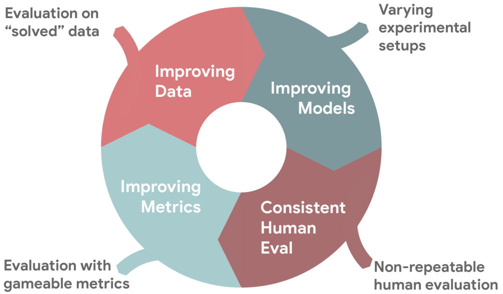
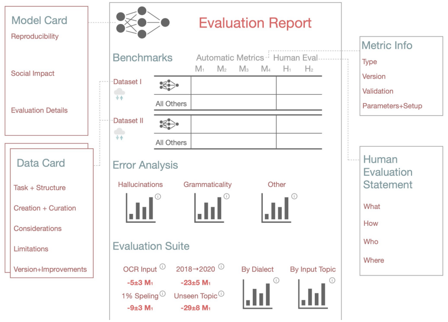
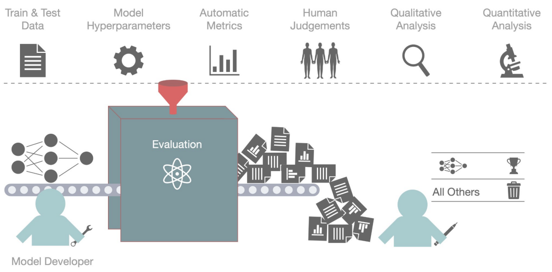






Do we know what we don't know? The state of evaluation in NLP

Stanford NLP Seminar

Sebastian Gehrmann | Google Research | s.gehrmann@outlook.com | @SebGehr

Generated Text has improved a lot

Summary Outputs

2016

norway grants diplomatic protest to russia

j.p. morgan chase may be considered for top job

Issues with fluency
and basic entity disambiguation

2019

Actor Roger Johnson, best known for his roles in Deadlier Than the Male and Some Girls Do, has died at the age of 93.

Fluent, but issues with faithfulness

2022

Rais wa Tanzania John Magufuli ameonekana katika uwanja wa Taifa wa mpira wa miguu jijini Dar es Salaam akiwa amevaa jezi ambayo upande ni sare ya Simba na upande ni sare ya Yanga.

Largely faithful, little-to-no training data,
and support for many languages

How do we evaluate the latest models?

Turing-NLG

7 Qualitative Examples for MT-NLG Generation Capabilities

As an addition to quantitative evaluation and analysis on benchmark datasets, we also qualitatively examined the language generation capabilities on novel scenarios. To our pleasant surprise, MT-NLG is quite capable in solving riddles, answering Jeopardy questions and even generating code off-the-shelf. We present some examples of each category below.

U-PaLM

5.3 Improved Diversity for Open-ended Generation

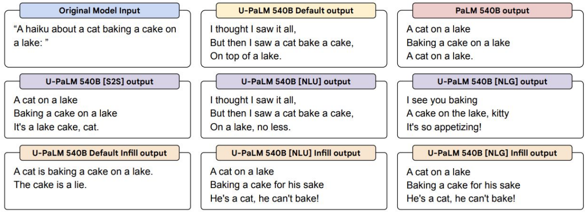


Figure 7: An example of querying U-PaLM for multiple diverse outputs by using different prompt mode token and LM/infill combinations.

Qualitative examples

Turing-NLG

7 Qualitative Examples for MT-NLG Generation Capabilities

As an addition to quantitative evaluation and analysis on benchmark datasets, we also qualitatively examined the language generation capabilities on novel scenarios. To our pleasant surprise, MT-NLG is quite capable in solving riddles, answering Jeopardy questions and even generating code off-the-shelf. We present some examples of each category below.

U-PaLM

5.3 Improved Diversity for Open-ended Generation

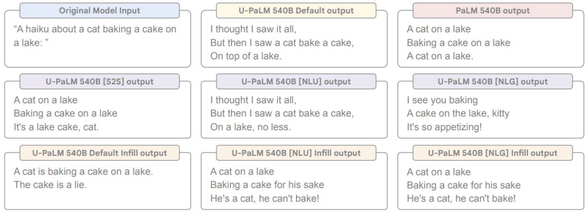


Figure 7: An example of querying U-PaLM for multiple diverse outputs by using different prompt mode token and LM/infill combinations.

GLaM

Natural Language Generative tasks. We compare the language sequences decoded by the models to the ground truth in generative tasks. These tasks are TriviaQA, NQS, WebQS, SQuADv2, LAMBADA, DROP, QuAC and CoQA. The performance is measured by the accuracy of exact match (EM) and F1 score, following the standard for each task in Brown et al. (2020). We use beam search with a width of 4 to generate the sequences.

PaLM

Model	Avg NLG	Avg NLU
GPT-3 175B	52.9	65.4
GLaM 64B/64E	58.4	68.7
PaLM 8B	41.5	59.2
PaLM 62B	57.7	67.3
PaLM 540B	63.9	74.7

Table 5: Average (Avg) Natural Language Generation (NLG) and Natural Language Understanding (NLU) results across 29 benchmarks using 1-shot evaluation. NLG benchmarks include eight tasks – TriviaQA, NQS, WebQS, SQuADv2, LAMBADA, DROP, QuAC and CoQA – while the remaining are NLU benchmarks. Results for GPT-3 and GLaM are from Du et al. (2021).

Language modeling or short-answer QA as proxy for NLG performance

Turing-NLG

7 Qualitative Examples for MT-NLG Generation Capabilities

As an addition to quantitative evaluation and analysis on benchmark datasets, we also qualitatively examined the language generation capabilities on novel scenarios. To our pleasant surprise, MT-NLG is quite capable in solving riddles, answering Jeopardy questions and even generating code off-the-shelf. We present some examples of each category below.

U-PaLM

5.3 Improved Diversity for Open-ended Generation

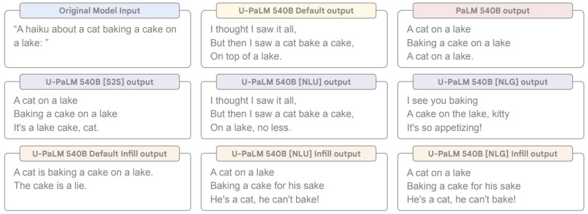


Figure 7: An example of querying U-PaLM for multiple diverse outputs by using different prompt mode token and LM/infill combinations.

GLaM

Natural Language Generative tasks. We compare the language sequences decoded by the models to the ground truth in generative tasks. These tasks are TriviaQA, NQS, WebQS, SQuADv2, LAMBADA, DROP, QuAC and CoQA. The performance is measured by the accuracy of exact match (EM) and F1 score, following the standard for each task in Brown et al. (2020). We use beam search with a width of 4 to generate the sequences.

PaLM

Model	Avg NLG	Avg NLU
GPT-3 175B	52.9	65.4
GLaM 64B/64E	58.4	68.7
PaLM 8B	41.5	59.2
PaLM 62B	57.7	67.3
PaLM 540B	63.9	74.7

Table 5: Average (Avg) Natural Language Generation (NLG) and Natural Language Understanding (NLU) results across 29 benchmarks using 1-shot evaluation. NLG benchmarks include eight tasks – TriviaQA, NQS, WebQS, SQuADv2, LAMBADA, DROP, QuAC and CoQA – while the remaining are NLU benchmarks. Results for GPT-3 and GLaM are from Du et al. (2021).

OPT

		Perplexity (↓)					Unigram F1 (↑)				
Model	**Eval**	**C2**	**WW**	**ED**	**BST**	**WoI**	**C2**	**WW**	**ED**	**BST**	**WoI**
Reddit 2.7B	Unsup.	18.9	21.0	11.6	17.4	18.0	.126	.133	.135	.133	.124
BlenderBot 1	Sup.	**10.2**	12.5	**9.0**	11.9	14.7	.183	.189	.192	.178	.154
R2C2 BlenderBot	Sup.	10.5	**12.4**	9.1	**11.7**	14.6	**.205**	**.198**	**.197**	**.186**	**.160**
OPT-175B	Unsup.	10.8	13.3	10.3	12.1	**12.0**	.185	.152	.149	.162	.147

Table 2: **Dialogue Evaluations.** OPT-175B, in a fully unsupervised setting, performs competitively against fully supervised models.

Perplexity of ground truth outputs

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?

…

Do any of the LLM strategies answer these questions?

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

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

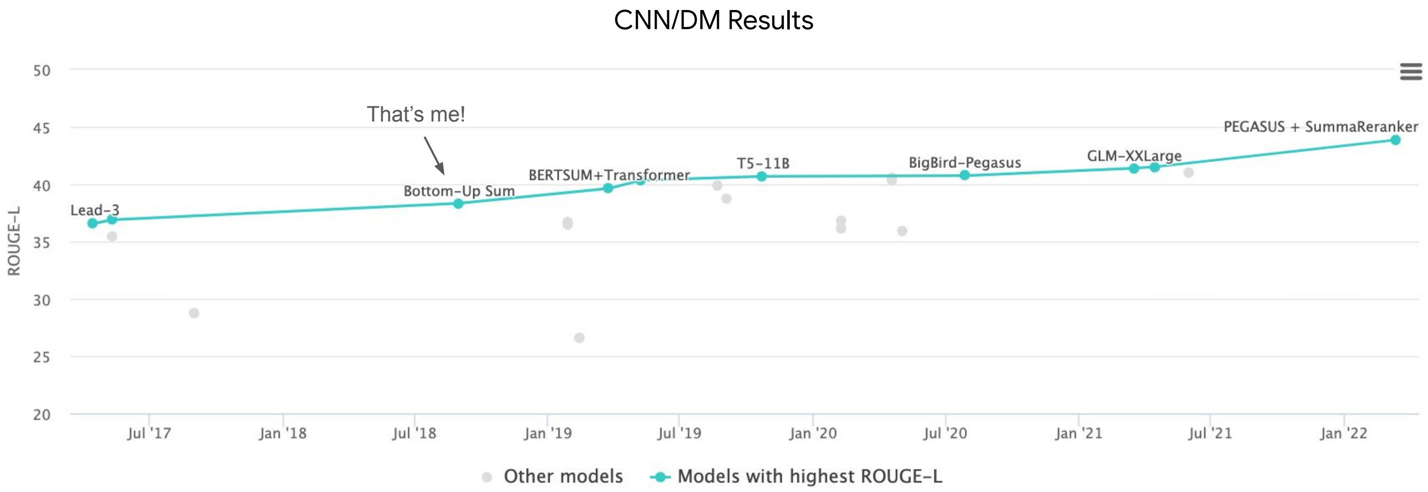
- Does
- What
- How does the model perform on "real-world" data?

…

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

What evidence is presented to make claims about quality?

Gehrmann, Clark, and Sellam, 2022



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

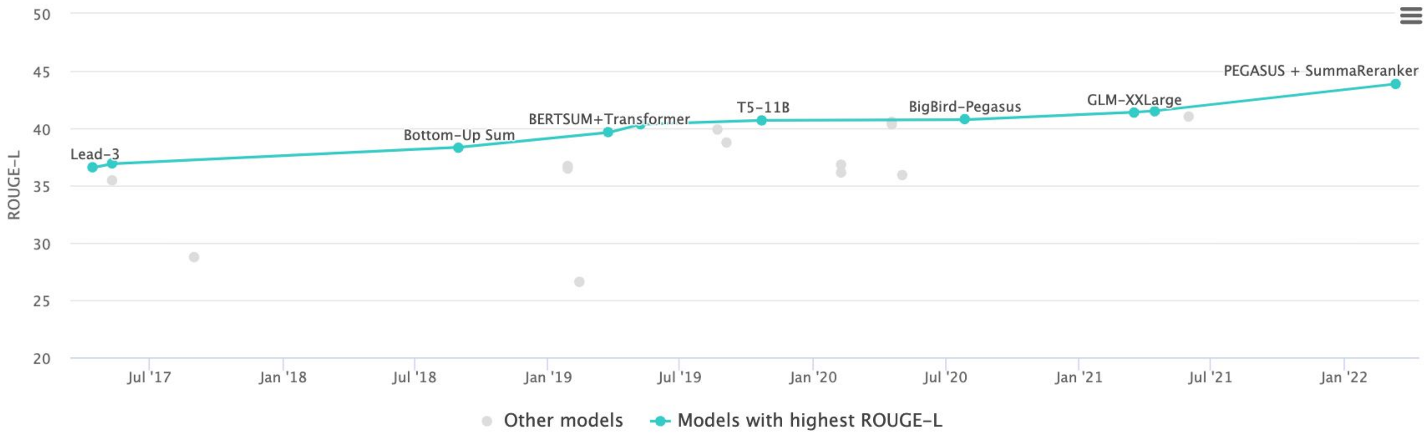
- 100% of summarization papers report ROUGE, 69% report **only** ROUGE
- Together, CNN/DM and XSum are used by 40%+ of papers

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

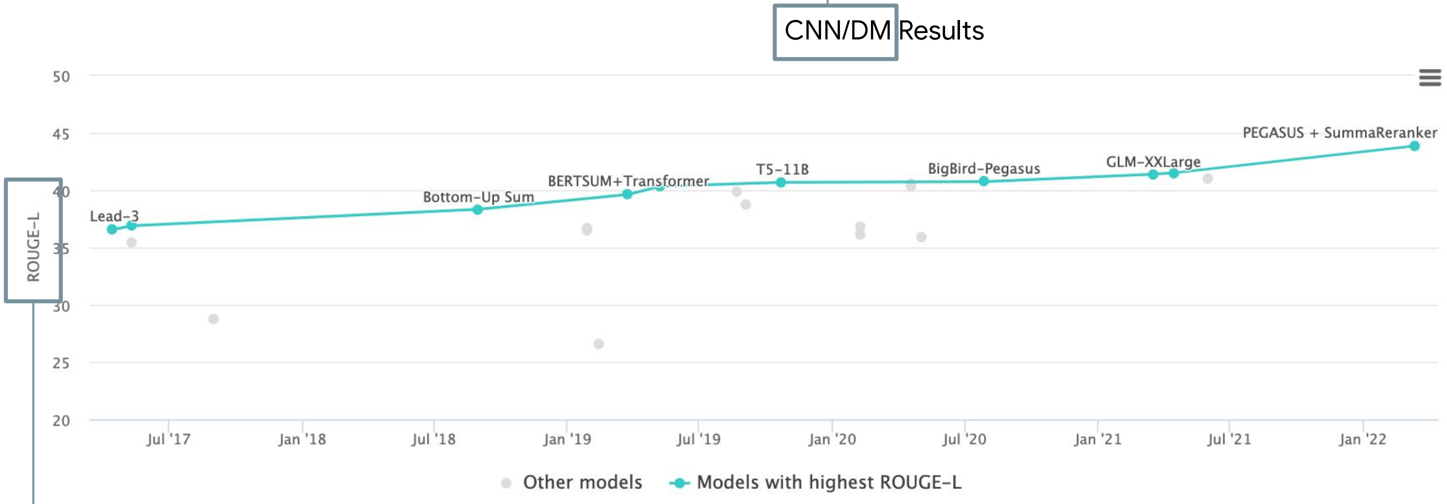
CNN/DM Results



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

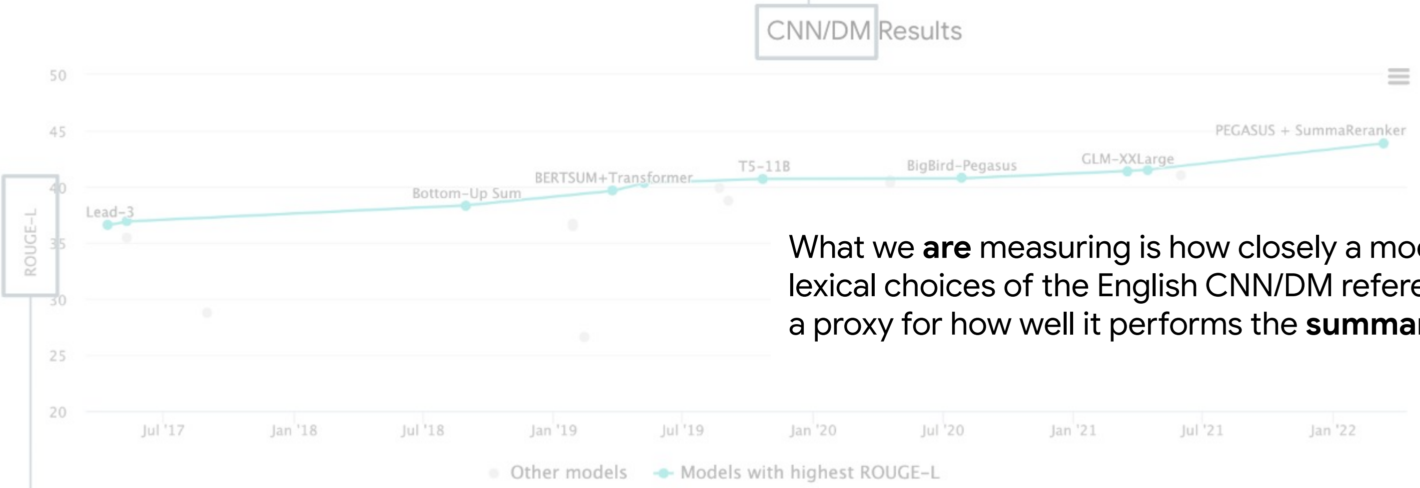


... is not the best possible ROUGE configuration.
... has low correlation with different quality aspects (e.g., faithfulness).
... Increases based on similarity to a reference and is thus confounded by its style and errors.
...

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

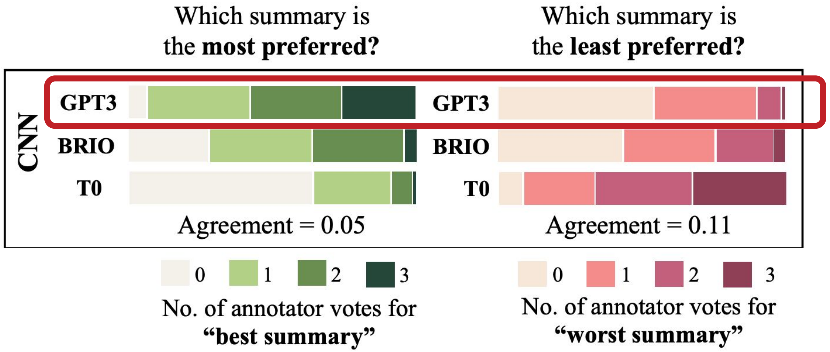


What we **are** measuring is how closely a model can match the lexical choices of the English CNN/DM references, but this is not a proxy for how well it performs the **summarization task**.

... 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.
...

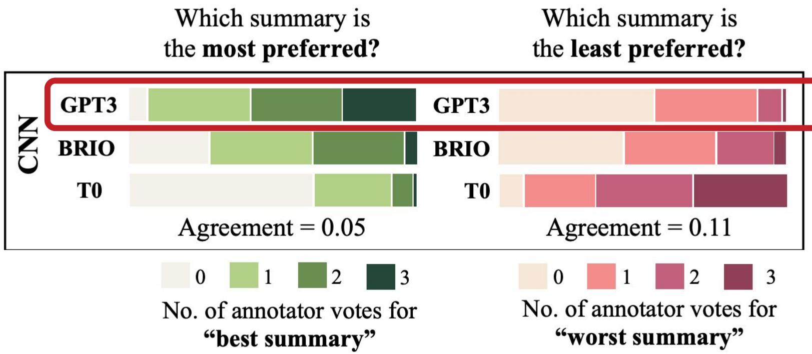
https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Gehrmann, Clark, and Sellam, 2022



Humans rank GPT-3 created summaries as best

Goyal et al., 2022, Rashkin et al., 2021

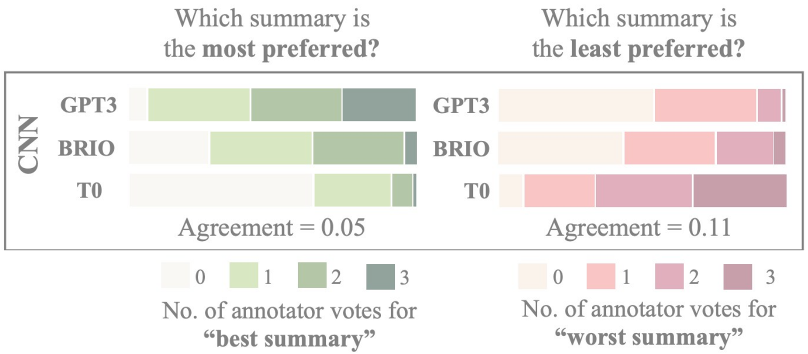


Humans rank GPT-3 created summaries as best

But metrics as worst….

Dataset	Model	Overlap-Based			Similarity-Based		QAEval	
		ROUGE(1/2/L)	METEOR	BLEU	BERTScore	MoverScore	EM	F1
CNN	PEGASUS	34.85/14.62/28.23	.24	7.1	.858	.229	.105	.160
	BRIO	38.49/17.08/31.44	.31	6.6	.864	.261	.137	.211
	T0	35.06/13.84/28.46	.25	5.9	.859	.238	.099	.163
	GPT3-D2	31.86/11.31/24.71	.25	3.8	.858	.216	.098	.159

Goyal et al., 2022, Rashkin et al., 2021



Model	Approach	Int	AIS
MatchSum (Zhong et al. 2020)	Extractive	90.0	99.4
Pointer-Gen (See, Liu, and Manning 2017)	Hybrid	90.0	97.8
BigBird (Zaheer et al. 2020)	Abstractive	90.0	87.2*
Reference	-	86.0	54.1*

Only 54.1% of references in the dataset are faithful to the underlying article.

Dataset	Model	Overlap-Based			Similarity-Based		QAEval	
		ROUGE(1/2/L)	METEOR	BLEU	BERTScore	MoverScore	EM	F1
CNN	PEGASUS	34.85/14.62/28.23	.24	7.1	.858	.229	.105	.160
	BRIO	38.49/17.08/31.44	.31	6.6	.864	.261	.137	.211
	T0	35.06/13.84/28.46	.25	5.9	.859	.238	.099	.163
	GPT3-D2	31.86/11.31/24.71	.25	3.8	.858	.216	.098	.159

Goyal et al., 2022, Rashkin et al., 2021

Lesson 1

Be mindful of what your metrics are (not) measuring

Lesson 2

Issues in the data will hide issues in models

Lesson 1

Be mindful of what your metrics are (not) measuring

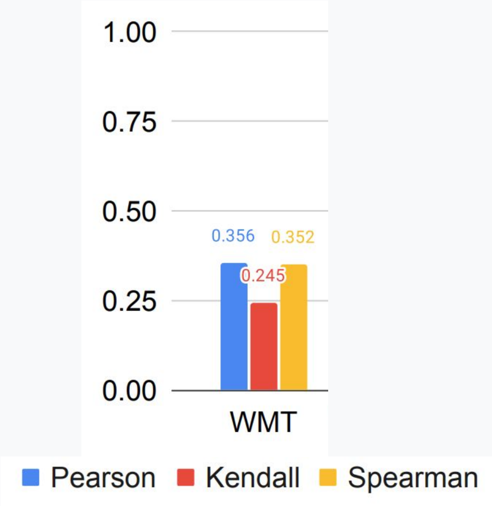
Can human evaluations solve this issue?

Lesson 2

Issues in the data will hide issues in models

It depends.

Agreement between individual ratings by linguists and those from non-expert crowdworkers can be extremely low.



Freitag et al., 2021

It depends.

Automatic metrics don't have a good correlation with human judgments, even on the system level.

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
ROUGE-su*	0.1912	0.2941	0.4354	0.3235
ROUGE-w	0.0000	0.3971	0.3764	0.1618
ROUGE-we-1	**0.2647**	0.4559	**0.5092**	**0.4265**
ROUGE-we-2	-0.0147	0.5000	0.3026	0.1176
ROUGE-we-3	0.0294	0.3676	0.3026	0.1912
S^3-pyr	-0.0294	0.5147	0.3173	0.1324
S^3-resp	-0.0147	0.5000	0.3321	0.1471
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**
MoverScore	0.1912	-0.0294	0.2583	0.2941
SMS	0.1618	0.5588	0.3616	0.2353
SummaQA^	0.1176	**0.6029**	0.4059	0.2206
BLANC^	0.0735	0.5588	0.3616	0.2647
SUPERT^	0.1029	0.5882	0.4207	0.2353
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**
Length^	-0.0294	0.4265	0.2583	0.1618
Novel unigram^	0.1471	-0.2206	-0.1402	0.1029
Novel bi-gram^	0.0294	-0.5441	-0.3469	-0.1029
Novel tri-gram^	0.0294	-0.5735	-0.3469	-0.1324
Repeated unigram^	**-0.3824**	0.1029	-0.0664	-0.3676
Repeated bi-gram^	**-0.3824**	-0.0147	-0.2435	**-0.4559**
Repeated tri-gram^	-0.2206	0.1471	-0.0221	-0.2647
Stats-coverage^	-0.1324	0.3529	0.1550	-0.0294
Stats-compression^	0.1176	-0.4265	-0.2288	-0.0147
Stats-density^	0.1618	**0.6471**	0.3911	0.2941

Fabbri et al., 2021

What is even being measured?

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

Often, the details are not provided:

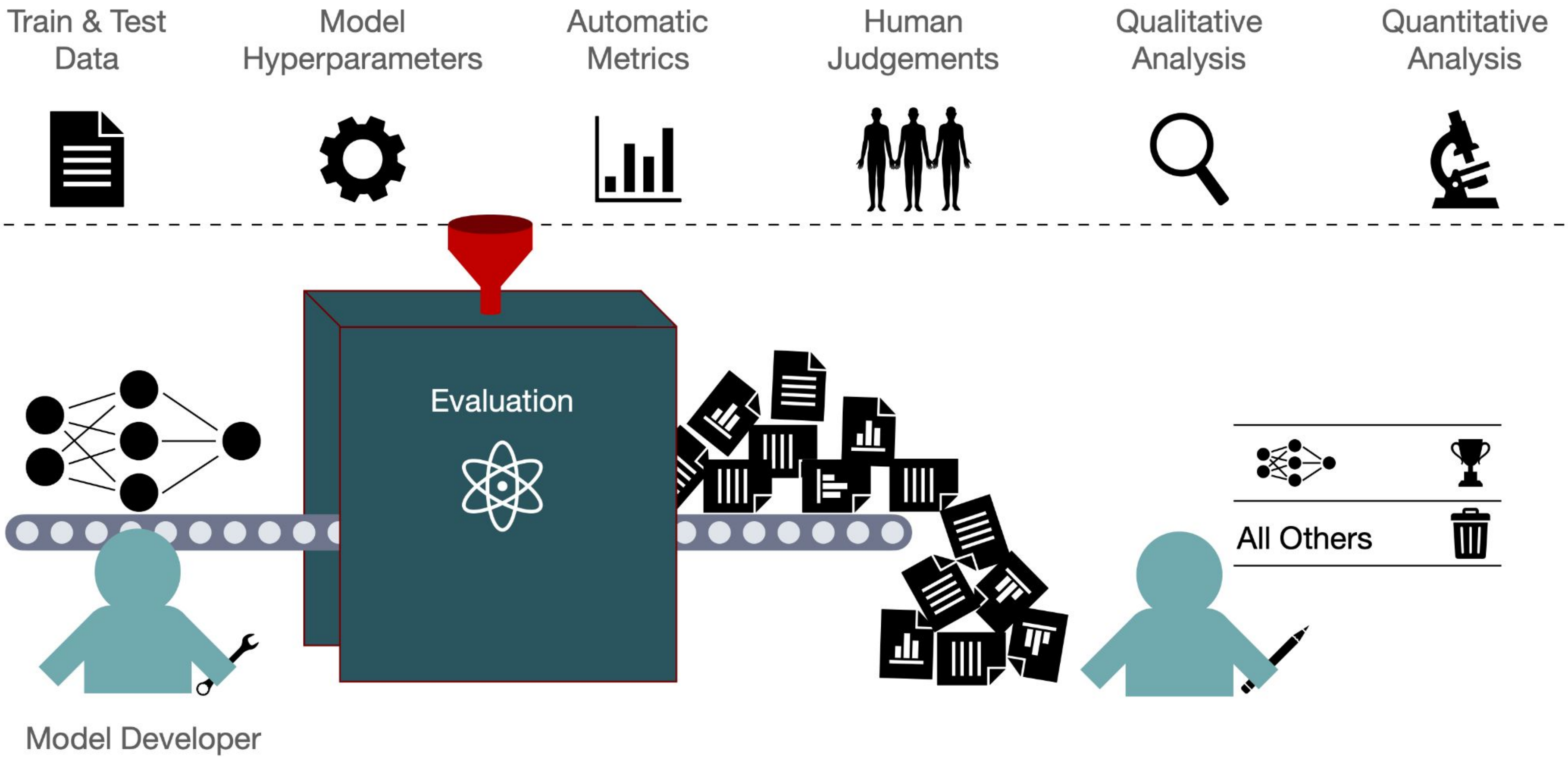
- >50% missing definitions
- ~66% missing prompts/questions
- 20% missing criteria names

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

Howcroft et al., 2020

Lesson 3

Human evaluations may not always be good and issues be hidden in the details


Train & Test Data
Model Hyperparameters
Automatic Metrics
Human Judgements
Qualitative Analysis
Quantitative Analysis
Evaluation
All Others
Model Developer

Agenda

01 Where do we want to be?

02 How do we get there?

03 New strategies for task-development in NLP

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

There is no equivalent of accuracy or F1 for NLG

Does y fulfill the **communicative goal?**

Is all information in y **attributable** to x?

Is y **fluent, natural**, and **free from grammatical erros**?

$$f : x \rightarrow y$$

Is y **not redundant** and contains the **most relevant content**?

Is f **robust to shifts** in the data distribution?

Does f work equally well for **all kinds of expected inputs**?

Does f produce **diverse outputs** in repeated interactions?

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

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 scenarios

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

Specific Metric(s)

1

Datasets

2

Human Evaluation and Automatic Metrics

3

Evaluation Suites

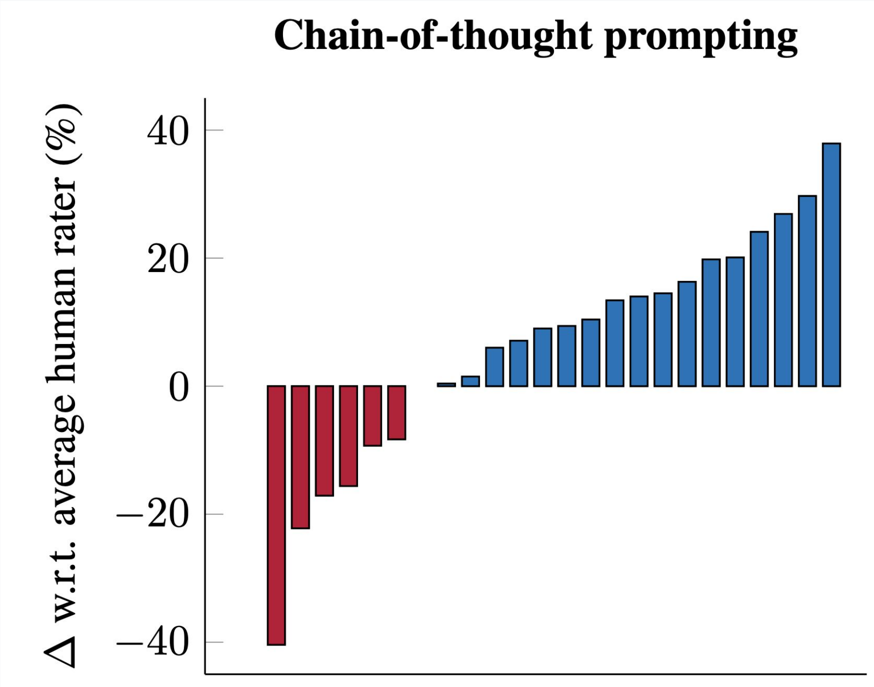
How do we get there?

Evaluation suite development in the age of LLMs

✓ No large training set needed

✗ Test set overlap
✗ Benchmarks are easily broken
✗ Metrics are still unclear

How to take advantage of LLMs?



The best current models already outperform humans on the most challenging out of 200+ tasks in BIG-bench.

Srivastava et al., 2022, Suzgun et al., 2022

Three opportunities for evaluation suite development

01 Curating existing resources (Gehrmann et al., 2021, 2022, Mille et al. 2021)

02 Human-AI collaboration (Yuan et al., 2021) ← This talk

03 New collection methodologies (Parikh et al., 2020, Gehrmann et al., 2022)

The WikiBio Task [Lebret et al., 2016]

Communicative Goal

Generate a brief description of a person grounded in descriptive attributes

Input / Target

Key-Value attribute pairs ⟶ ~1 paragraph biography

Challenges

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

Judy Garland

Garland c. 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

Judy Garland (born **Frances Ethel Gumm**; June 10, 1922 – June 22, 1969) was an American actress and singer. While critically acclaimed for many different roles throughout her career, she is widely known for playing the part of Dorothy Gale in *The Wizard of Oz* (1939).[2][3] 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]

The WikiBio Task [Lebret et al., 2016]

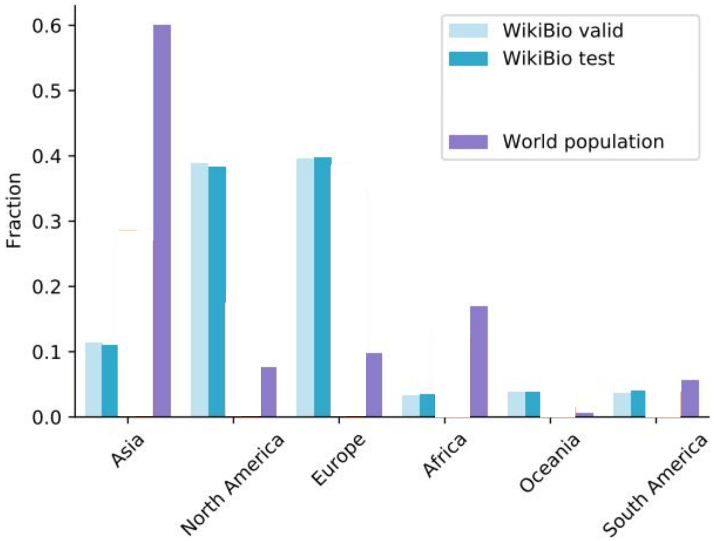
The task is very **noisy**

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

It does **not represent everyone**

	He / She / They / ?
WikioBio Valid	60 / 12 / 2 / 27
WikioBio Test	59 / 12 / 2 / 27

Type	WikiBio %
musical artist	11.7%
sportsperson	9%
scientist	4.4%
writer	3.6%
artist	2.5%
spy	0.03%
theologian	0.03%
mountaineer	0.009%



The WikiBio Task [Lebret et al., 2016]

The task is very **noisy**

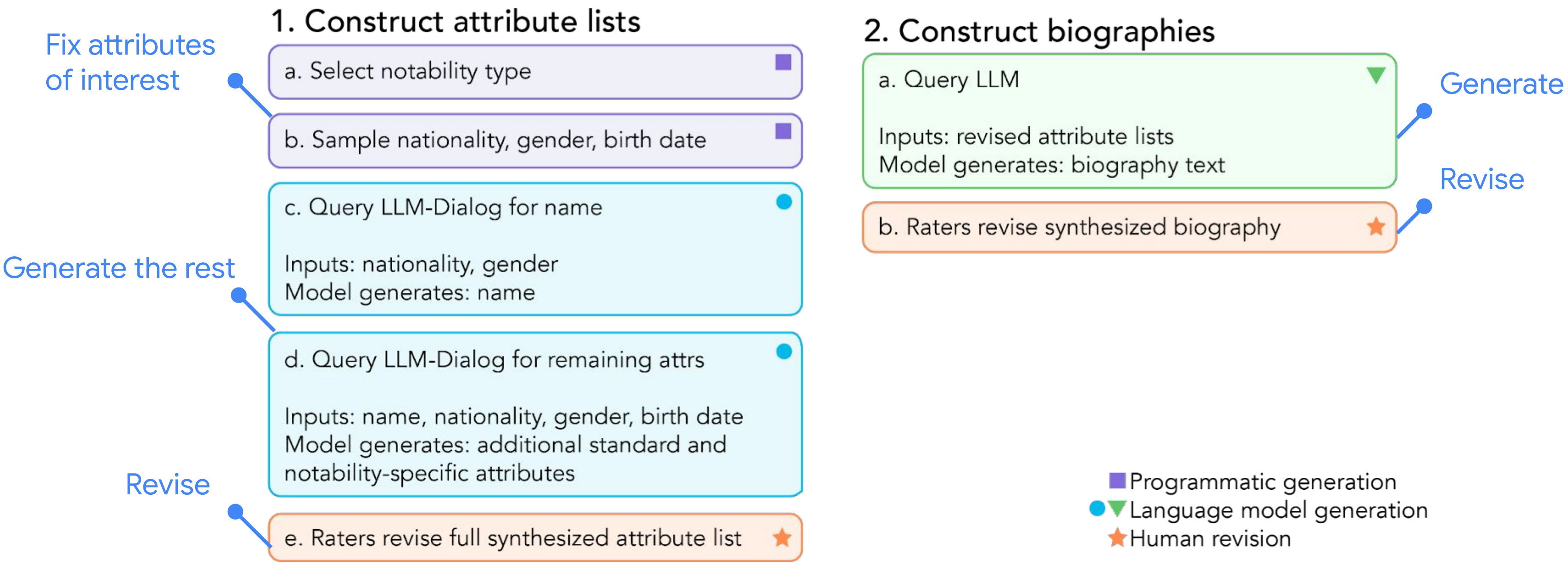
It does **not represent everyone**

Models are **already trained on Wikipedia**

On the right,
green text is found in the attribute box,
blue text is correct but not found.

Notability	Generated Bio
Author	**Ken MacLeod** is a Scottish writer, working in multiple genres. MacLeod began writing science fiction in 1997, with "The Stone Canal" – a nominee for a Nebula Award, Campbell Award, and BSFA Award. In 1999, MacLeod was awarded the Campbell Award for his novel "Newton's Wake" as well as the 2001 BSFA Award for 'The Cassini Division'.
Sportsperson	**Bill Kazmaier** is an American professional Powerlifter, strongman, and Weightlifter. During Kazmaiers career they held multiple world records across various strongman and powerlifting federations. Kazmaier began training in 1976, and competed until 1990 with multiple injuries.
Musician	**Al Alberts** was born on July 22, 1922, in Philadelphia, Pennsylvania. He learned piano as a child and became a vocalist in the late 1940s. He became most famous as the founder of The Four Aces.

Can we leverage a language model to create a test set without these issues?



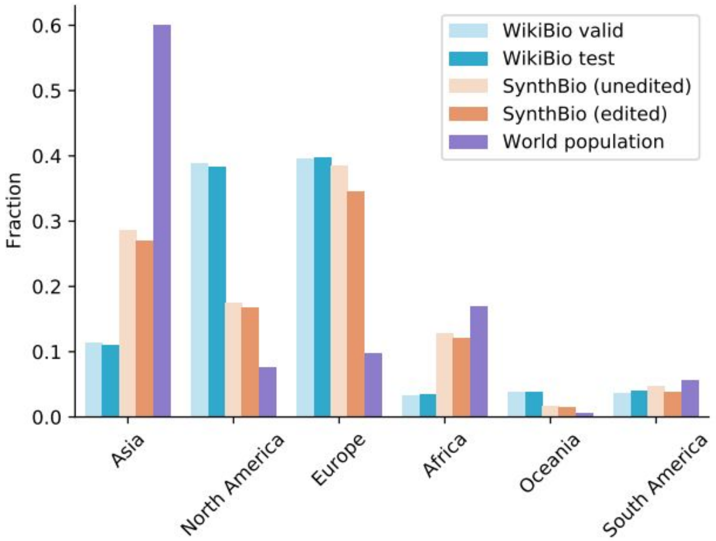
Result

Much better coverage and faithfulness

Dataset	Coverage	Faithfulness	Fluency
WikiBio	0.44±0.007	μ = 2.5	0.97
SynthBio	0.86±0.006	μ = 3.75	0.97

Much better representation

Type	WikiBio %	SynthBio %
musical artist	11.7%	12.5%
sportsperson	9%	12.5%
scientist	4.4%	12.5%
writer	3.6%	12.5%
artist	2.5%	12.5%
spy	0.03%	12.5%
theologian	0.03%	12.5%
mountaineer	0.009%	12.5%



Result

Much better coverage
and faithfulness

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

Much better representation

	He / She / They / ?
WikioBio Valid	60 / 12 / 2 / 27
WikioBio Test	59 / 12 / 2 / 27
SynthBio (unedited)	45 / 40 / 9 / 6
SynthBio (final)	38 / 37 / 23 / 2

Posthoc editing is necessary

What can we do with SynthBio?

Can we evaluate **language quality**?	No. We would overfit to the example-producing model.
Can we evaluate **coverage** and **faithfulness**?	Yes!

→ **5/6 metrics** produced a different rankings when the same models were evaluated on the old and new test sets.

https://storage.googleapis.com/gem-benchmark/SynthBio.json

New Dataset Collection Methodologies

Desiderata for a new data-to-text task.

✓ Focus on reasoning over multiple cells

✓ Multilingual and parallel to enable translation research

✓ Avoid Western-centric entities

✓ Avoid memorization

✓ High-quality references

✓ Clear evaluation approach

The DHS Program
Demographic and Health Surveys

Household Composition

The average household size in Kenya is 3.7 members. Nearly 1 in 3 households are headed by women (31%). Thirty-nine percent of the Kenyan population is under age 15.

Water, Sanitation, and Electricity

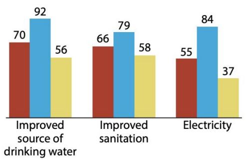
Seven in ten Kenyan households have access to an improved source of drinking water. Ninety-two percent of urban households and 56% of rural households have access to an improved source of drinking water.

Two-thirds of households in Kenya use an improved sanitation facility, including facilities shared with other households. Urban households are more likely than rural households to use improved sanitation facilities (79% versus 58%). Twenty-eight percent of households use unimproved sanitation, while 6% of households have no sanitation facility or openly defecate.

More than half of Kenyan households have electricity (55%). The majority of urban households have electricity (84%), compared to 37% of rural households.

Water, Sanitation, and Electricity by Residence

Percent of households with:



© 2014 Jonathan Torgovnik, Getty Images, Images of Empowerment

Ownership of Goods

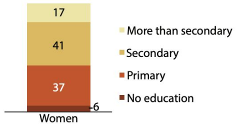
Most Kenyan households have a mobile phone (90%), 72% have a radio, and 49% have a television. More than half of Kenyan households own agricultural land (52%) or farm animals (56%). Urban households are more likely than rural households to own a mobile telephone, radio, or television. In contrast, rural households are more likely to own agricultural land or farm animals than urban households.

Education

Six percent of women age 15-49 in Kenya have no education. More than one-third of women (37%) have attended primary school, while 41% have attended secondary school. Only 17% of women have more than secondary education. Nearly 9 in 10 women (89%) are literate. More women in urban areas are literate, compared to rural women (95% versus 85%, respectively).

Education among Women

Percent distribution of women age 15-49 by highest level of education attended



Infographic-to-text

Communicative Goal

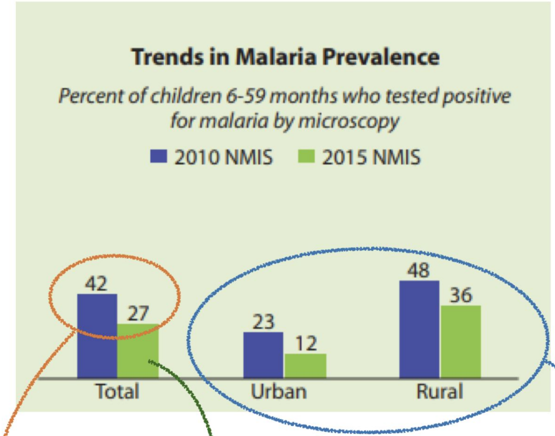
Given a tabular representation of an infographic, generate a short description.

Input / Target

A table with column and row labels and values
⟶ a single sentence in a specified language

Challenges

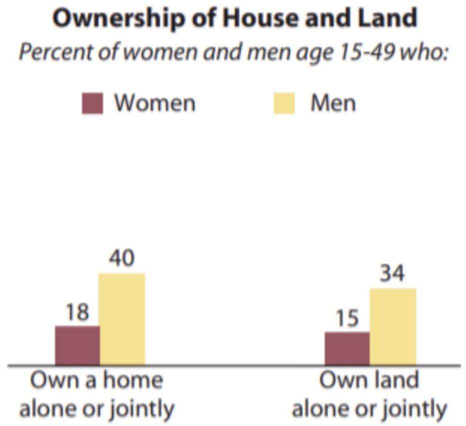
- Select relevant cells
- Compare and contrast cells in natural language
- Do not hallucinate



More than one quarter of children under 5 tested positive for malaria by microscopy.

Prevalence is higher among rural households than urban households.

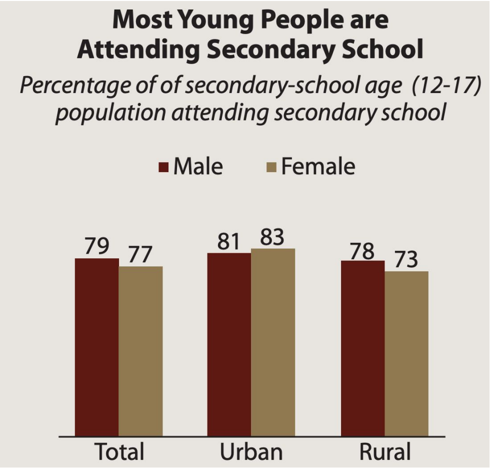
The malaria prevalence among children has decreased by 15% between 2010 and 2015.

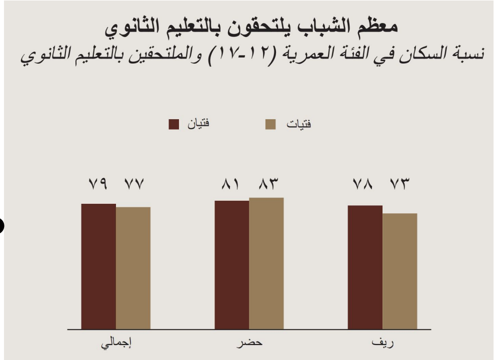


(1) We transcribe everything into tables and extract descriptive sentences

Title Ownership of House and Land
Unit of Measure Percent of women and men age 15-49 who:

	Women	Men
Own a home alone or jointly	18	40
Own land alone or jointly	15	34





(2) We get parallelism between two languages by design, and use professional translators for all others

TaTA: Table-to-Text in African languages

TaTA supports 8 languages.
Every example is available in all of them.

The references are largely faithful (but not perfect).
75% of outputs require reasoning over μ=8 cells.

Only 1.5% of 15-grams in references exist in mC4.
For the same languages in universal dependencies, the average is 45%.

Language	# Transcribed / # Translated
Arabic	157 / 711
English	903 / 0
French	88 / 778
Hausa	62 / 804
Igbo	32 / 834
Portuguese	23 / 833
Swahili	68 / 800
Yorùbá	25 / 841

	Faithfulness	Reasoning	# Cells
Reference		0.75	$8.0_{6.7}$

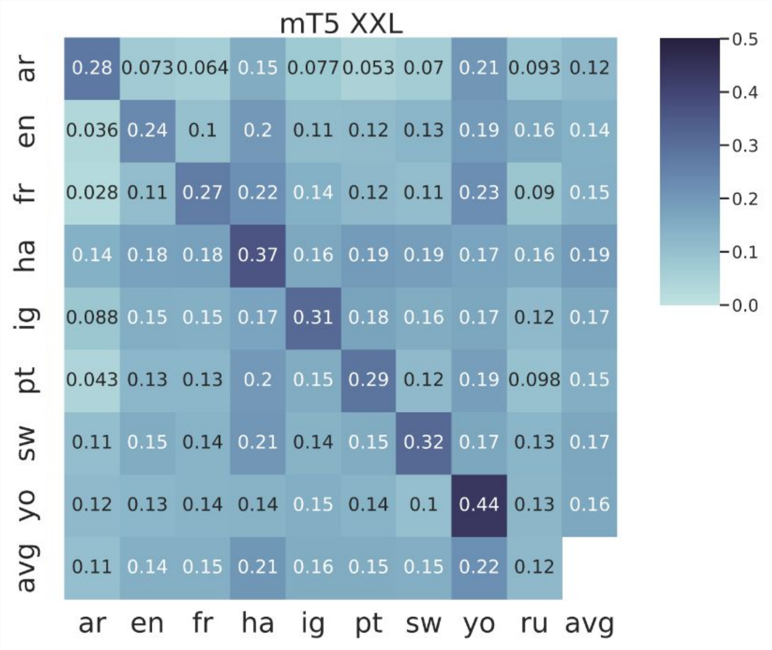
The old problem with the metrics

All standard metrics disagree with each other.

Cross-lingual experiments led to confusing findings:

- Hausa is the best language to train on
- Models trained on any language performs well on Yorùbá

???



Standard metric performance when a model trained on language A (left) is evaluated on language B (right)

A New Paradigm for Metrics

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

$$f : x \rightarrow y$$

A metric that measures
a **particular quality aspect**

The **metric score**

$$g : x, \hat{y}, y \rightarrow \mathbb{R}$$

Source, Reference, and System Output

A new paradigm for metrics

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
ROUGE-su*	0.1912	0.2941	0.4354	0.3235
ROUGE-w	0.0000	0.3971	0.3764	0.1618
ROUGE-we-1	**0.2647**	0.4559	**0.5092**	**0.4265**
ROUGE-we-2	-0.0147	0.5000	0.3026	0.1176
ROUGE-we-3	0.0294	0.3676	0.3026	0.1912
S^3-pyr	-0.0294	0.5147	0.3173	0.1324
S^3-resp	-0.0147	0.5000	0.3321	0.1471
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**
MoverScore	0.1912	-0.0294	0.2583	0.2941
SMS	0.1618	0.5588	0.3616	0.2353
SummaQA^	0.1176	**0.6029**	0.4059	0.2206
BLANC^	0.0735	0.5588	0.3616	0.2647
SUPERT^	0.1029	0.5882	0.4207	0.2353
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**
Length^	-0.0294	0.4265	0.2583	0.1618
Novel unigram^	0.1471	-0.2206	-0.1402	0.1029
Novel bi-gram^	0.0294	-0.5441	-0.3469	-0.1029
Novel tri-gram^	0.0294	-0.5735	-0.3469	-0.1324
Repeated unigram^	**-0.3824**	0.1029	-0.0664	-0.3676
Repeated bi-gram^	**-0.3824**	-0.0147	-0.2435	**-0.4559**
Repeated tri-gram^	-0.2206	0.1471	-0.0221	-0.2647
Stats-coverage^	-0.1324	0.3529	0.1550	-0.0294
Stats-compression^	0.1176	-0.4265	-0.2288	-0.0147
Stats-density^	0.1618	**0.6471**	0.3911	0.2941

Existing metrics try to do everything, but do nothing well.

General-purpose metrics cannot give us the performance breakdown we desire.

	Ensemble	Q^2_{metric}	ANLI	SC_{ZS}	F1	BLEURT	QuestEval	FactCC	$BART_{score}$	$BERT_{score}$
FRANK	91.2	87.8	**89.4**	89.1	76.1	82.8	84.0	76.4	86.1	84.3
SummEval	82.9	78.8	80.5	**81.7**	61.4	66.7	70.1	75.9	73.5	77.2
MNBM	76.6	68.7	**77.9****	71.3	46.2	64.5	65.3	59.4	60.9	62.8
QAGS-C	87.7	**83.5**	82.1	80.9	63.8	71.6	64.2	76.4	80.9	69.1
QAGS-X	84.8	70.9	**83.8**	78.1	51.1	57.2	56.3	64.9	53.8	49.5
BEGIN	86.2	79.7	82.6	82.0	86.4	86.4	84.1	64.4	86.3	**87.9**
$\mathbf{Q^2}_{dataset}$	82.8	**80.9***	72.7	77.4	65.9	72.4	72.2	63.7	64.9	70.0
DialFact	90.4	**86.1****	77.7	84.1	72.3	73.1	77.3	55.3	65.6	64.2
PAWS	91.2	**89.7****	86.4	88.2	51.1	68.3	69.2	64.0	77.5	77.5
FEVER	94.7	88.4	**93.2****	~~93.2~~	51.8	59.5	72.6	61.9	64.1	63.3
VitaminC	96.1	81.4	**88.3****	~~97.9~~	61.4	61.8	66.5	56.3	63.2	62.5
Avg. w/o VitC, FEVER	86.0	80.7	**81.5**	81.4	63.8	71.4	71.4	66.7	72.2	71.4

Fabbri et al., 2020, Honovich et al., 2022

A new paradigm for metrics

What if, instead of relying on existing metrics,
a benchmark can be released with its own metrics?

We are saving a ton by not needing large training corpora.
So **let's collect human annotations as metric training data**.

Annotate validation outputs to train metrics,
and test outputs to evaluate systems AND the new metrics

Gehrmann et al., 2022

Applying this to TaTA

Conventional metrics fail to capture attribution and/or understandability.

The dataset-specific metrics have **high correlations**

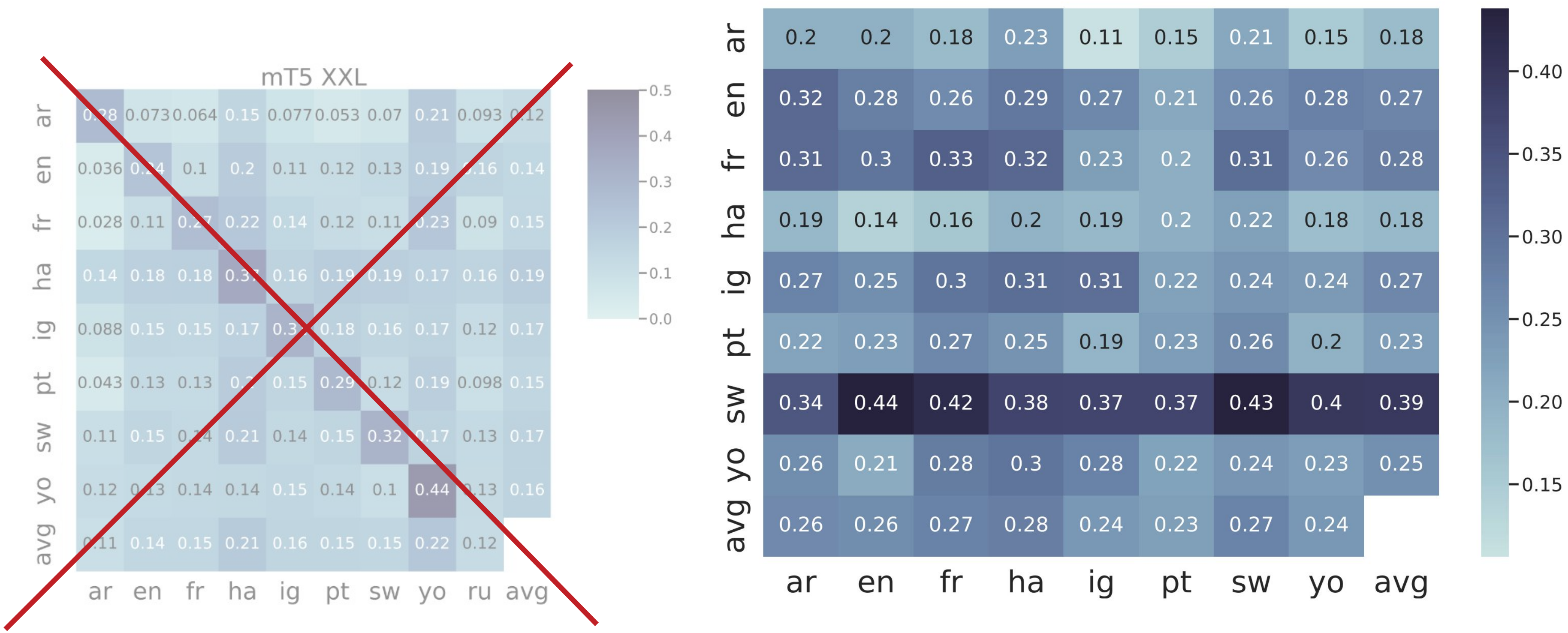
The best metric needs **no references**!

Generic metrics

	Correlation with U+A
BLEURT-20	0.12
ROUGE-1 P/R/F	0.07 / 0.09 / 0.11
ROUGE-2 P/R/F	0.12 / 0.11 / 0.13
ROUGE-L P/R/F	0.08 / 0.11 / 0.13
TABLE P/R/F	0.02 / 0.06 / 0.05
CHRF	0.16
STATA QE	**0.66**
STATA QE+REF	0.61
STATA REF	0.53

Dataset-specific

Better metrics lead to better science



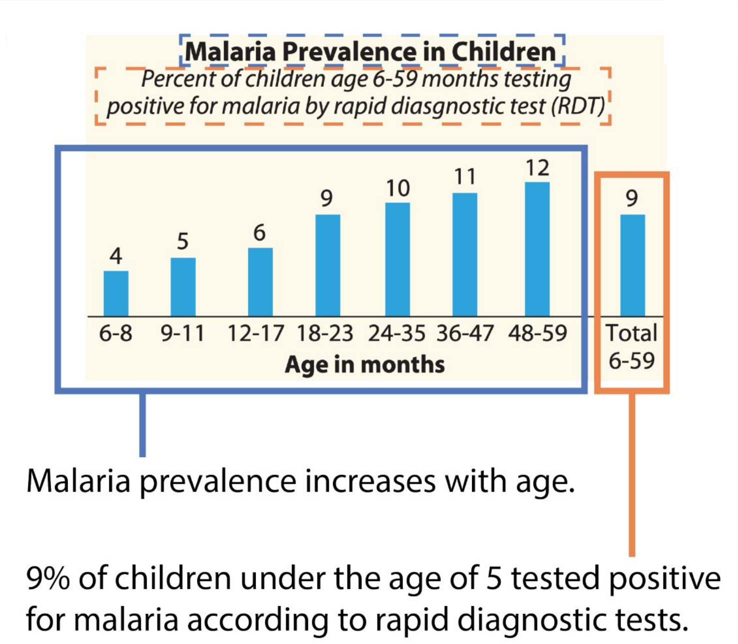
What does this mean?

Validate that metrics are measuring what you want them to measure.

Invest into good human evaluations by focusing on test set collection instead of training set collection.

Release metrics alongside datasets.

Datasets in 1-2 years may just be a collection of
Dev and test inputs and **human annotations**



https://huggingface.co/datasets/GEM/TaTA

Conclusion

What can **you** do to improve evaluations?

Treat evaluation as an equal partner to model development, not an afterthought.

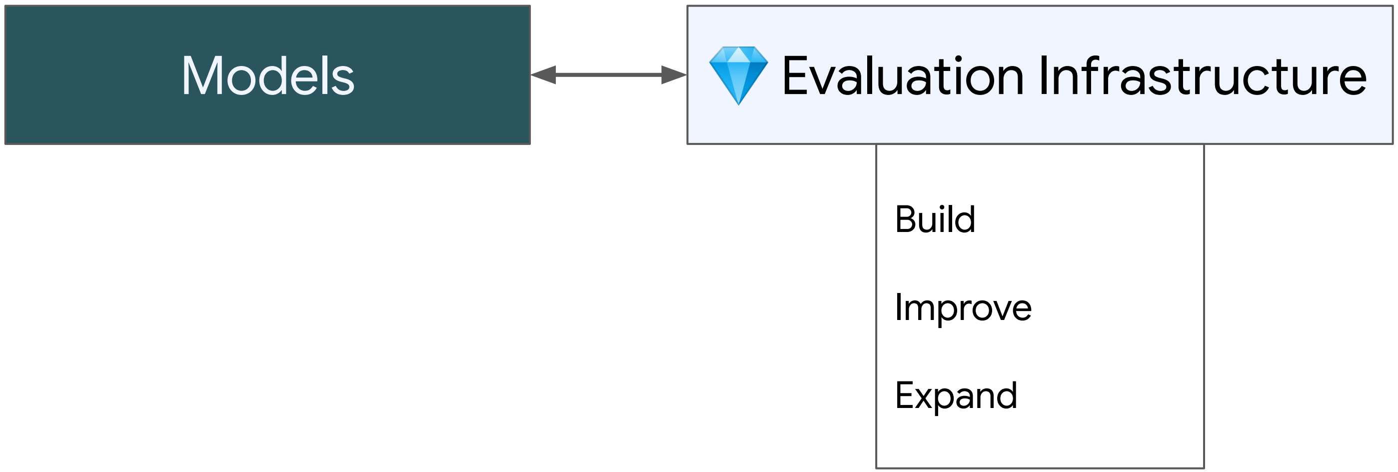
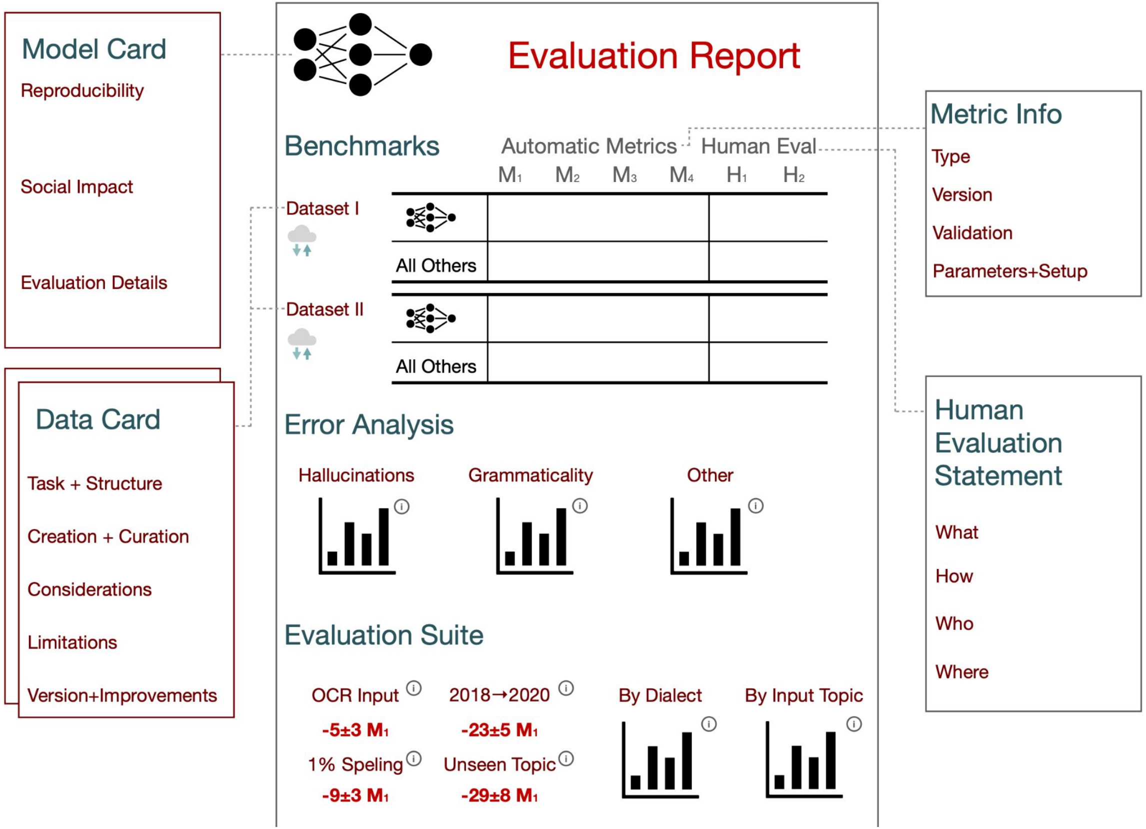


Figure from gem-benchmark.com

Contribute to **evaluation suites**



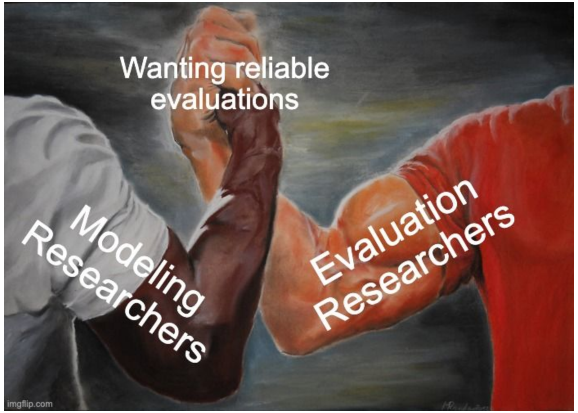
Follow best practices

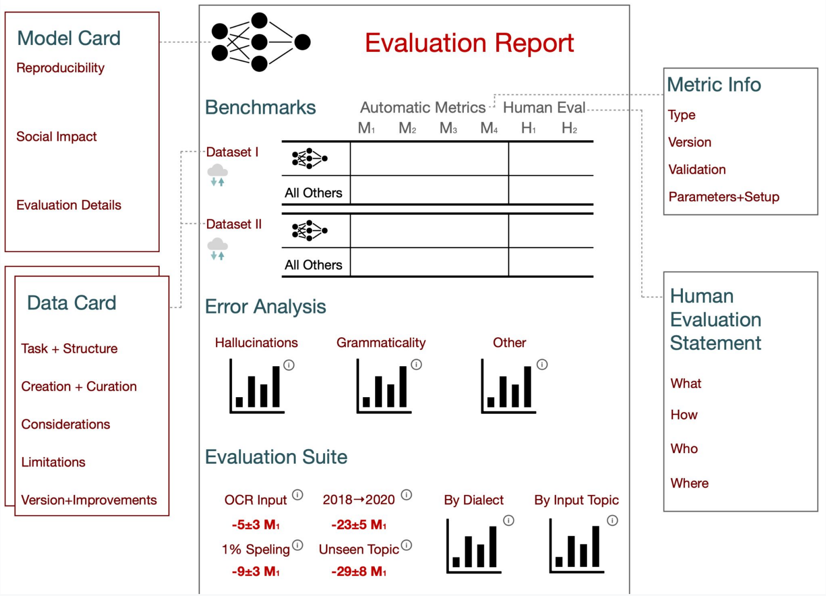
Are you just following the prior work or are you thinking about the evaluation design choices you are making?

Best Practice & Implementation	Yes	No	%
Make informed evaluation choices and document them			
Evaluate on multiple datasets	47	9	83.9
Motivate dataset choice(s)	21	34	38.2
Motivate metric choice(s)	20	46	30.3
Evaluate on non-English language	19	47	28.8
Measure specific generation effects			
Use a combination of metrics from at least two different categories	36	27	57.1
Avoid claims about overall "quality"	34	31	52.3
Discuss limitations of using the proposed method	19	46	29.2
Analyze and address issues in the used dataset(s)			
Discuss or identify issues with the data	19	47	28.8
Contribute to the data documentation or create it if it does not yet exist	1	58	1.7
Address these issues and release an updated version	3	10	23.1
Create targeted evaluation suite(s)	14	52	21.2
Release evaluation suite or analysis script	3	63	4.5
Evaluate in a comparable setting			
Re-train or -implement most appropriate baselines	40	19	67.8
Re-compute evaluation metrics in a consistent framework	38	22	63.3
Run a well-documented human evaluation			
Run a human evaluation to measure important quality aspects	48	18	72.7
Document the study setup (questions, measurement instruments, etc.)	40	9	81.6
Document who is participating in the study	28	20	58.3
Produce robust human evaluation results			
Estimate the effect size and conduct a power analysis	0	48	0.0
Run significance test(s) on the results	12	36	25.0
Conduct an analysis of result validity (agreement, comparison to gold ratings)	19	29	39.6
Discuss the required rater qualification and background	10	38	20.8
Document results in model cards			
Report disaggregated results for subpopulations	13	53	19.7
Evaluate on non-i.i.d. test set(s)	14	52	21.2
Analyze the causal effect of modeling choices on outputs with specific properties	16	50	24.2
Conduct an error analysis and/or demonstrate failures of a model	15	51	22.7
Release model outputs and annotations			
Release outputs on the validation set	1	65	1.5
Release outputs on the test set	2	63	3.1
Release outputs for non-English dataset(s)	1	25	3.8
Release human evaluation annotations	1	47	2.1

Gehrmann et al., 2022

Thank you!





Sebastian Gehrmann
Google Research
s.gehrmann@outlook.com
@SebGehr