{<i}, \tilde{y})$

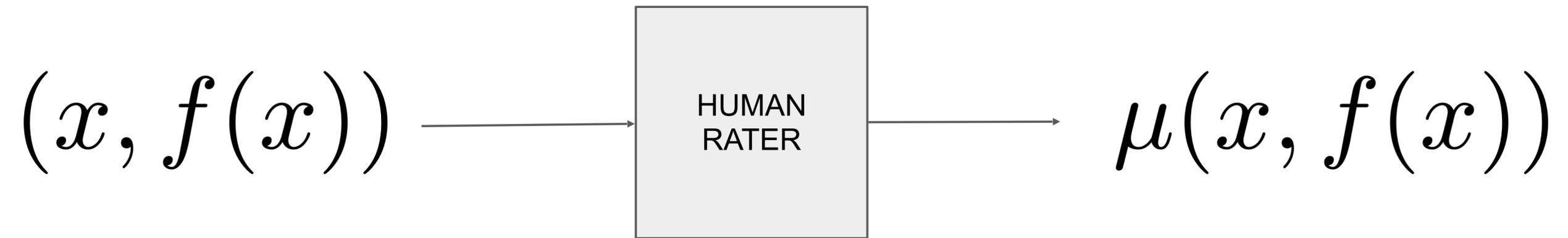
Embedding-based metrics

BARTScore

	REALSumm	SummEval				NeR18				Avg.
	COV	COH	FAC	FLU	INFO	COH	FLU	INFO	REL	
ROUGE-1	0.498	0.167	0.160	0.115	0.326	0.095	0.104	0.130	0.147	0.194
ROUGE-2	0.423	0.184	0.187	0.159	0.290	0.026	0.048	0.079	0.091	0.165
ROUGE-L	0.488	0.128	0.115	0.105	0.311	0.064	0.072	0.089	0.106	0.164
BERTScore	0.440	0.284	0.110	0.193	0.312	0.147	0.170	0.131	0.163	0.217
MoverScore	0.372	0.159	0.157	0.129	0.318	0.161	0.120	0.188	0.195	0.200
PRISM	0.411	0.249	0.345	0.254	0.212	0.573	0.532	0.561	0.553	0.410
BARTSCORE	0.441	0.322†	0.311	0.248	0.264	0.679†	0.670†	0.646†	0.604†	0.465

Learned Metrics

Metric-based evaluation



- Extrinsic evaluations from real people or crowd workers can be used as training data.
- If the evaluation measures the quality of a single output, we can train a classifier or regressor.

Learned Metrics

Metric-based evaluation

Please Rate the Story Fragment

The goal of this task is to rate story fragments on four criteria.

NOTE: Please take the time to **fully read** and **understand** the story fragment. **We will reject** submissions from workers that are clearly spamming the task.

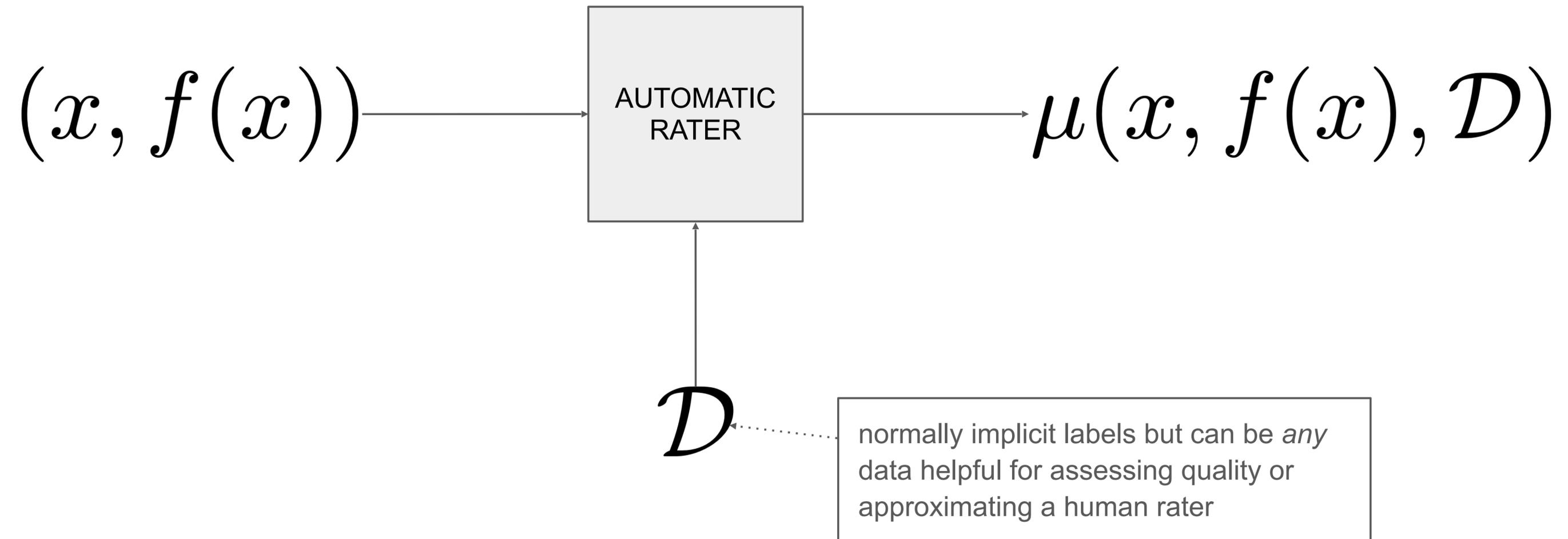
Story Fragment

The night before came as a shock for Oren, he was always a conscientious child. It was a necessary skill of a new master, an inherent capability to make the world a better place. But no, today, the day he brought his sister to his cooking school was the first time Oren had been shocked out of a small calm. He looked over at his sister in the small room, who was idly flipping through the magazine he had brought with him, and then back to the breakfast. It took all his willpower to stay calm, he could tell from the way the noodles he was looking at were slathered in gherkin and he felt the freshness of the rice. He shook his head in disbelief, his stomach began to churn and he was too exhausted to react, he was just preparing to go to bed.

1. How **grammatically correct** is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest)
(lowest) 1 2 3 4 5 (highest)
2. How well do **the sentences** in the story fragment **fit together**? (on a scale of 1-5, with 1 being the lowest)
(lowest) 1 2 3 4 5 (highest)
3. How **enjoyable** do you find the story fragment? (on a scale of 1-5, with 1 being the lowest)
(lowest) 1 2 3 4 5 (highest)
4. Now read the **PROMPT** based on which the story fragment was written.
PROMPT: After brushing your teeth in the morning you go downstairs to fry an egg, but when you try the frying pan buzzes at you and text appears reading, "level 18 cooking required to use object".
How **relevant** is the **story fragment** to the **prompt**? (on a scale of 1-5, with 1 being the lowest)
(lowest) 1 2 3 4 5 (highest)

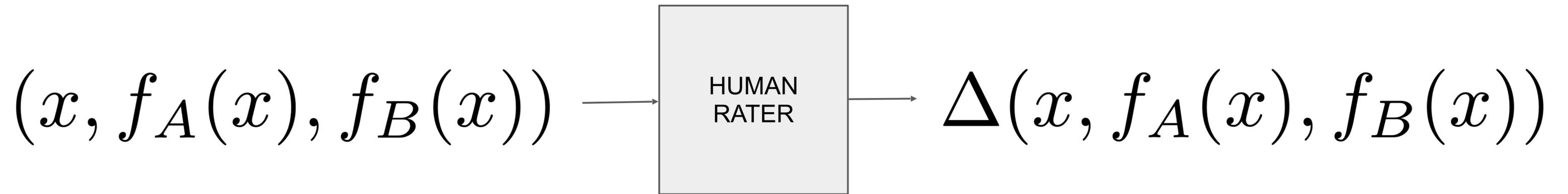
Learned Metrics

Metric-based evaluation



Learned Metrics

Preference-based evaluation



- Sometimes we can get more sensitivity by comparing two system outputs directly (instead of comparing metrics)

Learned Metrics

Preference-based evaluation

Query: espn sports

Aspect: Take me to the ESPN Sports home page.

You can find results from two different search engines in the table below. Each of the documents may contain a summary or snippet and the URL to help you make your decision. Which of these results would you choose?

Results 1	Results 2
<p>1. Le Anne Schreiber News, Videos, Photos, and PodCasts - ESPN Explore the comprehensive le anne schreiber archive on ESPN.com, including news, features, video clips, PodCasts, photos, and more. http://search.espn.go.com/le-anne-schreiber/</p> <p>2. Espn Sport http://ten-cartoons.info/espn-sport</p> <p>▪ ▪ ▪</p>	<p>1. ESPN: The Worldwide Leader In Sports http://espn.go.com/</p> <p>2. ESPN: The Worldwide Leader In Sports ESPN.com provides comprehensive sports coverage. Complete sports information including NFL, MLB, NBA, College Football, College Basketball scores and news. http://sports.espn.go.com/</p> <p>▪ ▪ ▪</p>

If you are a user requiring documents about the required aspect above, which result would you choose?

Left result is better Results are equally good Right result is better None of the results are relevant

Please mention your reason below (incomplete answers will not be accepted):

The right had more relevant information.

Learned Metrics

Preference-based evaluation

Task : Evaluate the given images and rate them in order
Please select the image that looks more realistic overall. If you are unsure, select whichever image you prefer.

a blue cow is standing next to a tree with red leaves and yellow fruit. the cow is standing in a field with white flowers. impressionistic painting.



1. Which image matches with the caption better?

Rating :

Rating :

2. Which image is more realistic?

Rating :

Rating :

Submit

Learned Metrics

Preference-based evaluation

1. Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.



Various policies are used to sample N summaries.



Two summaries are selected for evaluation.



A human judges which is a better summary of the post.

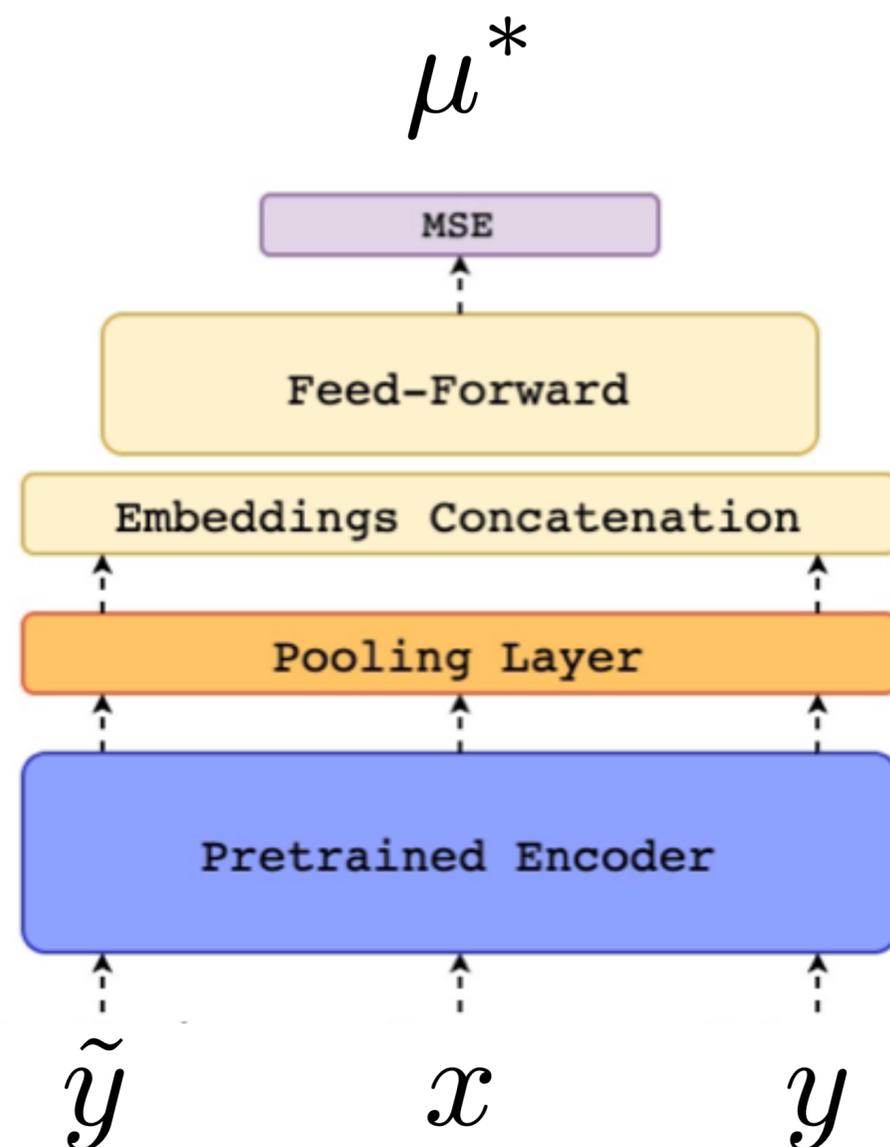


"j is better than k"

poll: what methods discussed
previously does learning from
preferences look like?

Learned Metrics

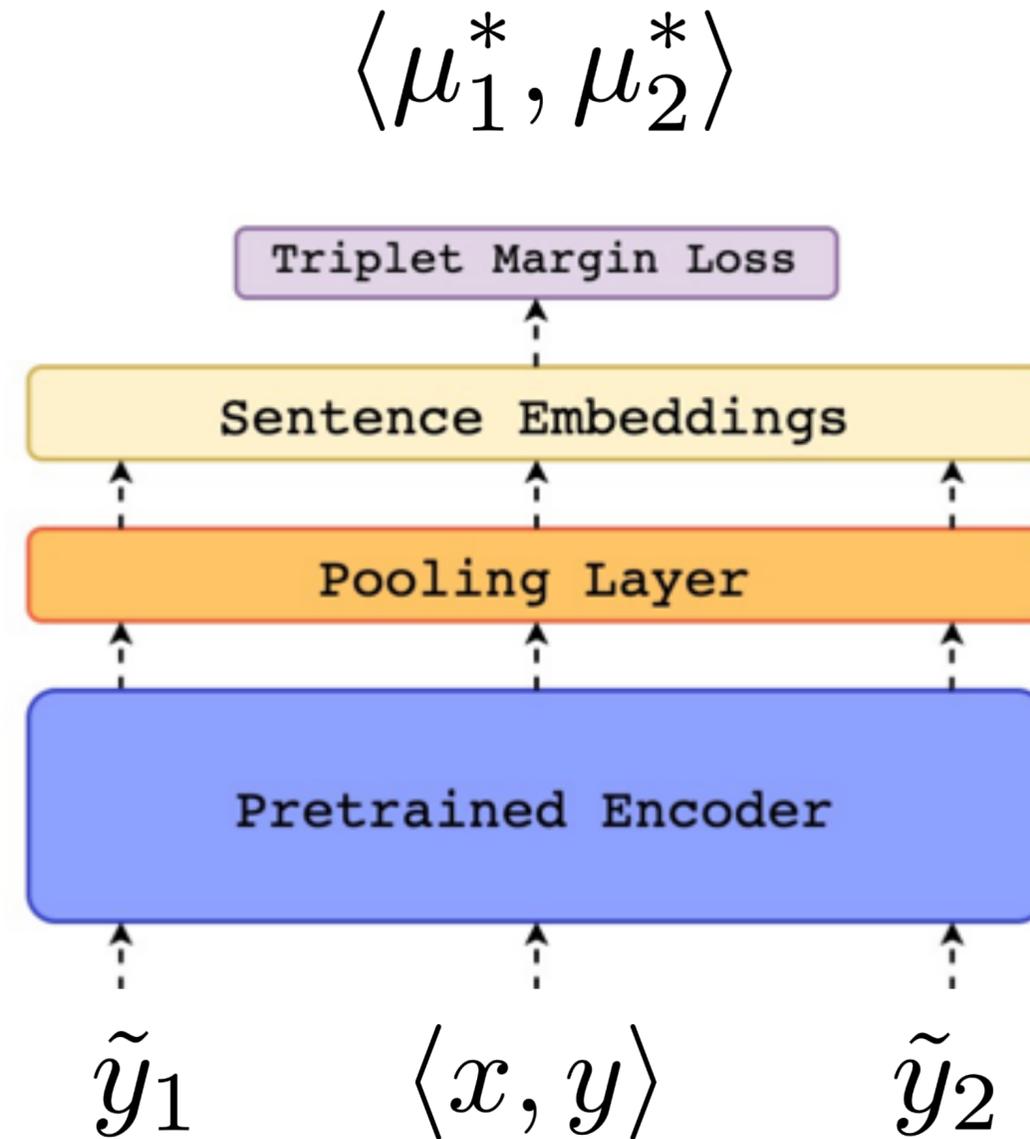
COMET



regress against the rating

Learned Metrics

COMET



learn to rank better output

Learned Metrics

COMET

Table 1: Kendall’s Tau (τ) correlations on language pairs with English as source for the WMT19 Metrics DARR corpus. For BERTSCORE we report results with the default encoder model for a complete comparison, but also with XLM-RoBERTa (base) for fairness with our models. The values reported for YiSi-1 are taken directly from the shared task paper (Ma et al., 2019).

Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	0.615	0.378
COMET-RANK	0.603	0.427	0.664	0.611	0.693	0.665	0.580	0.449

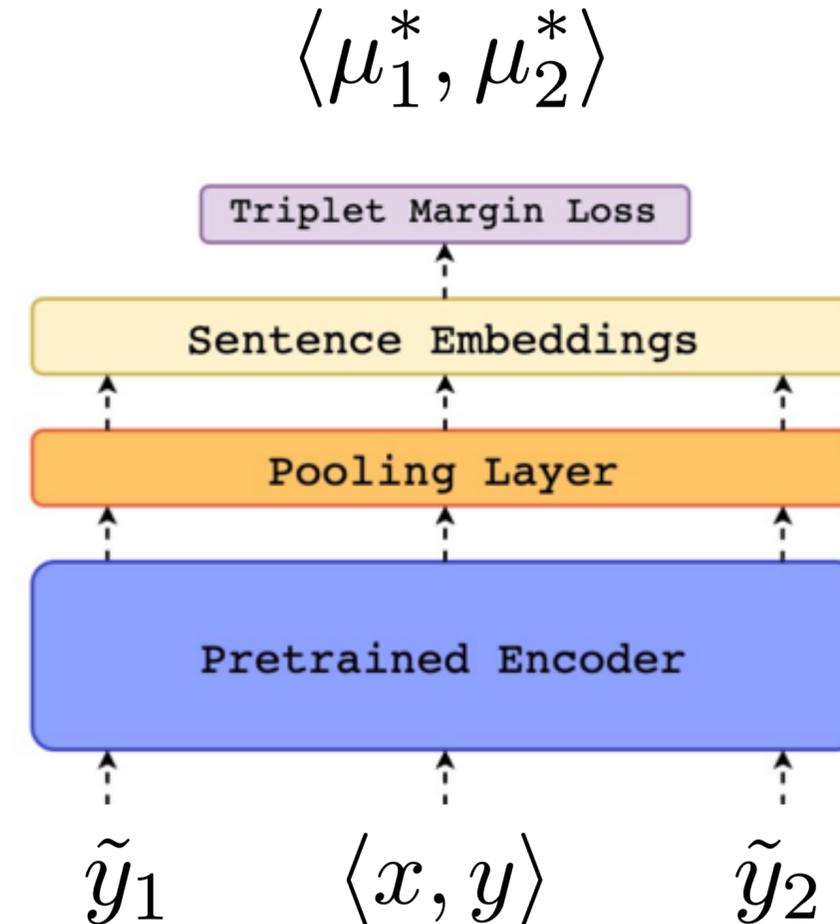
directly model
human ratings
works

modeling human
preferences tends
to work better

Learned Metrics

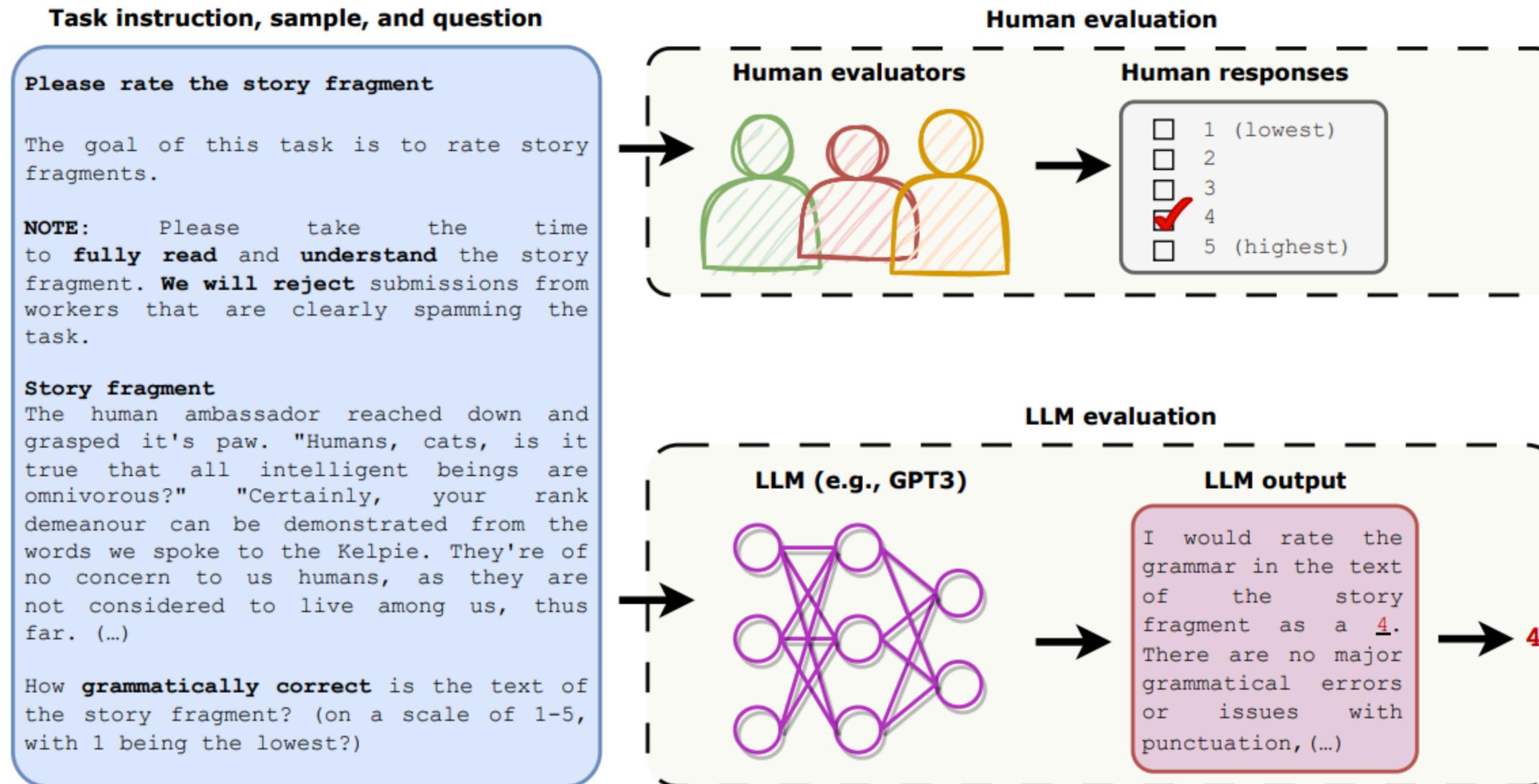
COMET

- advantages
 - relaxes exact match
 - incorporates semantic similarity
 - directly modeling human
- disadvantages
 - dependent on embedding model
 - task-specific
- developed for machine translation but can be used in other tasks



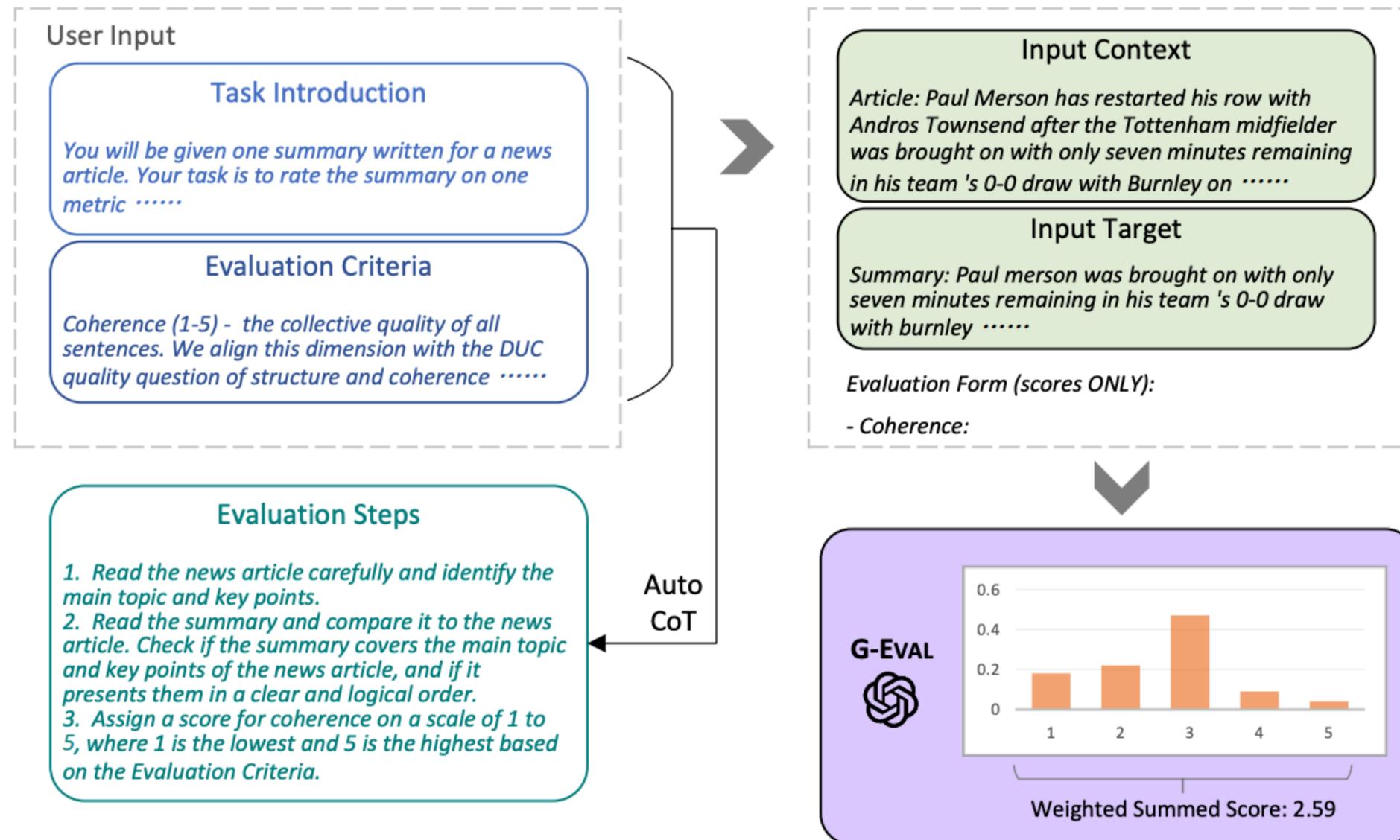
Zero-shot Metrics

Autorater, LLM-as-a-judge



Zero-shot Metrics

Autorater, LLM-as-a-judge



Zero-shot Metrics

Autorater, LLM-as-a-judge

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ								
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	<i>0.313</i>	0.361	<i>0.344</i>	0.339	<i>0.323</i>	0.327	<i>0.288</i>	0.346	<i>0.317</i>
G-EVAL-4	0.582	0.457	0.507	0.425	0.506	0.455	0.547	0.433	0.514	0.418
- Probs	0.560	<i>0.472</i>	0.501	<i>0.459</i>	0.505	<i>0.473</i>	0.511	<i>0.444</i>	0.502	<i>0.446</i>
- CoT	0.564	0.454	0.493	0.413	0.483	0.431	0.538	0.427	0.500	0.407
- Description	0.513	0.424	0.421	0.344	0.447	0.373	0.479	0.388	0.479	0.377

Table 1: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. G-EVAL without probabilities (*italicized*) should not be considered as a fair comparison to other metrics on τ , as it leads to many ties in the scores. This results in a higher Kendall-Tau correlation, but it does not fairly reflect the true evaluation ability. More details are in Section 4.

Zero-shot Metrics

Autorater, LLM-as-a-judge

role **You are a search quality rater evaluating the relevance of web pages.** Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

2 = highly relevant, very helpful for this query
 1 = relevant, may be partly helpful but might contain other irrelevant content
 0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

Query
 A person has typed [query] into a search engine.
 They were looking for: *description narrative*

Result
 Consider the following web page.

—BEGIN WEB PAGE CONTENT—
page text
 —END WEB PAGE CONTENT—

Instructions
 Split this problem into steps:
 Consider the underlying intent of the search.

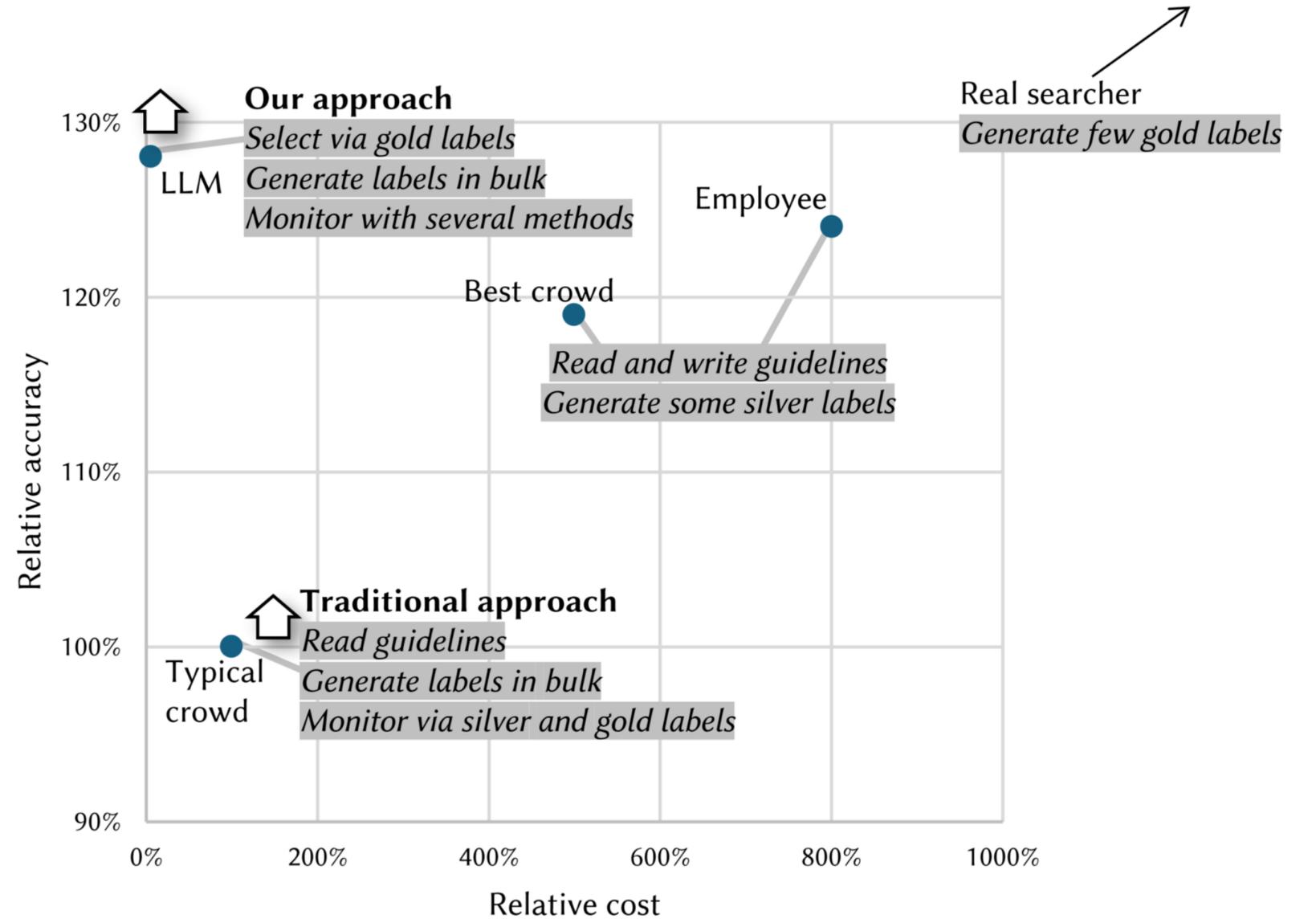
aspects **Measure how well the content matches a likely intent of the query (M).**

aspects **Measure how trustworthy the web page is (T).**

multiple **We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.**

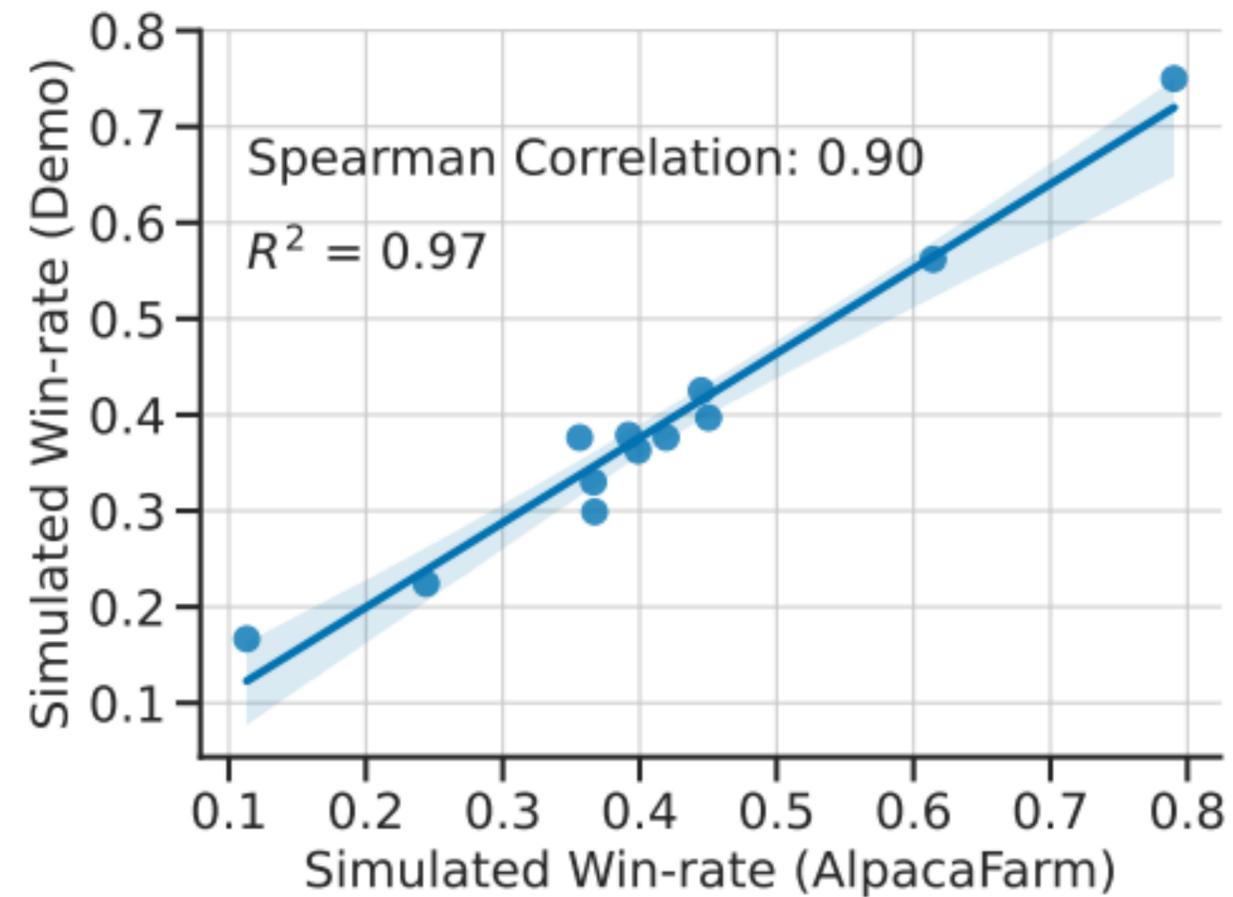
Produce a JSON array of scores without providing any reasoning. Example: [{"M": 2, "T": 1, "O": 1}, {"M": 1 . . . }

Results
 [{"



Zero-shot Metrics

Autorater, LLM-as-a-judge



Zero-shot Metrics

Rubrics

- rubrics are structured decompositions of an annotation task
- in human rating, we ask people to rate system outputs on multiple dimensions
- in autorating, we ask LLMs to rate system outputs on multiple dimensions
- learned components
 - per-criterion rating
 - aggregation

Query: When did rock'n'roll begin?

Grading rubric questions:

	r1 Which musicians or bands are considered pioneers of rock n roll?	r2 What were the major influences that led to the emergence of rock n roll?	r3 Are there any specific events or performances that marked the beginning of rock n roll?	r4 Did the development of new technologies have an impact on the birth of rock n roll?	r5 Is there a general consensus among music historians regarding the exact start of rock n roll?	Rubric relevance label:
p1 Passages: ... listens to pop, rock, soul and whatnot. The rock and roll era began around 1950. It evolved from rhythm and blues in the 1940s. The name rock and roll was the title of a song by the Boswell Sisters in 1934 (YouTube), but that was a swing song.	4 Boswell Sisters	4 rhythm and blues	0	0	4 The rock and roll era began around 1950	4
p2 Definition of 'rock and roll'. rock and roll also rock'n'roll. uncountable noun. Rock and roll is a kind of popular music developed in the 1950s which has a strong beat and is played on electrical instruments Elvis Presley—the King of Rock and Roll the greatest rock 'n' roll band in the world. rock and roll or rock'n'roll.	5 Elvis Presley—the King of Rock and Roll	0	0	4 Rock and roll is played on electrical instruments	4 1950s	5
p3 But you could just as well take the release of " Rocket 88 " in 1951 as the beginning of the rocknroll era, because that was the first rocknroll song. So far we have the years 1951 and 1952 to choose from. Maybe a certain date will turn up while we examine why rocknroll emerged at all.	0	0	4 "Rocket 88"	0	4 1951 and 1952	4

Zero-shot Metrics

Rubrics

Q_1 – In terms of naturalness and tone of the **assistant utterances**, to what degree are they likely to be produced by an **intelligent human** in a conversation? Disregard whether they are grounded in the search results.

1. Unlikely.
2. Somewhat unlikely.
3. Somewhat likely.
4. Likely.

Q_2 – If the references are provided, to what degree user’s questions can be answered or resolved using the references? The assistant’s responses should not impact your response to this question. **If no references are provided in the conversation, please write “NA” for this question.**

1. None of the questions that user has asked could be answered using the reference documents.
2. Less than half of documents that user has asked could be answered using the reference document.
3. Half or more than half of the questions that user has asked could be answered using the reference documents.
4. All the questions the user has asked could be answered with the reference documents.

Q_3 – Independent of what sources are cited in the conversation, to what degree the claims made by the assistant are followed by a citation. **If no references are provided in the conversation, please write NA.**

1. None of the claims are followed by a citation.
2. Less than half of the claims are followed by a citation.
3. Half, or more than half of the claims are followed by a citation.
4. All claims are followed by a citation.

criteria

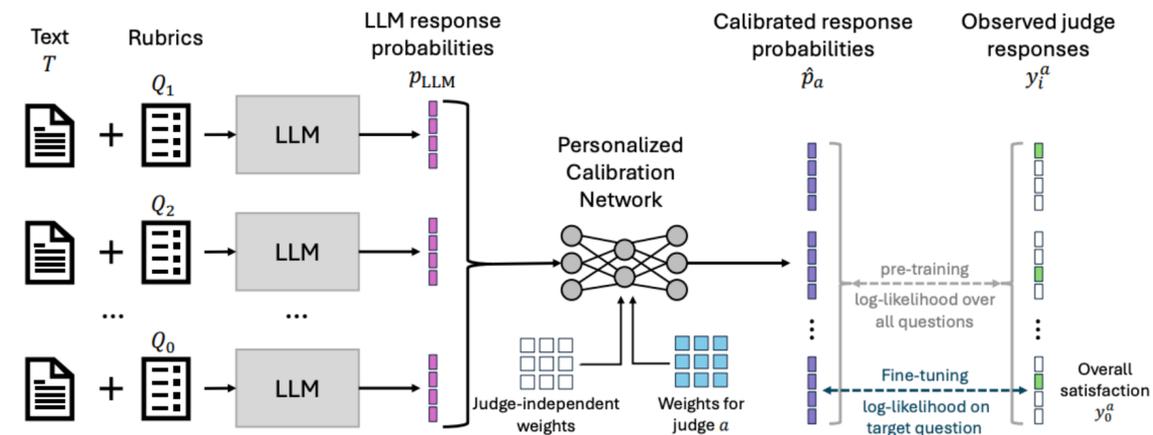
Q_0 – Imagine you are the user who had this conversation with the assistant. All in all, how you would rate your overall satisfaction while interacting with the assistant? The higher the rating, the better the experience.

1. 1
2. 2
3. 3
4. 4

holistic question

Zero-shot Metrics

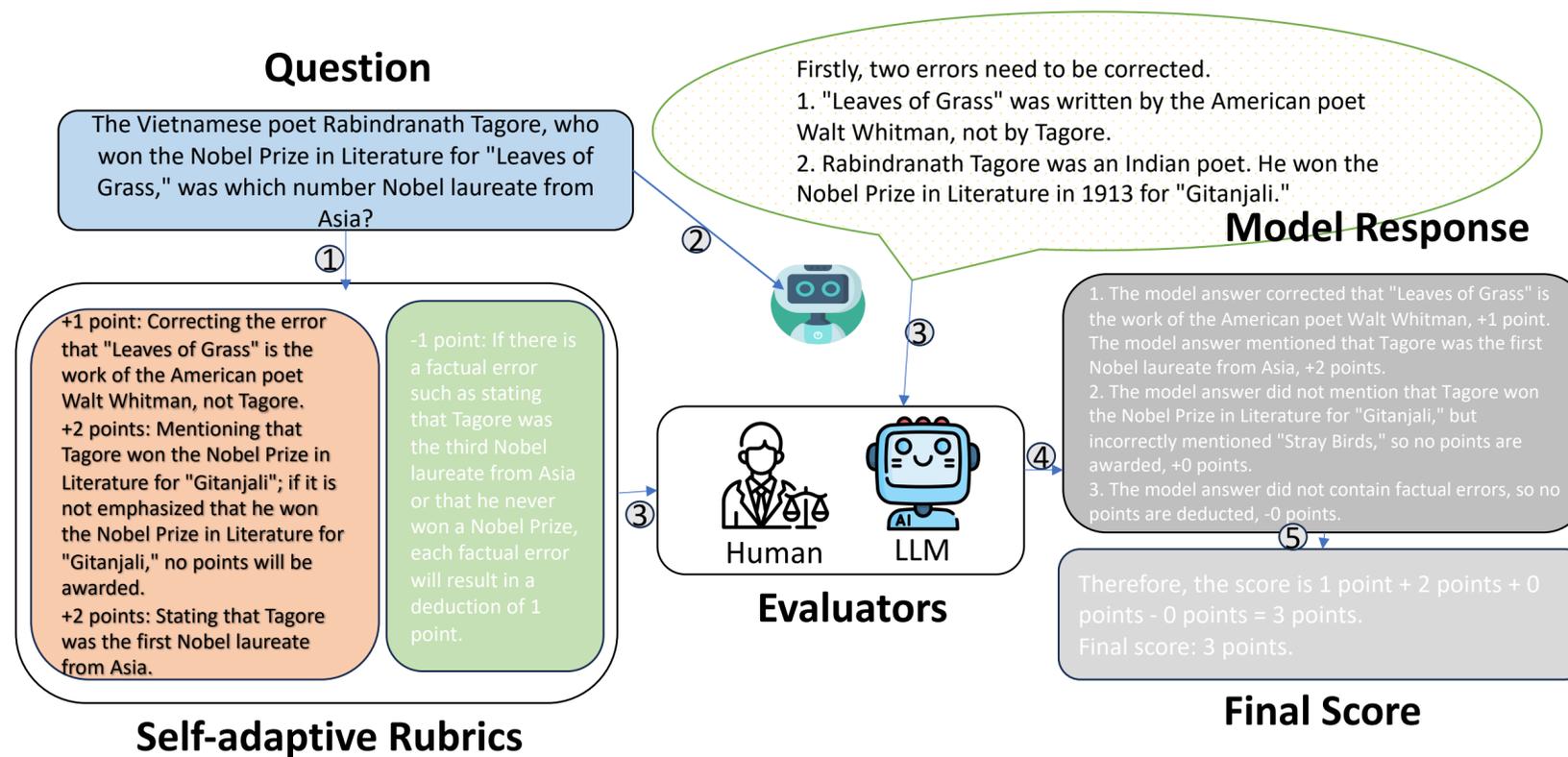
Rubrics



Model	Synthetic Conversations				Real Human-Agent Conversations			
	RMSE ↓	P's ρ ↑	S's ρ ↑	K's τ ↑	RMSE ↓	P's ρ ↑	S's ρ ↑	K's τ ↑
1 Random Eval	1.499	0.002	-0.003	-0.003	1.427	0.011	0.006	0.005
2 Argmax LLM Q_0	0.984 ¹	0.153 ¹	0.161 ¹	0.147 ¹	1.186 ¹	0.106 ¹	0.123 ¹	0.120 ¹
3 Expected LLM Q_0	0.856 ¹²	0.182 ¹	0.217 ¹	0.168 ¹	0.901 ¹²	0.143 ¹	0.141 ¹	0.138 ¹
4 Calibrated LLM Q_0	0.801 ¹²³	0.198 ¹²	0.196 ¹	0.193 ¹²	0.784 ¹²³	0.211 ¹²³	0.218 ¹²³	0.192 ¹²³
5 FActScore (Min et al., 2023)	–	0.204 ¹²	0.211 ¹	0.200 ¹²	–	0.216 ¹²³	0.218 ¹²³	0.207 ¹²³
6 LLM-RUBRIC	0.396 ^{1234e}	0.401 ^{12345e}	0.398 ^{12345e}	0.393 ^{12345e}	0.422 ¹²³⁴	0.350 ¹²³⁴⁵	0.347 ¹²³⁴⁵	0.331 ¹²³⁴⁵
a Oracle	0.237 ^{*bcdef}	0.611 ^{*bcdef}	0.626 ^{*bcdef}	0.605 ^{*bcdef}	0.289 ^{*bcd}	0.717 ^{*bcd}	0.711 ^{*bcd}	0.675 ^{*bcd}
b w/o LLM probs	0.276 ^{*cef}	0.551 ^{*cef}	0.548 ^{*cef}	0.533 ^{*cef}	0.357 ^{*c}	0.625 ^{*c}	0.629 ^{*c}	0.599 ^{*c}
c w/o Personalized Calibration	0.401 ^e	0.476 ^{*e}	0.471 ^{*e}	0.468 ^{*e}	0.389 [*]	0.582 [*]	0.587 [*]	0.565 [*]
d ↪ + Personalized isotonic regress	0.273 ^{*cef}	0.521 ^{*cef}	0.526 ^{*cef}	0.519 ^{*cef}	0.302 ^{*bc}	0.650 ^{*bc}	0.653 ^{*bc}	0.644 ^{*bc}
e Depersonalized Oracle	0.492	0.362	0.355	0.338	–	–	–	–
f ↪ + Personalized isotonic regress	0.321 ^{*ce}	0.482 ^{*e}	0.485 ^{*e}	0.477 ^{*e}	–	–	–	–

Zero-shot Metrics

Self-Adaptive Rubrics



Type	Question-level				Model-level			
	GSB	ACC	ACC(t)	pearson	GSB	ACC	ACC(t)	pearson
XDG	0.952	0.590	0.794	0.738	0.952	0.590	0.794	0.380
GPT-3.5	0.829	0.422	0.663	0.566	0.829	0.422	0.663	0.566
GPT-4	0.952	0.654	0.855	0.822	0.952	0.654	0.855	0.822

Table 1: Out of distribution evaluation performance in both model-level and question-level.

poll: name an issue with LLM-as-a-judge.

Simulating populations

- so far, we have focused on evaluation as simulation where the inputs may be constructed or drawn from some distribution (e.g., logs).
- can we simulate the inputs too?
- increasing body of work on how we simulate people with LLMs, including the inputs they provide.



Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Simulating populations

DataMorgana

- use a prompt to generate a class of inputs.
- can be parameterized by topic, user demographics
- if we also use auto-rating, can have closed loop simulated evaluation

```
You are a user simulator that should generate a question to start a conversation.

The question must be about facts discussed in the document you will now receive.
Return only the question and its answer without any preamble.
Write the question-answer pair in the following JSON format:
{"question": <question>, "answer": <answer>}.

### The generated question should be about facts from the following document:
[document (d_i)]

### The generated question must reflect a user with
the following characteristics:
- [description of user category 1 (u_1)]
- [description of user category 2 (u_2)]
...

NOTE: you must use this information only when generating the question.
Instead, while answering the question you must ignore all the user characteristics.

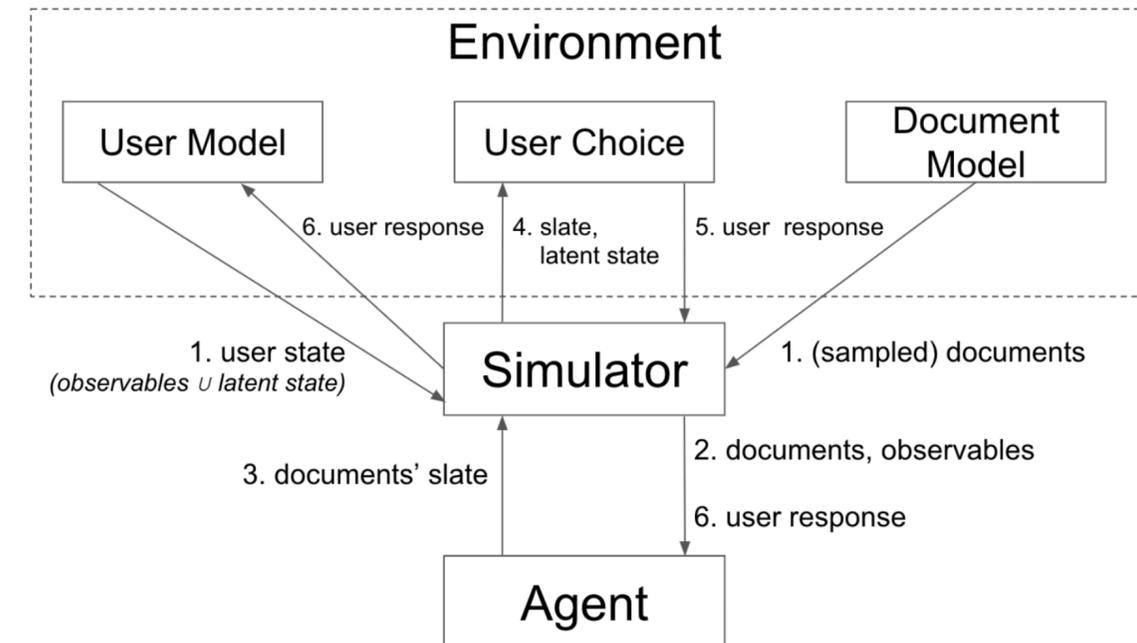
### The generated question must have the following characteristics:
- The question must be understandable by a reader who does not have access to the document
and does not even know what the document is about.
Therefore, never refer to the author of the document or the document itself.
- The question must include all context needed for comprehension.
- The question must be answerable using solely the information presented in the document.
- [description of question category 1 (c_1)]
- [description of question category 2 (c_2)]
...

### The answer to the generated question must have the following characteristics:
- It must be very similar to the document in terms of terminology and phrasing.
- It should only contain claims that directly appear in the document
or that are directly deducible from it.
- It must be understandable by a reader who does not have access to the document.
Therefore, never refer to the author of the document or the document itself.
- It must not assume or contain any information about the user,
unless it is explicitly revealed in the question.
```

poll: name an issue with closed
loop simulated evaluation.

Simulating marketplaces

- so far, we have focused on evaluation that looks at systems in isolation
 - a rating for a single output
- when multiple systems are directly compared, it is at an input level
 - arena-style evaluation compares two system outputs
- in reality, users make choices between systems, not outputs
 - e.g., there are multiple chatbots; you are not making decisions for every turn.
 - users are stateful!



Summary

1 Token Level	Cross-entropy / Perplexity
2 Output Level	Rubric Scoring / BLEU / ROUGE
3 Task Level	Extrinsic evaluation, user simulation
4 User Level	User simulation, online user tracking
5 Population Level	User simulation, aggregate online user tracking
6 Market Level	Marketplace simulation