Measuring the Quality of Natural Language Generation Systems

SICSA/SIGGEN webinar

Sebastian Gehrmann | Google Research | gehrmann@google.com | @SebGehr

Background: What is Natural Language Generation?

An NLG system
with an explicit **communicative goal**

Natural Language - fluent, understandable,
in accordance with the communicative goal

$$f : x \rightarrow y$$

Structured or textual information
that defines the output space

(News) Summarization

Communicative Goal

Succinctly present the main ideas of an article

Input

A news article of about 600-800 words

Target

A 2-3 sentence summary

Challenges

- Identify and select "important" content
- Plan the summary structure
- Actualize the plan in natural language
- Do not hallucinate, i.e., generate ungrounded content

Written by
Oscar Holland, CNN
Junko Ogura, CNN

One of Japan's most distinctive works of contemporary architecture, the Nakagin Capsule Tower in Tokyo, will be demolished this month, according to the building's new owners.

The decision ends years of uncertainty surrounding the eye-catching structure, which once offered a futuristic vision of urban living but had recently fallen into disrepair.

Completed in 1972, the tower comprises 144 factory-built units arranged around two concrete cores. Each 10-square-meter (108-square-foot) "capsule" features a porthole-style window, with appliances and furniture built into the structure of each home.

A furnished capsule room inside Nakagin Capsule Tower. Credit: Carl Court/Getty Images

The building is considered a prime example of Metabolism, an architectural movement that emerged from the ruins of World War II with a radical new vision for Japan's cities. As well as embracing technology and mass production, the avant-garde group's members looked to nature for inspiration, with structural components treated like organic cells that could be "plugged" into a larger whole or later replaced.

Related:
The Japanese architects who treated buildings like living organisms

The building's designer, Kisho Kurokawa -- one of Metabolism's youngest adherents -- had originally envisaged the Tokyo tower's capsules being replaced every 25 years. But they instead grew dilapidated and outdated, with many of the apartments now sitting empty, used for storage and office space, or rented out to architecture enthusiasts on a short-term basis.

In 2007, the owners' association voted to sell the tower to a property developer that intended to demolish and replace it. But the firm filed for bankruptcy during the 2008 recession, and the site's fate was thrown into years-long limbo.

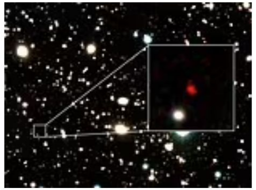
Preservationists hope some of the the 108-square-foot capsules can be saved and repurposed or acquired by museums. Credit: Carl Court/Getty Images

Owners again agreed to sell in 2021, and the building was acquired by a group of real estate firms operating under the name Capusule Tower Building (CTB). A spokesperson for the joint venture, Takashi Shindo, told CNN over the phone that the last residents moved out last month, with demolition scheduled to begin April 12.

Preservationists had long expressed hope that the building might be saved -- including Kurokawa, before his death in 2007. Petitions and campaigns have called for the structure to be protected as an example of Japan's architectural heritage. (Although the Metabolism movement proved influential, very few of its proposals were ever realized, making Nakagin Capsule Tower a rare living example of the group's philosophy.)

The organization behind the conservation campaign, the Nakagin Capsule Tower Building Preservation and Regeneration Project, asked city authorities to intervene -- and even considered applying for protected status with UNESCO. But neither approach proved successful, according to project member Tatsuyuki Maeda, who acquired 15 of the capsules between 2010 and the building's sale last year.

Astronomers discover the most distant galaxy yet - a whopping 13.5 billion light years from Earth



Astronomers from the University of Tokyo have discovered the most distant galaxy ever found, a whopping 13.5 billion light years from Earth.

Judy Garland

Garland in the 1940s

Born	Frances Ethel Gumm June 10, 1922 Grand Rapids, Minnesota, U.S.[1]
Died	June 22, 1969 (aged 47) London, England
Resting place	Hollywood Forever Cemetery
Occupation	Actress · singer · dancer · vaudevillian · television and radio presenter
Years active	1924–1969
Height	4 ft 11½ in (151 cm)
Political party	Democratic
Spouse(s)	David Rose (m. 1941; div. 1944) Vincente Minnelli (m. 1945; div. 1951) Sidney Luft (m. 1952; div. 1965) Mark Herron (m. 1965; div. 1969) Mickey Deans (m. 1969)
Children	3, including Liza Minnelli and Lorna Luft

mmarization

Judy Garland (born **Frances Ethel Gumm**; June 10, 1922 – June 22, 1969) was an American actress and singer. She is widely known for playing the role of Dorothy Gale in *The Wizard of Oz* (1939).[2][3] With a career spanning 45 years, she attained international stardom as an actress in both musical and dramatic roles, as a recording artist, and on the concert stage. Renowned for her versatility, she received an Academy Juvenile Award, a Golden Globe Award, and a Special Tony Award.[4][5][6] Garland was the first woman to win the Grammy Award for Album of the Year, which she won for her 1961 live recording titled *Judy at Carnegie Hall*.[7]

ne plan in natural language
ucinate, i.e., generate
d content

Biography Generation

Communicative Goal
Generate a brief description of a person grounded in descriptive attributes

Input
Key-Value attribute pairs

Target
A ~1 paragraph biography

Challenges

- Plan the biography structure to incorporate the entirety of the input attributes
- Actualize the plan in natural language
- Do not hallucinate, i.e., generate ungrounded content

Agenda

01 How are NLG Systems evaluated?

02 Common Pitfalls in NLG Evaluation

03 Implementing Best Practices in GEMv2

What should our results tell us about a model?

Researcher:

- Can we **confirm the claims** made about the model performance?
- Is this the **currently best approach** to address the particular problem?
- What are **shortcomings** future researchers should work on?

What should our results tell us about a model?

Researcher:

- Can we confirm the claims made about the model performance?
- Is this the currently best approach to address the particular problem?
- What are shortcomings future researchers should work on?

Product Manager:

- Does the model meet the **quality requirements** we set?
- What are **catastrophic failures** of a model?
- How does the model perform on **“real-world” data**?
- How is the performance on **different user personas**?

...

What do we want to measure?

There is no equivalent of accuracy or F1 for NLG. We could measure the following aspects...

- Fulfilling a **communicative goal**
- Remaining **faithful** to the input information
- **Grammaticality**, **fluency** and **naturalness**
- **Readability** and **simplification** (structure, content)
- **Compactness** of summarization with correct **focus** and **non-redundancy**
- Intra- and inter-sentential/dialogue turn **cohesion**
- **Robustness** to shifts in the data distribution
- **Diversity** in repeated interactions

Some goals are **task specific** and some are more general.

Progress on goal A could lead to degradation on goal B.

There is no one-size-fits-all evaluation.

1

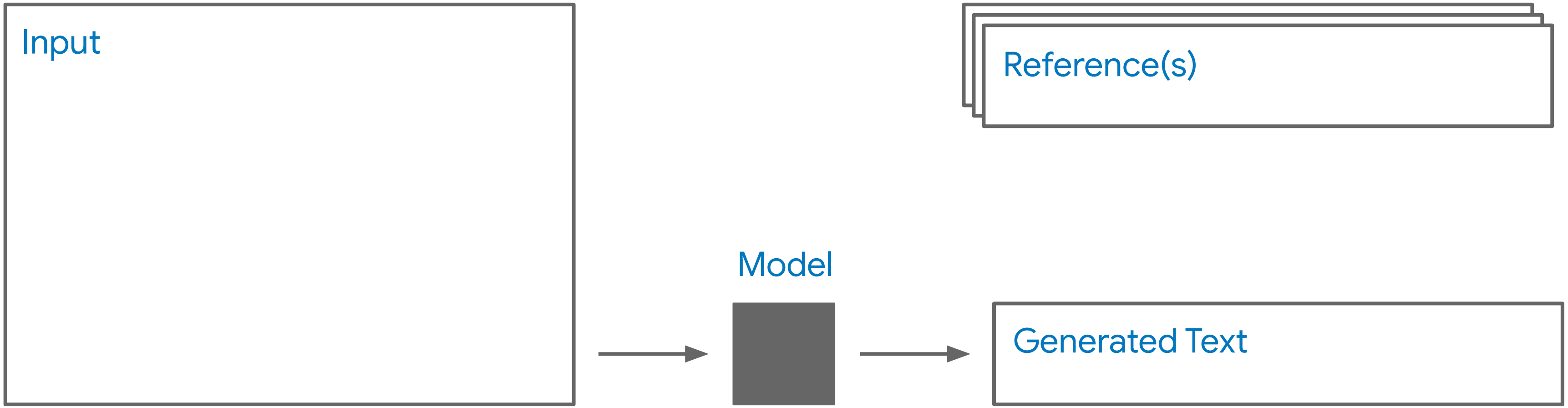
Automatic Evaluation

2

Human Evaluation

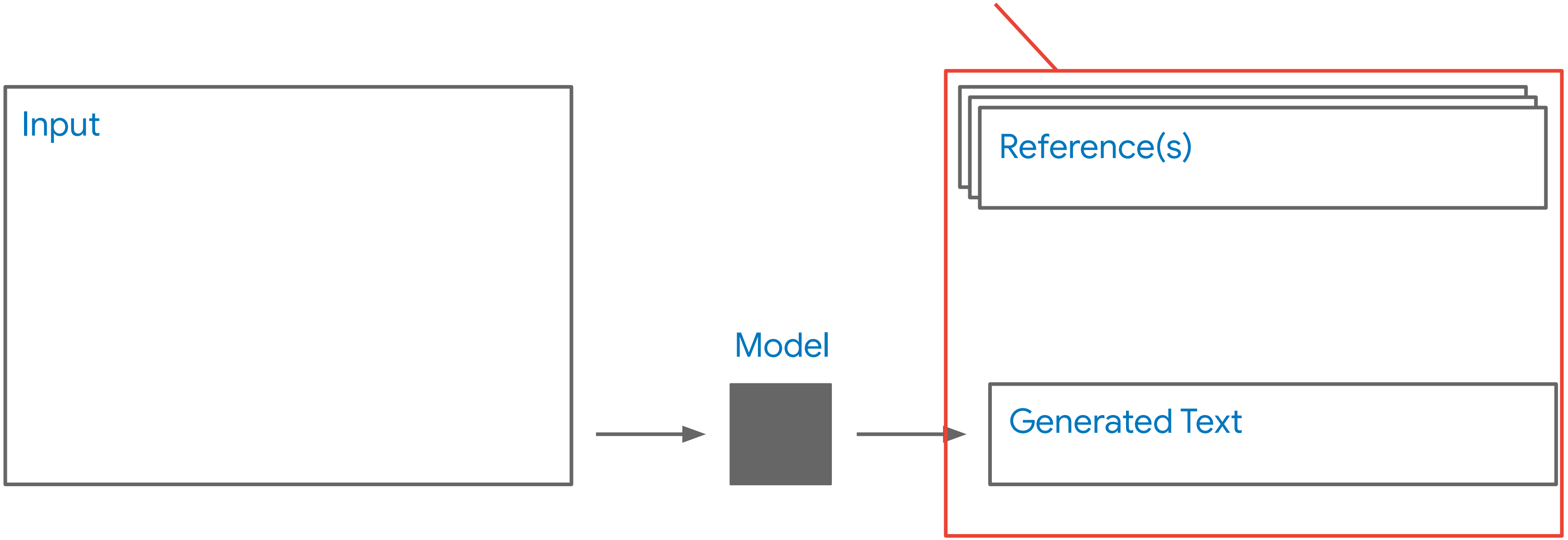
3

Evaluation Suites



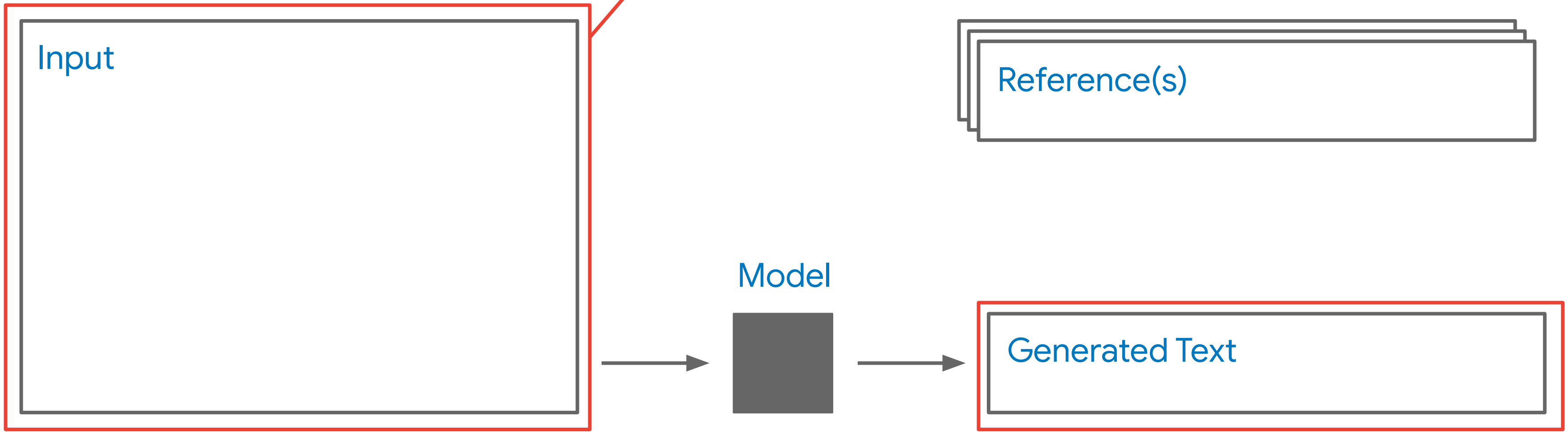
Automatic Metrics

Assumption: Higher similarity to human-written references means that the system is better

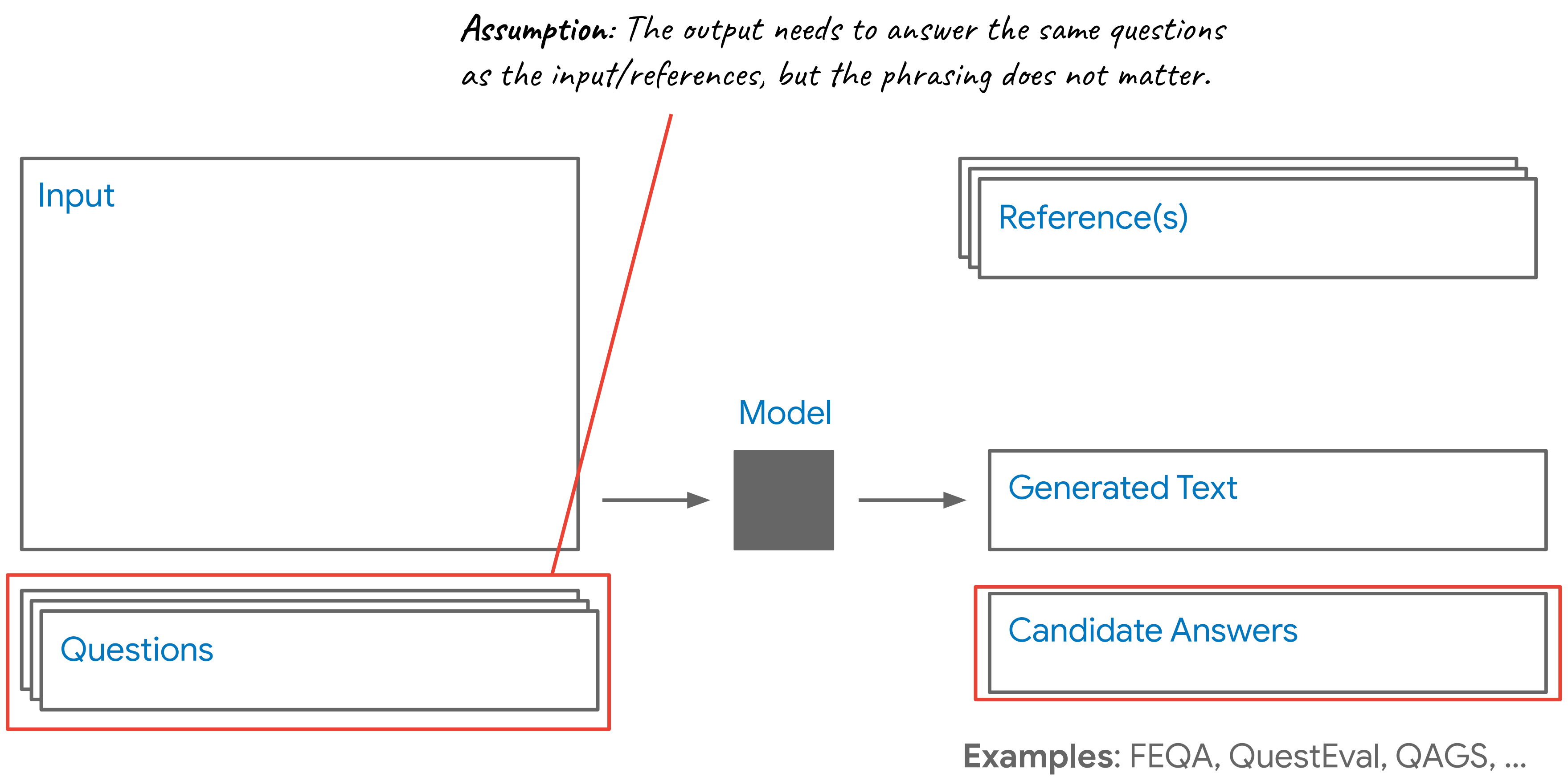


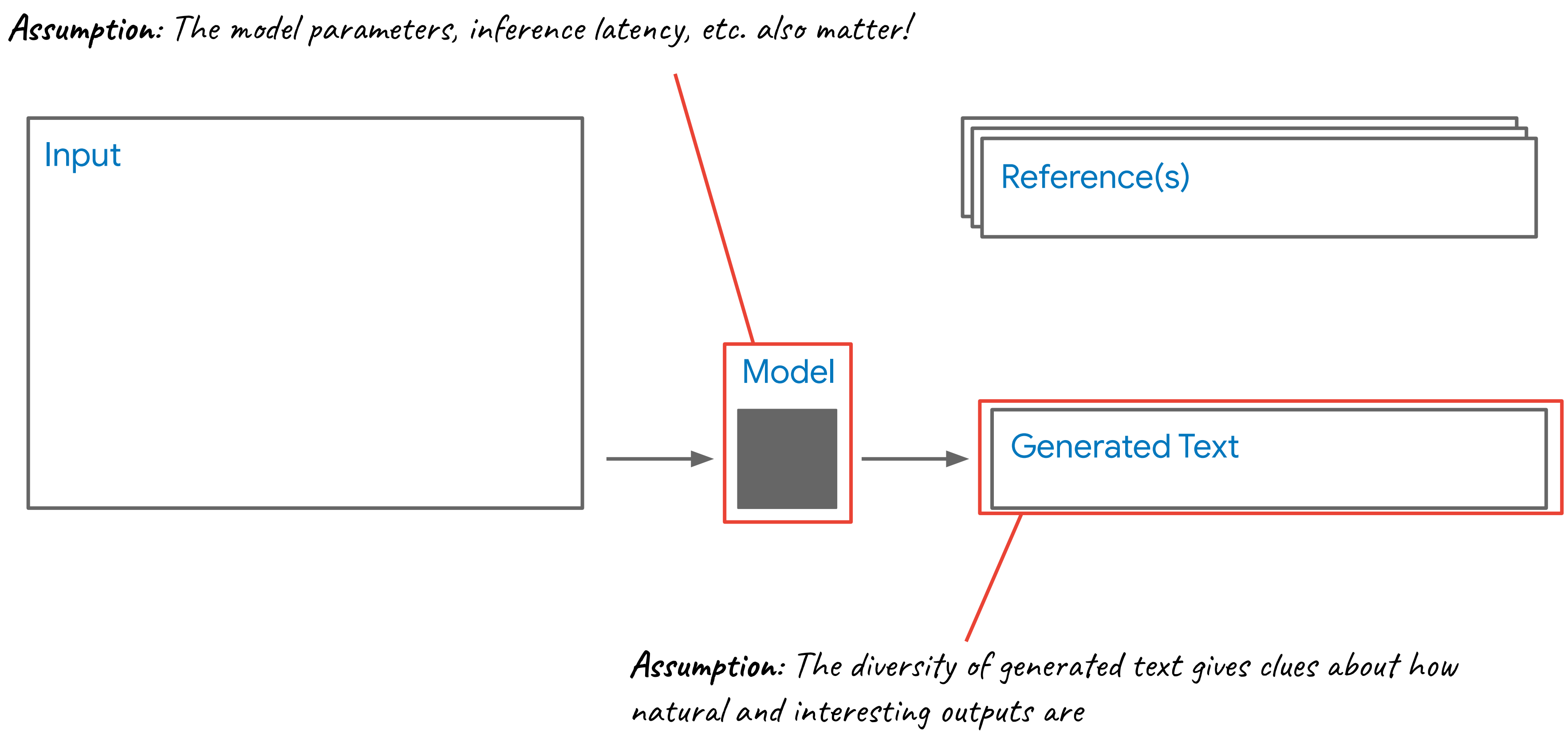
Lexical Similarity: ROUGE, BLEU, ...
Semantic Similarity: BERT-score, BLEURT, ...

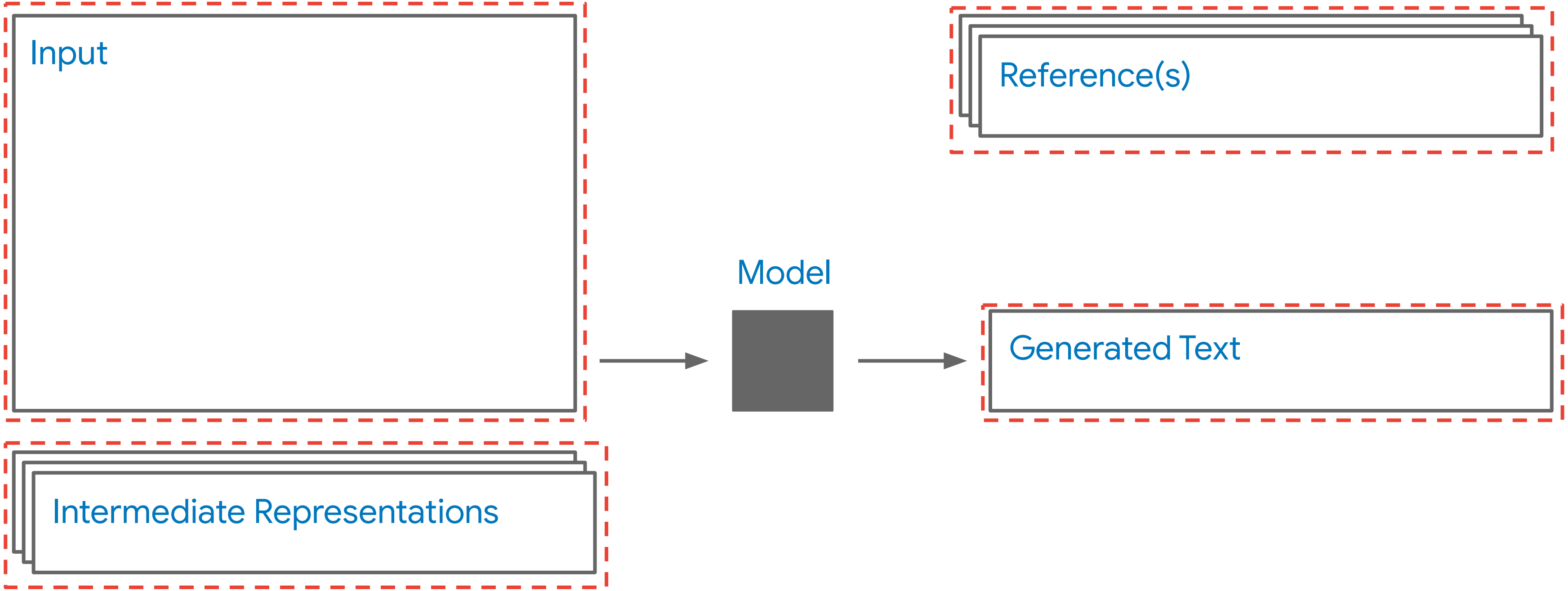
Assumption: The input itself already has all the necessary information. We can use a second model to predict quality.



Examples: COMET-QE, YiSi-2, NLI models

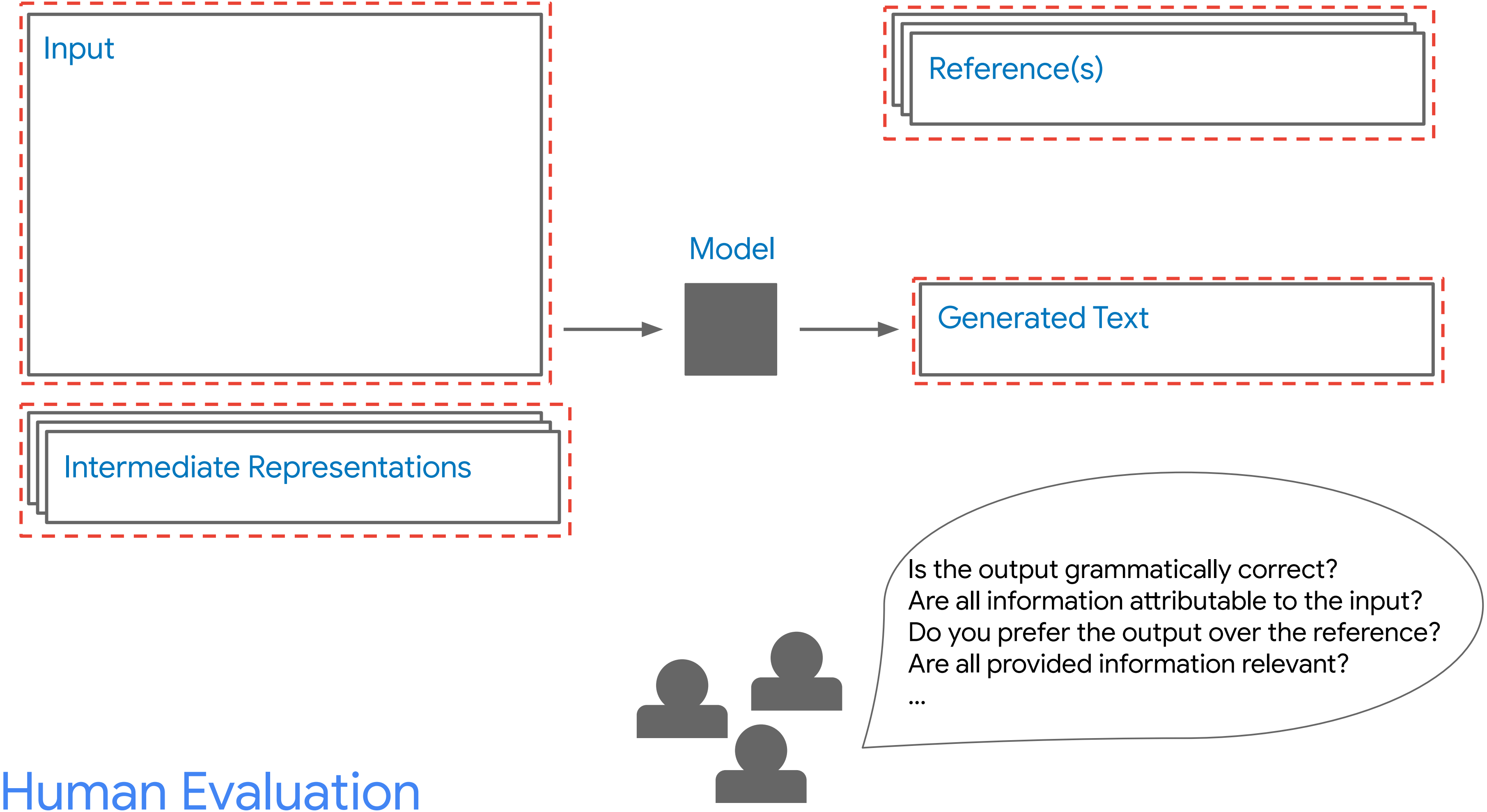

Assumption: The output needs to answer the same questions
as the input/references, but the phrasing does not matter.
Input
Reference(s)
Model
Generated Text
Questions
Candidate Answers
Examples: FEQA, QuestEval, QAGS, ...


Assumption: The model parameters, inference latency, etc. also matter!
Input
Reference(s)
Model
Generated Text
Assumption: The diversity of generated text gives clues about how natural and interesting outputs are

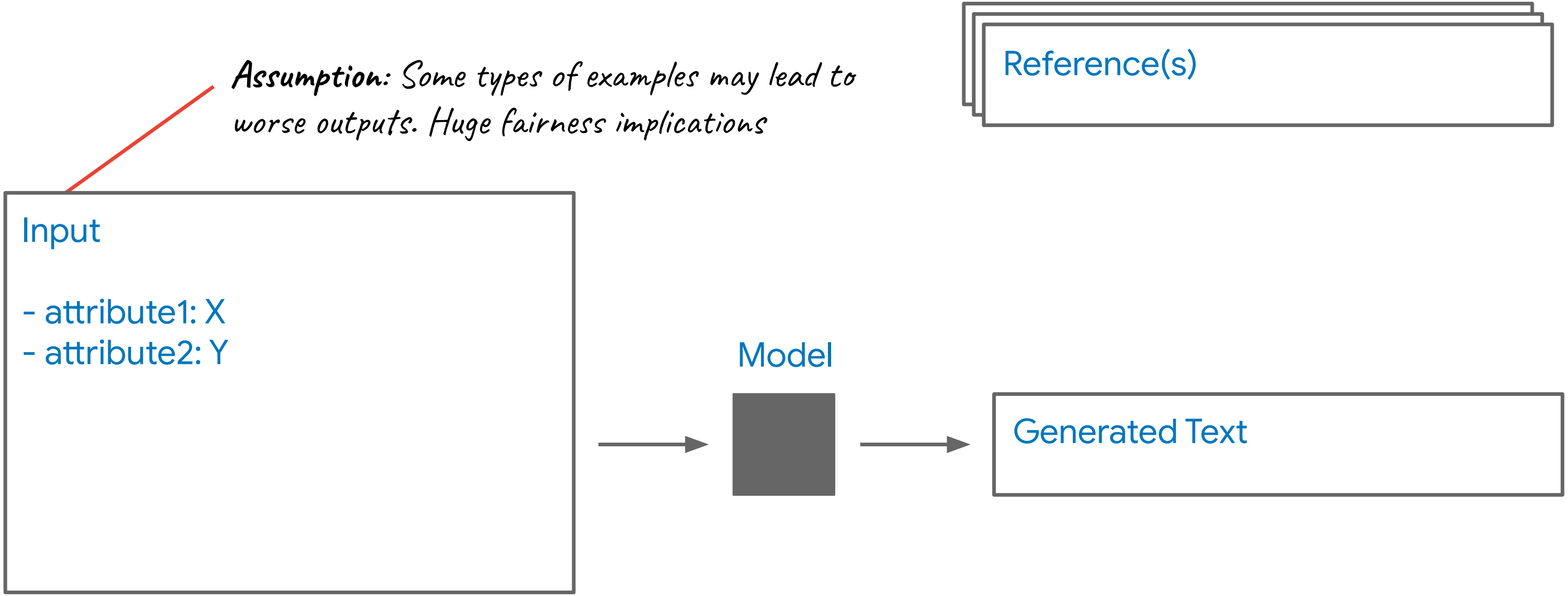


The same diversity of information to take into account exists for human evaluations.

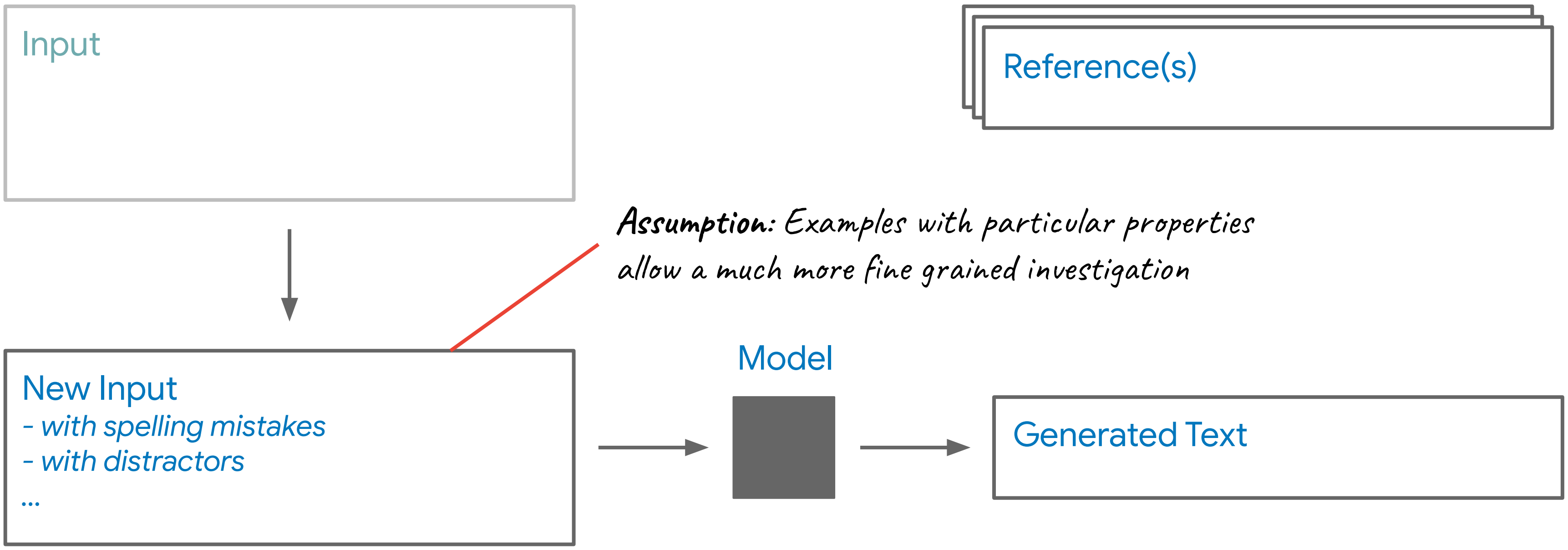
Human Evaluation



Human Evaluation

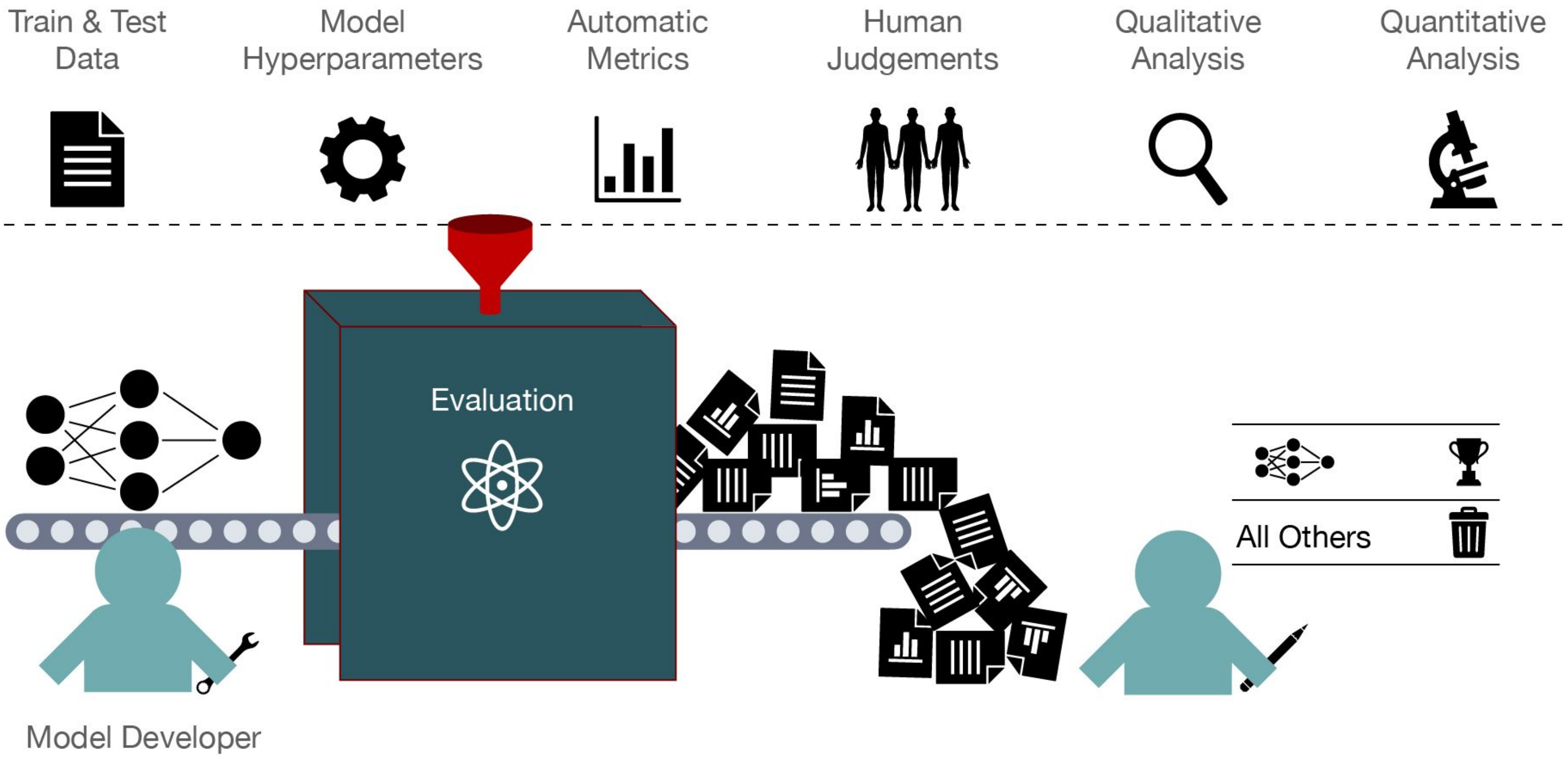


Evaluation Suites - Subpopulations



Evaluation Suites - Challenge Sets

There is no one-size-fits-all evaluation and the possibilities are limitless.



Gehrmann, Clark, and Sellam, 2022

Common Pitfalls in NLG Evaluation

What should our results tell us about a model?

Researcher:

- Do the results confirm the claims made about the model performance?
- **Is this the currently best approach to address the particular problem?**
- What are shortcomings future researchers should work on?

Product Manager:

- Does the model meet the quality requirements we set?
- What are catastrophic failures of a model?
- How does the model perform on “real-world” data?

...

What should our results tell us about a model?

Researcher:

- Do the results confirm the claims made about the model performance?
- **Is this the currently best approach to address the particular problem?**
- What are shortcomings future researchers should work on?

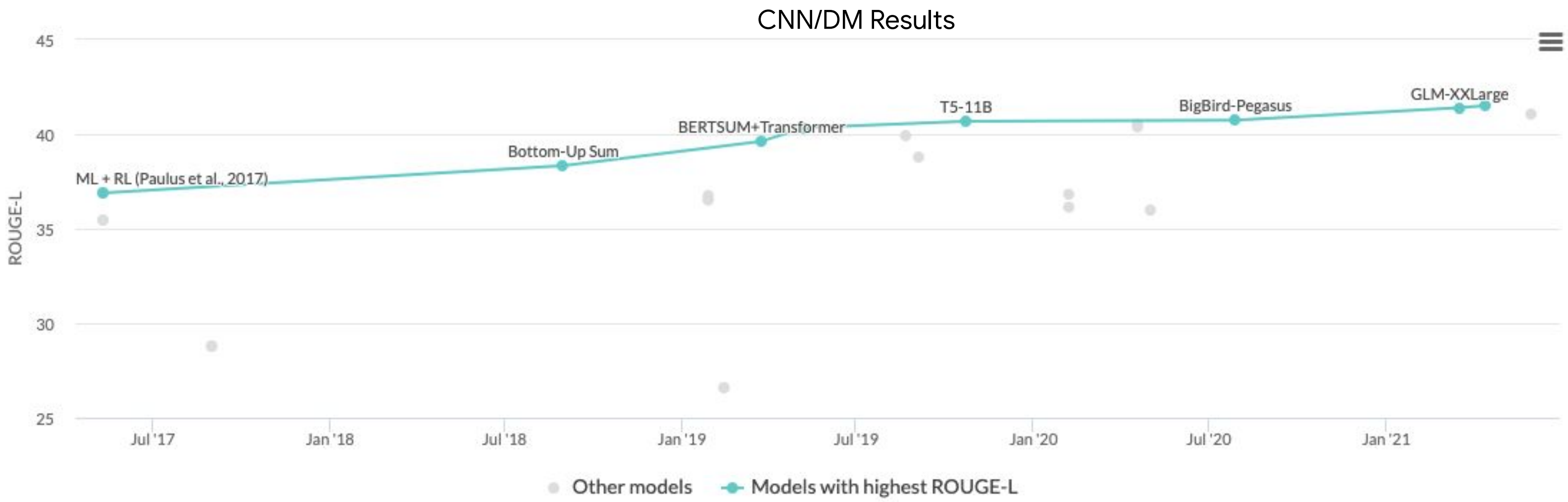
Product M

- Does
- What
- How does the model perform on “real-world” data?

...

48% of NLG papers published at *CL conferences in 2021 make claims about a systems overall “quality”.

Gehrmann, Clark, and Sellam, 2022



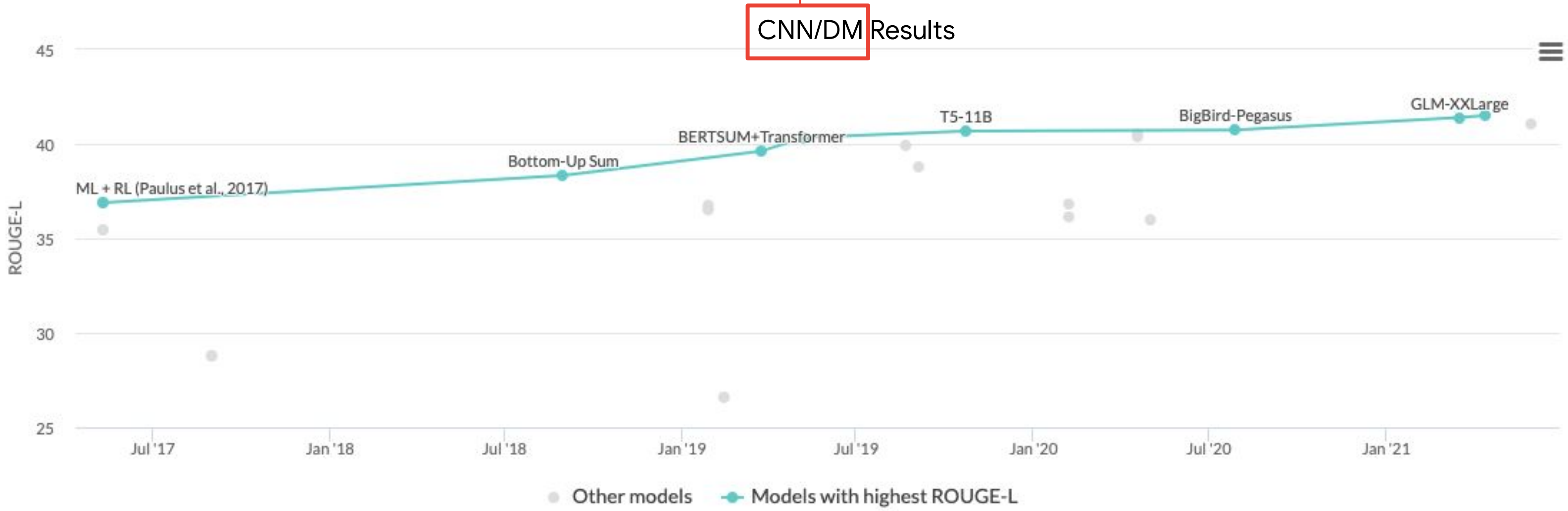
Measuring ROUGE-L on CNN/DM is the de-facto summarization benchmark.

- 100% of summarization papers report ROUGE, 69% report **only** ROUGE
- Together, CNN/DM and XSum are used by 40%+ of papers

https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Gehrmann, Clark, and Sellam, 2022

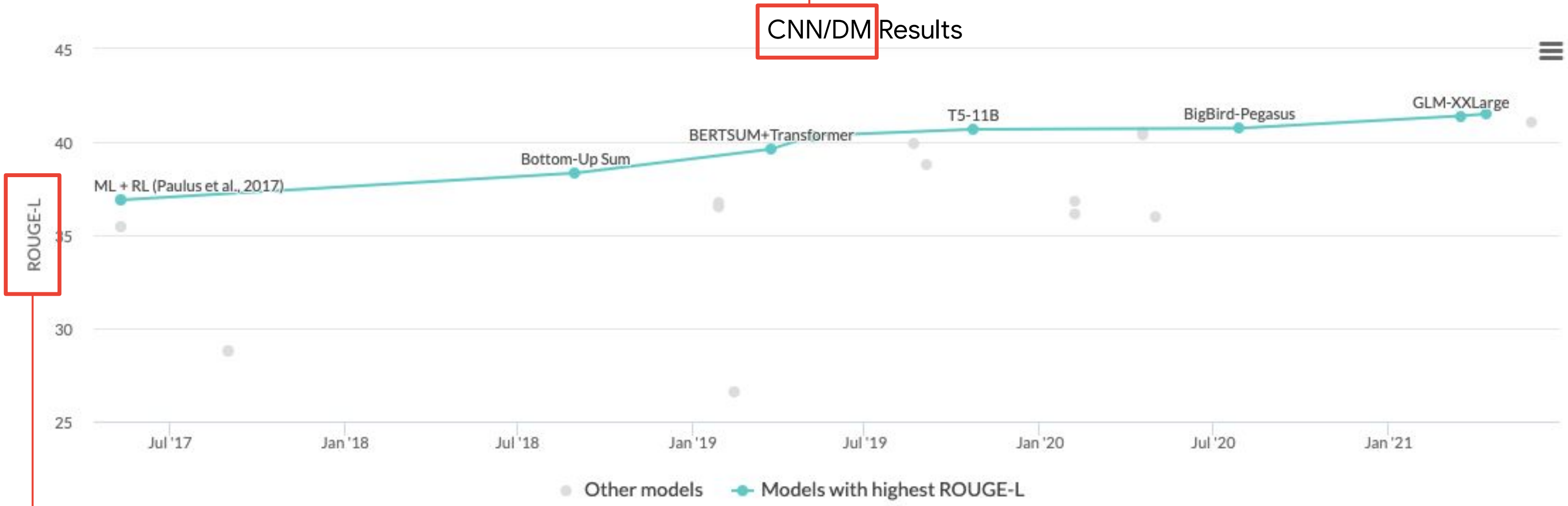
... is an English-only corpus
... Its references were never designed to be a summary
⟶ First three sentences are rated as a better one
⟶ References contain non-attributable facts



https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Gehrmann, Clark, and Sellam, 2022

- ... is an English-only corpus
- ... Its references were never designed to be a summary
- ⟶ First three sentences are rated as a better one
- ⟶ References contain non-attributable facts

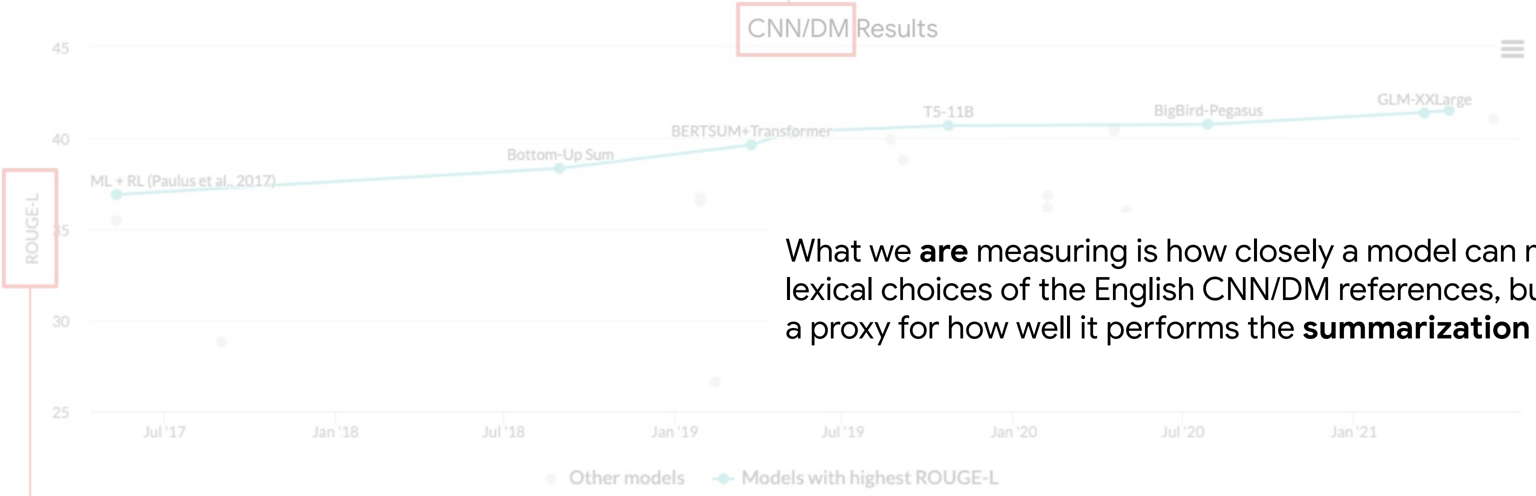


- ... is not the best possible ROUGE configuration.
- ... has low correlation with different quality aspects (e.g., faithfulness).
- ... Increases based on similarity to a reference and is thus confounded by its style and errors.
- ...

https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Gehrmann, Clark, and Sellam, 2022

... is an English-only corpus
... Its references were never designed to be a summary
⟶ First three sentences are rated as a better one
⟶ References contain non-attributable facts



What we **are** measuring is how closely a model can match the lexical choices of the English CNN/DM references, but this is not a proxy for how well it performs the **summarization task**.

... is not the best possible ROUGE configuration.
... has low correlation with different quality aspects (e.g., faithfulness).
... Increases based on similarity to a reference and is thus confounded by its style and errors.
...

https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Gehrmann, Clark, and Sellam, 2022

Popular metrics only assess form, not content.

Reference Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the Seattle band Silkworm.

Candidates	BLEU	ROUGE
Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the California band Grateful Dead.	**0.79**	0.77
Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer.	0.71	**0.79**
Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer from Seattle, Washington.	0.73	0.70

Dhingra et al., 2019

The correlation between human judgments and metrics is poor.

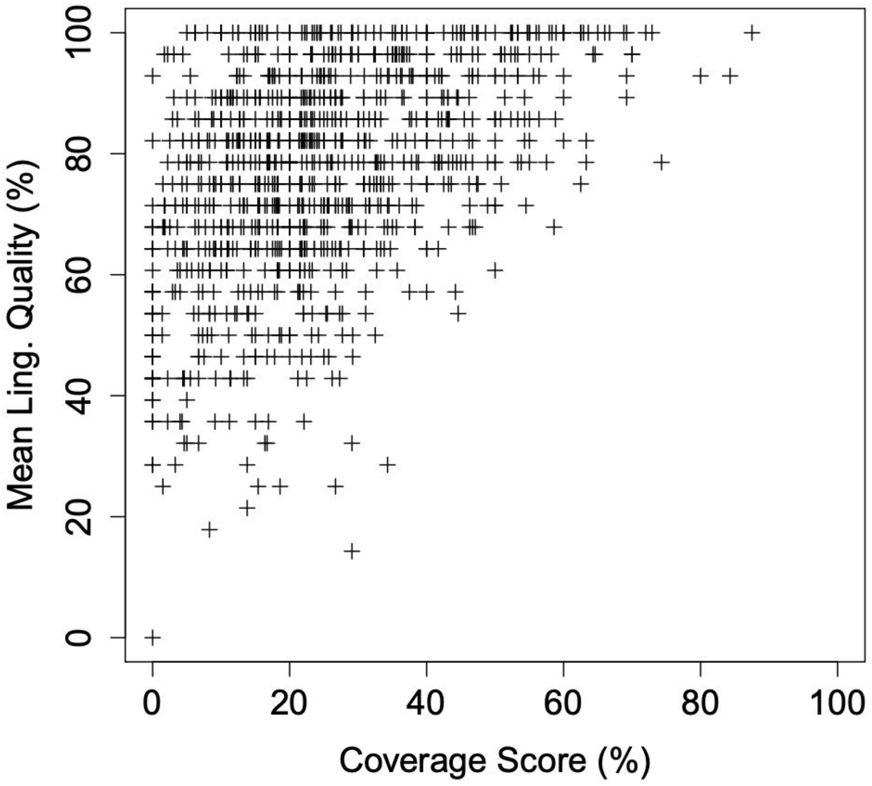
Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-1	0.2500	0.5294	**0.5240**	0.4118
ROUGE-2	0.1618	0.5882	0.4797	0.2941
ROUGE-3	0.2206	**0.7059**	**0.5092**	0.3529
ROUGE-4	**0.3088**	0.5882	**0.5535**	0.4118
ROUGE-L	0.0735	0.1471	0.2583	0.2353
BertScore-p	0.0588	−0.1912	0.0074	0.1618
BertScore-r	0.1471	**0.6618**	0.4945	0.3088
BertScore-f	0.2059	0.0441	0.2435	**0.4265**
BLEU	0.1176	0.0735	0.3321	0.2206
CHRF	**0.3971**	0.5294	0.4649	**0.5882**
CIDEr	0.1176	−0.1912	−0.0221	0.1912
METEOR	0.2353	**0.6324**	**0.6126**	**0.4265**

Kendall-Tau rank correlation of different metrics

Kryściński et al., Fabbri et al., 2021

A single metric is not enough.

Multiple studies found a lack of correlation between linguistic and content quality.



Pitler et al. (2010), Graham, 2015

We rely on **flawed** references.

Dataset	Coverage	Faithfulness	Fluency
WikiBio	0.44±0.007	$\mu = 2.5$ 	0.97

Model	ROUGE
WikiBio	
T5 (77M) beam	40.0
T5 (248M) beam	41.2
T5 (783M) beam	41.9

Yuan et al., 21

Unfaithful references can lead to **inflated performance numbers**.

Dataset	Coverage	Faithfulness	Fluency
WikiBio	0.44±0.007	$\mu = 2.5$	0.97
SynthBio	0.86±0.006	$\mu = 3.75$	0.97

Model	ROUGE
WikiBio	
T5 (77M) beam	40.0
T5 (248M) beam	41.2
T5 (783M) beam	41.9
SynthBio	
T5 (77M) beam	19.7
T5 (248M) beam	20.2
T5 (783M) beam	20.4

Yuan et al., 21

Lesson 1

Be mindful of what your metrics are (not) measuring

Lesson 2

Issues in the data will hide issues in the model

Lesson 1

Be mindful of what your metrics are (not) measuring

Can human evaluations solve this issue?

Lesson 2

Issues in the data will hide issues in the model

What is being measured

In 478 INLG papers, there were 71 different quality aspects. 👉

Often, the details are not provided:

- >50% missing definitions
- ~66% missing prompts/questions
- 20% missing criteria names

Criterion Paraphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
goodness of outputs in their own right (both form and content)	13
referent resolvability	11
usefulness (nonspecific)	11
appropriateness (content)	10
naturalness	10
user satisfaction	10
wellorderedness	10
correctness of outputs in their own right (form)	9
correctness of outputs relative to external frame of reference (content)	8
ease of communication	7
humanlikeness	7
appropriateness	6
understandability	6
nonredundancy (content)	6
goodness of outputs relative to system use	5
appropriateness (both form and content)	5

Table 4: Occurrence counts for normalised criterion names.

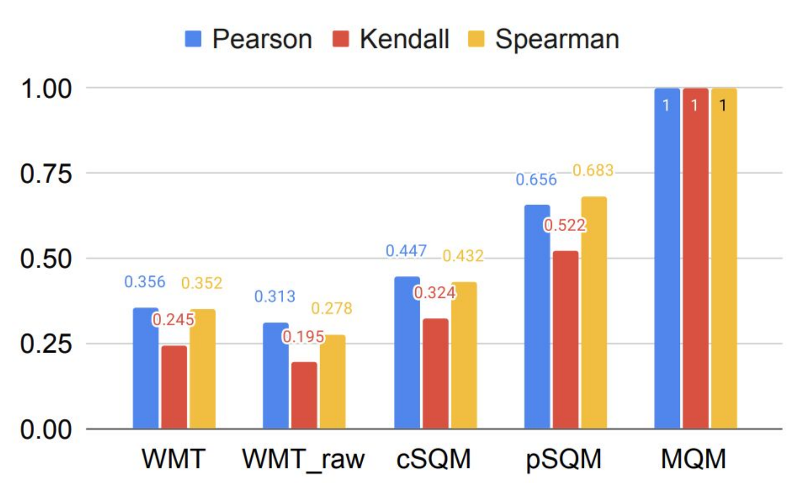
Howcroft et al., 2020

Who are the raters?

Agreement between ratings by linguists and those from crowdworkers can be extremely low.

Eval	Judges	Topics	Systems
TAC	0.28	$\boxed{0.40}$	0.13
MTurk	$\boxed{0.44}$	0.13	0.05

Variance in ratings can be explained by topics (Experts) or judges (non-Experts).

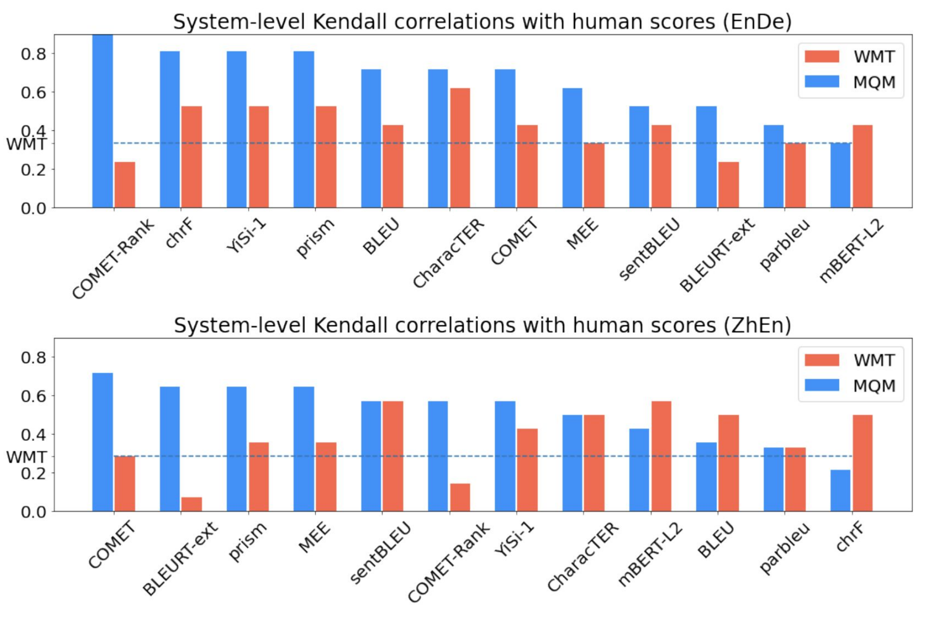


Expert annotations (MQM) have a low correlation with non-expert annotation schemes.

Gillick + Liu, 2010, Freitag et al., 2021

Metrics may be better than non-experts.

Metrics agree more with the high-quality annotations more than with their training data.



Freitag et al., 2021

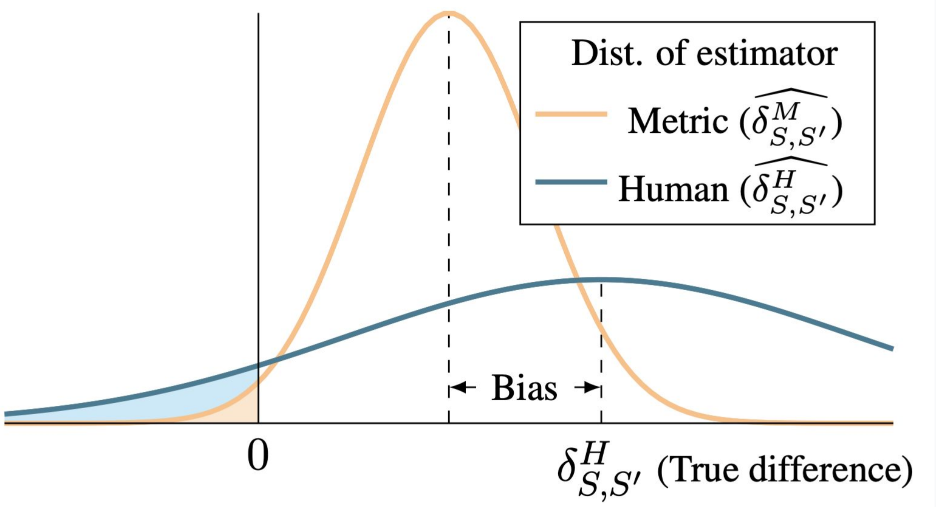
How many annotations do we need?

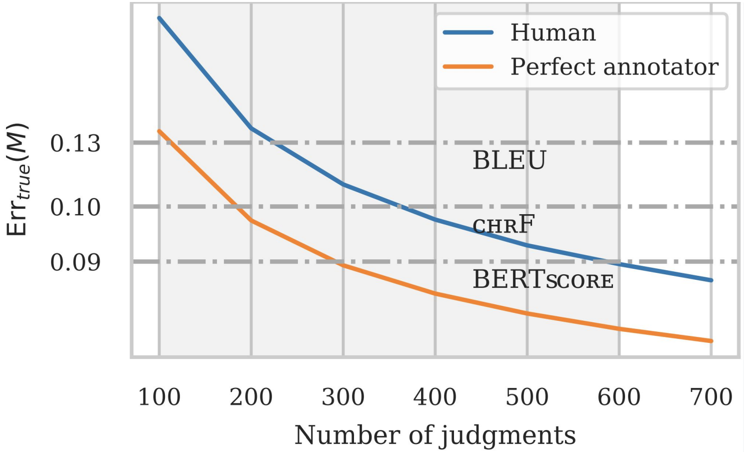
Humans measure the “true” difference between two systems, but have **high variance**. Metrics have lower variance, but are **biased**. Both are sources of errors.

Better models lead to smaller differences and we need more annotator judgments.

To detect a difference of 1 point on a 1-100 scale in WMT, we need 10,000 perfect annotator judgements.

Yet, **the median number of human annotations is 100**.





[Wei + Jia, 2021](), [van der Lee, 2019]()

Lesson 3

Human evaluations are often less reliable than metrics

Lesson 4

Issues with human eval are hidden in the details

We need to explain our data choices.

~29% of NLG papers evaluate on non-English.

"Standard" datasets have significant noise.

Only 38% of papers explicitly state why they chose the datasets they did

We need to move beyond *"previous work used X"* as excuse to continue to work on the same flawed English datasets.

read : falcao still ' has faith ' that he could continue at man utd next season. click here for the latest manchester united news.

Models	Hallucinated I	Hallucinated E	Hallucinated I ∪ E	Faith.	+Fact.
GOLD	**7.4**	73.1	76.9	23.1	—

Maynez, Narayan, et al., 2020

We need to expand our data choices.

NLU researchers have come up with ways to probe for specific model capabilities and failures.

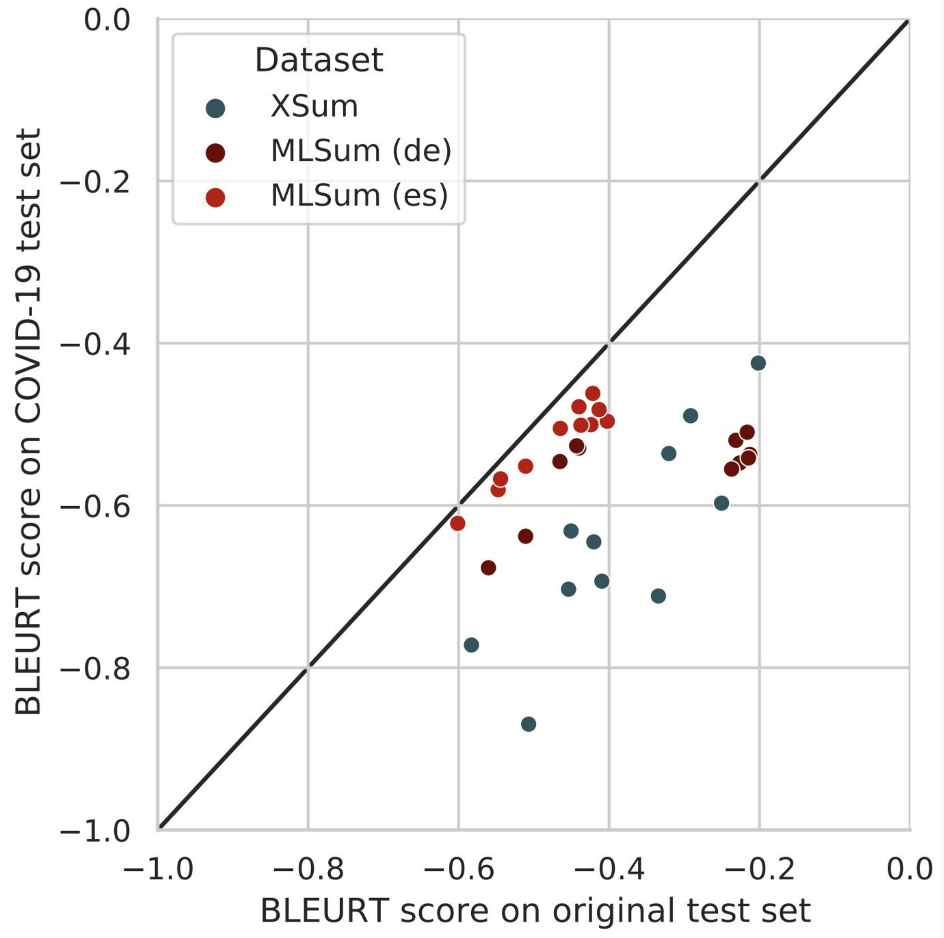
We should do the same for generation.

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

Ribeiro et al., 2020

We need to update our data choices.

Models degrade as time passes but our test sets remain static.



Gehrmann et al., 2021, Mille et al., 2021

We need to contribute to data.

<2% of model developers contribute to data documentation.

20% create evaluation suites, but only 5% release them.

Lesson 5

Testing on new, especially non-English datasets should be normal and as easy as possible

Lesson 6

Datasets and their documentation need version control

Implementing Best Practices with GEMv2

Lessons.

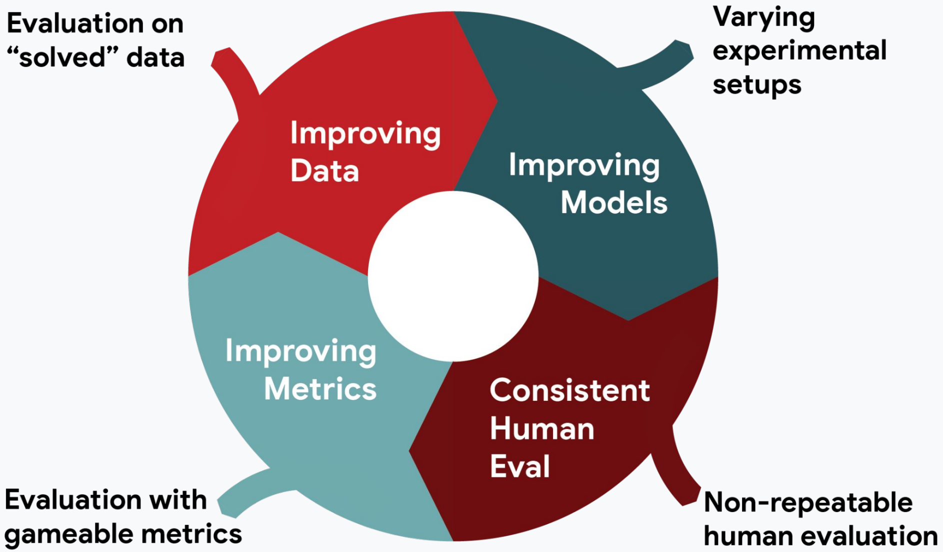
1) Be mindful of what your metrics are (not) measuring
2) Issues in the data will hide issues in the model
3) Human evaluations are often less reliable than metrics
4) Issues with human eval are hidden in the details
5) Testing on new, especially non-English datasets should be normal and as easy as possible
6) Datasets and their documentation need versioning

The ideas are out there but the implementation seems to be the bottleneck

We can only hold model developers accountable for bad evaluation practices if following good practices is possible

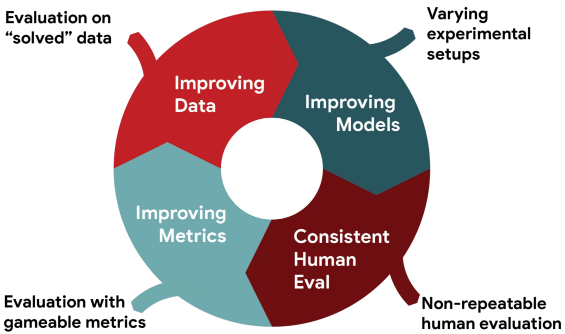
We need to break through this circular dependency.

At the moment, we can't identify whether and how our models **fail**, or whether failure is **attributable** to the data, model, or evaluation.



There is an opportunity to create a positive feedback loop.

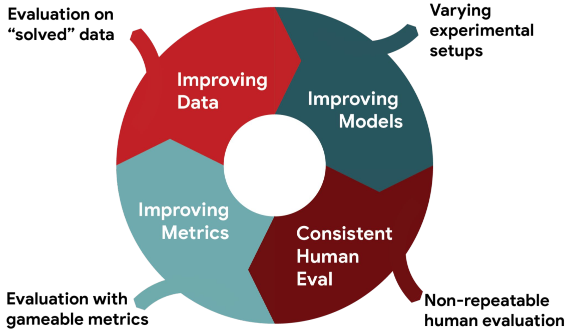
Better evaluation practices ⟵⟶ better evaluation techniques ⟵⟶ better models



- **Call out limitations** of our methods
- Point out and **fix** issues in the data
- **Use combinations** of different metrics/human assessments
- **Release** model and evaluation outputs (esp. non-English)

Gehrmann et al., 2021, Gehrmann, Clark, and Sellam, 2022

Expanding on the lessons, we can derive a list of best practices.



Best Practice & Implementation
Make informed evaluation choices and document them
Evaluate on multiple datasets
Motivate dataset choice(s)
Motivate metric choice(s)
Evaluate on non-English language
Measure specific generation effects
Use a combination of metrics from at least two different categories
Avoid claims about overall "quality"
Discuss limitations of using the proposed method
Analyze and address issues in the used dataset(s)
Discuss or identify issues with the data
Contribute to the data documentation or create it if it does not yet exist
Address these issues and release an updated version
Create targeted evaluation suite(s)
Release evaluation suite or analysis script
Evaluate in a comparable setting
Re-train or -implement most appropriate baselines
Re-compute evaluation metrics in a consistent framework
Run a well-documented human evaluation
Run a human evaluation to measure important quality aspects
Document the study setup (questions, measurement instruments, etc.)
Document who is participating in the study
Produce robust human evaluation results
Estimate the effect size and conduct a power analysis
Run significance test(s) on the results
Conduct an analysis of result validity (agreement, comparison to gold ratings)
Discuss the required rater qualification and background
Document results in model cards
Report disaggregated results for subpopulations
Evaluate on non-i.i.d. test set(s)
Analyze the causal effect of modeling choices on outputs with specific properties
Conduct an error analysis and/or demonstrate failures of a model
Release model outputs and annotations
Release outputs on the validation set
Release outputs on the test set
Release outputs for non-English dataset(s)
Release human evaluation annotations

Gehrmann et al., 2021, Gehrmann, Clark, and Sellam, 2022

GEM Motivation: We can help implement best practices

Without dictating an evaluation approach, how do we make it possible to choose the most appropriate one for any project?

GEM Motivation: We can help implement best practices

Without dictating an evaluation approach, how do we make it possible to choose the most appropriate one for any project?

Goal 1 The core of evaluation is data. We need

- Consistent data loaders and documentation
- Ways to update data and their documentation over time
- Modularity to easily expand to new datasets

Goal 2 Computing automatic metrics should be possible in a consistent and replicable environment across all supported datasets

This is what we are doing with the **G**eneration, **E**valuation, and **M**etrics Benchmark.

Models

💎 Evaluation Infrastructure

Build

Improve

Expand

How can **evaluation practices** benefit from improved **evaluations**?

Data	Automatic Metrics	Human Evaluation
Select and document tasks **Build**	Extensive metrics libraries **Build**	Reproducible framework **Build**
Improve datasets over time **Improve**	Fine-grained breakdowns **Improve**	Consistent definitions and quality control **Improve**
Collect new datasets **Expand**	Development of new metrics **Expand**	Collect annotation corpora **Expand**

Models

Gehrmann et al., 2021

Datasets
We support 35 of them in 45+ languages. Many with new splits, challenge sets, linearization processes...
Summarization, Data-to-Text, Paraphrasing, Simplification, Dialog.

Loaders available at: huggingface.co/GEM

Data Cards
We created an expanded template and interactive form for anyone to use:
huggingface.co/spaces/GEM/DatasetCardForm

We will release an HTML rendering tool for them next month.

Metrics
You don't have to run (most) metrics locally anymore:
huggingface.co/spaces/GEM/submission-form

We now support running metrics in docker to avoid dependency issues:
github.com/GEM-benchmark/GEM-metrics

This is GEMv2.

```python
import datasets

data = datasets.load_dataset("GEM/wiki_lingua", "en")
```

For detailed tutorials, see gem-benchmark.com/tutorials.

```python
submission_dict = {
    "submission_name": "BART-base",
    "param_count": sum(p.numel() for p in model.parameters()),
    "description": "Baseline for the task based on BART-base.",
    "tasks": {
      "common_gen_validation": {"values": valid_formatted, "keys": valid_keys},
      "common_gen_test": {"values": test_formatted, "keys": test_keys},
      "common_gen_challenge_train_sample": {"values": challenge_train_sample_formatted,
                                            "keys": challenge_train_sample_keys}
    }
}
```

```
python run_metrics.py -s outputs.json  -r targets.json -o predictions.json
```

This is GEMv2.

What comes next?

GEM Workshop 2022 at EMNLP. Look out for our call for papers!

GEM Shared Task on multilingual summarization. More to come soon!

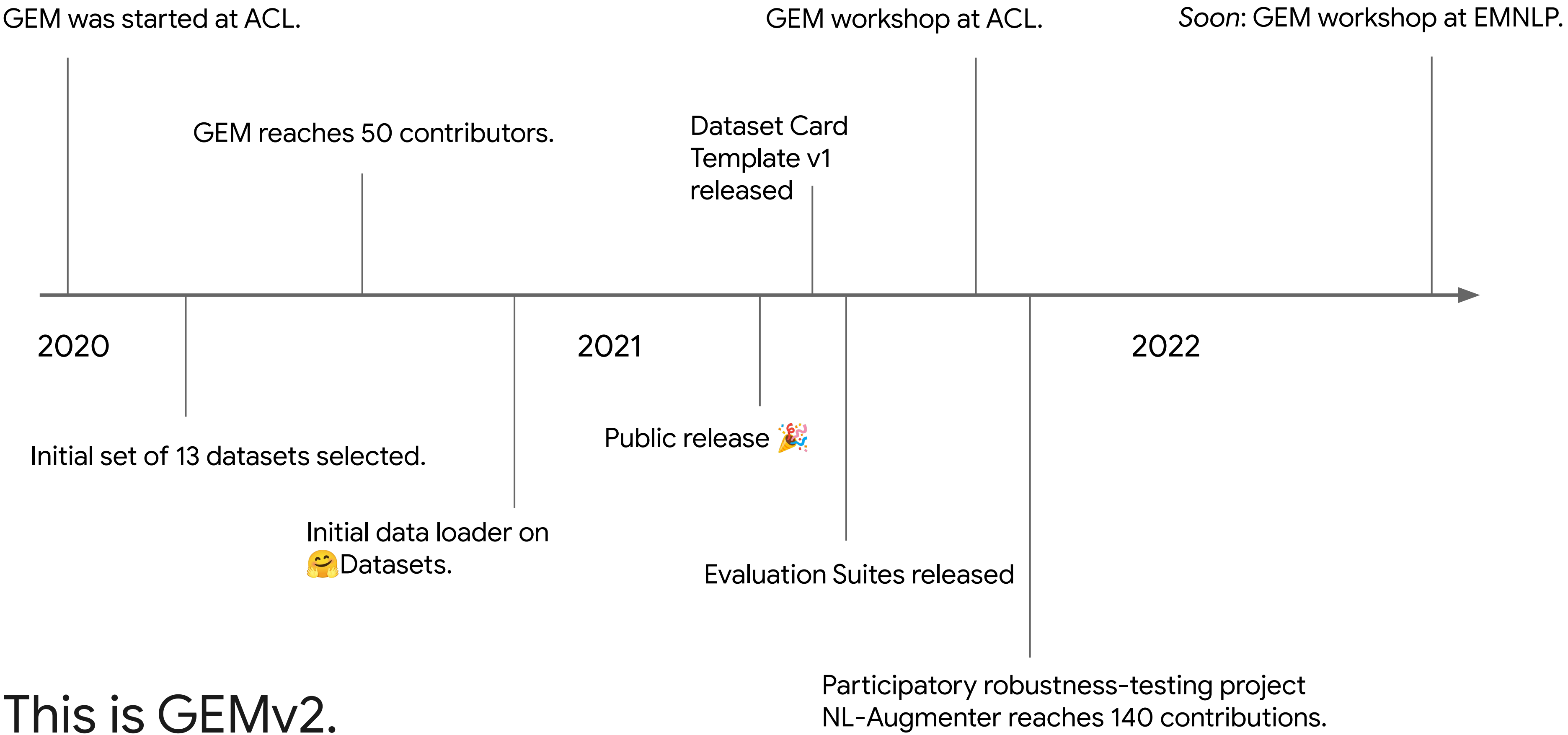
Release of 50k human annotations across all tasks? Sooner than you think 😊

Better human evaluation infrastructure? In the works

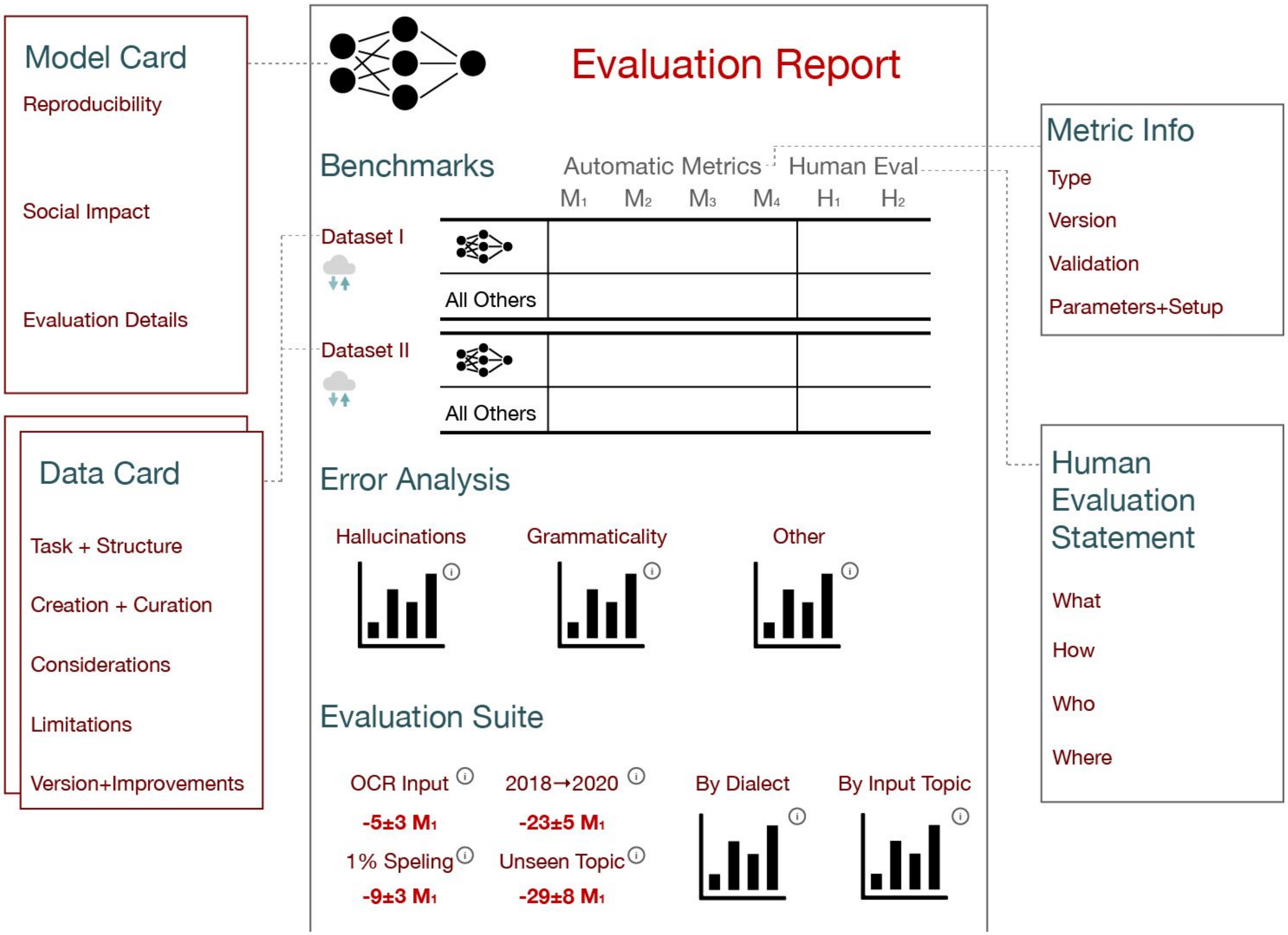
Interactive result investigation? Wouldn't that be nice.

Participatory collection of multilingual and multi-dialectal data? Soon™

This is GEMv2. There is much left to be done. gem-benchmark.com/team/join



This is GEMv2.



Model Card
Reproducibility
Social Impact
Evaluation Details
Data Card
Task + Structure
Creation + Curation
Considerations
Limitations
Version+Improvements
Evaluation Report
Benchmarks
Automatic Metrics
Human Eval
M1 M2 M3 M4 H1 H2
Dataset I
All Others
Dataset II
All Others
Error Analysis
Hallucinations
Grammaticality
Other
Evaluation Suite
OCR Input
-5±3 M1
2018→2020
-23±5 M1
1% Speling
-9±3 M1
Unseen Topic
-29±8 M1
By Dialect
By Input Topic
Metric Info
Type
Version
Validation
Parameters+Setup
Human Evaluation Statement
What
How
Who
Where

Conclusion

We don’t really know how to evaluate models...

But we can do a better job at evaluation

- We can **write better documentation**
- We can **report more metrics**
- We can **frame model results around where they fail**

GEM can help you do this.

Sebastian Gehrmann

gehrmann@google.com

@SebGehr

Google Research

Backup

Abstract:
How good is a system that produces natural language and where does it fail? This question lies at the core of natural language generation research and motivates what systems we develop.

An answer involves deliberations of languages, datasets, metrics, human evaluations, and many more. Using the latest evaluation resources will lead to a more accurate and reproducible answer, but it also relies on keeping up to date with the constantly evolving and fragmented ecosystem of evaluation practices.

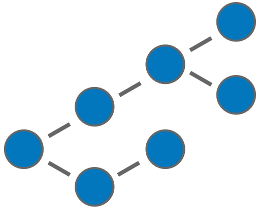
As a result, many system evaluations rely on anglo-centric corpora and well-established, but flawed, metrics. The Generation Evaluation and Metrics benchmark (GEM) is a participatory project aiming to make it easier to use the evaluation resources produced across the NLG community.

In GEMv2, our team of 120 researchers provide access to 35+ corpora in 45+ languages and all the latest metrics in a single line or even without any code. In the talk, I will provide an overview of evaluation challenges and of GEMv2 and discuss how better evaluation practices can lead to better NLG models.

Background: How do we build a simple neural NLG systems?

Step 1 Pick a model parameterized by θ 

Step 2 Train the model on a corpus to find $\underset{\theta}{\operatorname{argmax}}\, p_\theta(y|x)$



Step 3 Perform approximate inference through beam search

What should our results tell us about a model?

✗ System Foo performs the best.

✓ System Foo leads to consistent performance increases
in Bar-type metrics on challenges that measure Baz
while maintaining equal performance on most metrics of type Qux.

What should our results tell us about a model?

✗ System Foo performs the best.

Specific Claims

Multiple Experiments

✓ System Foo leads to consistent performance increases in Bar-type metrics on challenges that measure Baz while maintaining equal performance on most metrics of type Qux.

Acknowledge Limitations

Multiple Metrics