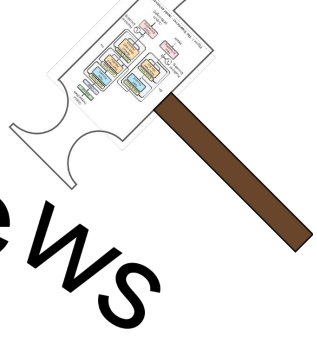
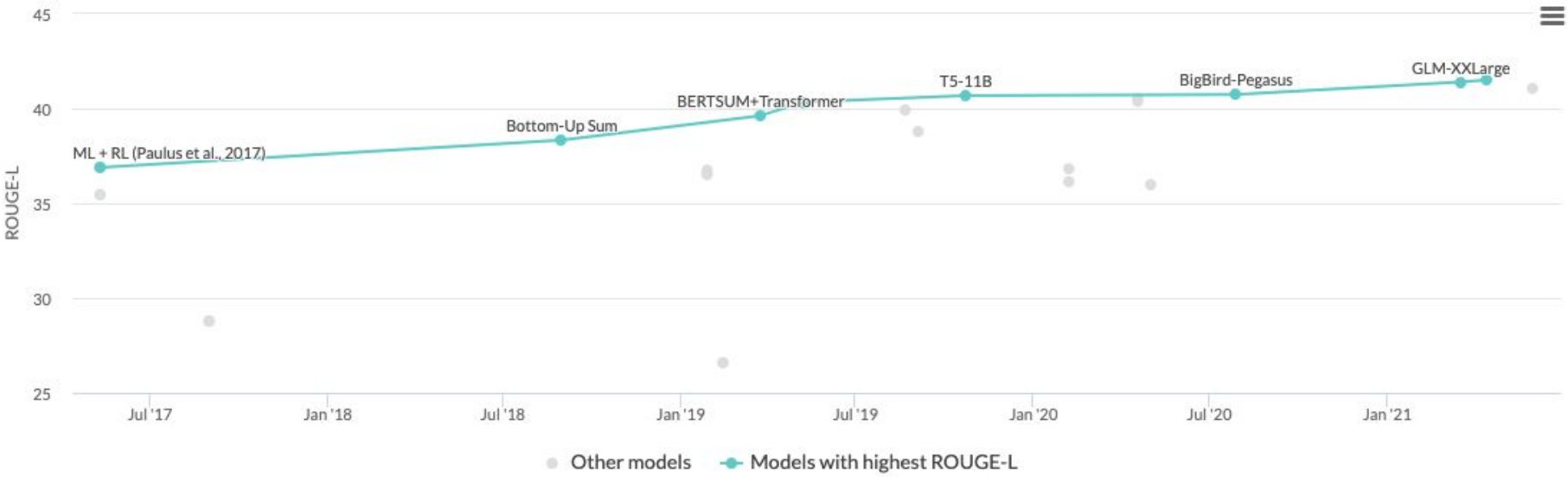


Breaking News

It’s time to fix evaluation of generated text

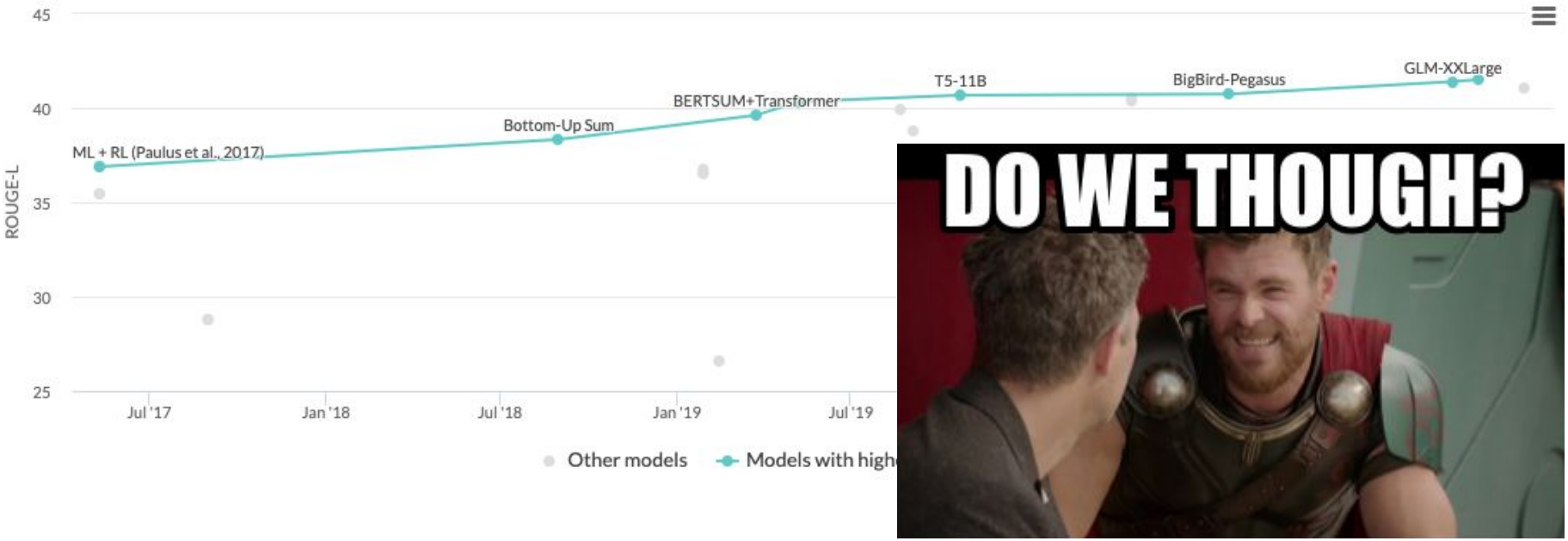
Sebastian Gehrmann
gehrmann@google.com
@SebGehr
Google Research



Good News! We are making progress toward solving summarization.

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

Sebastian Gehrmann, Google Research, 2021



Good News! We are making progress toward solving summarization.

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

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

Candidates

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the California band Grateful Dead.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer from Seattle, Washington.

Dhingra et al., 2019

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

Metrics prefer bad generations over good ones.

Dhingra et al., 2019

ROUGE and its problems...

"ROUGE may **not be a good method** for measuring the usefulness of summaries **when the summaries are not extractive**."
Dorr et al., 2005

A system's ability to produce human-like outputs may be completely **unrelated to its effect on human task-performance**.
Belz+Gatt, 2008

Metrics may provide a useful measure of language quality, although the evidence for this is not as strong as we would ideally like to see; however, **they do not provide a useful measure of content quality**.
Reiter+Belz, 2009

Luckily, we are not using ROUGE to measure content quality of abstractive summaries, right? ..... Right?

Agenda

01 A brief preview

02 Automatic evaluation is broken

03 Human evaluation is broken

04 Datasets are broken

05 How do we fix things?

02

Automatic Evaluation is broken

Sebastian Gehrmann, Google Research, 2021

We suspect that ROUGE is not great.

So let’s see what people are using.

Bottom-Up

Method	R-1	R-2	R-L

Pointer-Generator

ROUGE			METEOR	
1	2	L	exact match	+ stem/syn/para

GPT-2

	R-1	R-2	R-L	R-AVG

Big Bird

Model	Arxiv			PubMed			BigPatent		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L

UniLM

	RG-1	RG-2	RG-L

HiBERT

Model	R-1	R-2	R-L

T5

CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L

Gigaword Abstractive Model

	RG-1	RG-2	RG-L

BERTSum

Model	R1	R2	RL
ORACLE	52.59	31.24	48.87
LEAD-3	40.42	17.62	36.67
Extractive			
SUMMARUNNER (Nallapati et al., 2017)	39.60	16.20	35.30
REFRESH (Narayan et al., 2018b)	40.00	18.20	36.60
LATENT (Zhang et al., 2018)	41.05	18.77	37.54
NEUSUM (Zhou et al., 2018)	41.59	19.01	37.98
SUMO (Liu et al., 2019)	41.00	18.40	37.20
TransformerEXT	40.90	18.02	37.17
Abstractive			
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38
DRM (Paulus et al., 2018)	39.87	15.82	36.90
BOTTOMUP (Gehrmann et al., 2018)	41.22	18.68	38.34
DCA (Celikyilmaz et al., 2018)	41.69	19.47	37.92
TransformerABS	40.21	17.76	37.09
BERT-based			
BERTSUMEXT	43.25	20.24	39.63
BERTSUMEXT w/o interval embeddings	43.20	20.22	39.59
BERTSUMEXT (large)	43.85	20.34	39.90
BERTSUMABS	41.72	19.39	38.76
BERTSUMEXTABS	42.13	19.60	39.18

Summarization is dominated by ROUGE-1, -2, and -L.

Fun fact: The selection was popularized by Rush et al. (2015), who picked a subset of the DUC-2004 options which also included 3, 4, and LW.
But ROUGE-2 and ROUGE-SU4 were used in later DUC challenges.

A very scientific survey.

I read 20 modeling-focused summarization papers from ACL 2021 and recorded the following evaluation aspects:

1) Automatic metrics
2) Human evaluation criteria [if applicable]
3) Dataset(s)

Throughout the talk, I will show the results.

A very scientific survey.

I read 20 modeling-focused summarization papers from ACL 2021 and recorded the following evaluation aspects:

1) Automatic metrics
2) Human evaluation criteria [if applicable]
3) Dataset(s)

Throughout the talk, I will show the results.
On the right, you can see the metrics.

ROUGE 20 ← **100%**
BERT-score 7 ← **Never validated for summ.**
FeQA 1 ←
QAGS 1 ← **Summarization Metrics!**
MoverScore 1
Other-Diversity 2
Other-Entailment 1
Other-Faithfulness 2
Some kind of human eval* 9 ← **<50%**

But just how bad is ROUGE?

And how do you evaluate a metric?
Let’s look at some more recent studies.

ROUGE F-Scores may not be enough.

Recall, that the use of F-scores for ROUGE 1, 2, and L is essentially arbitrary. It may also be strictly suboptimal.

In a study correlation of assessment scores of all possible 192 ROUGE configurations found that the best performing one was to use BLEU instead.[1]

The best ROUGE was ROUGE-2 precision with stemming and removed stopwords.

→ If using ROUGE, consider reporting fine-grained scores.

N-gram Count	
R-3	28.7
R-2	25.0
R-4	18.8
R-1	7.5
R-L	7.5
R-W	7.5
R-S4	2.5
R-SU4	2.5

Stemming	
Not Stemmed	53.8
Stemmed	46.2

Stop-words	
Not Rem.	56.2
Removed	43.8

Summary-level Agg.	
Prec.	52.5
F-score	25.0
Recall	22.5

System-level Agg.	
Average	63.7
Median	36.3

Table 2: Proportions of optimal ROUGE variants attributed to each ROUGE configuration option (%).

Graham, 2015

[1]Not statistically significant, though.

How can we evaluate with lexical overlap, if humans don't even agree with each other?

Data: 100 samples from the CNN/DM test set.

Unconstrained: Every annotator selects sentences in the input they consider important.

Constrained: Every annotator selects sentences with answers to three questions related to the document.

Even when only 3/5 people have to agree on a sentence, there is 0.6 sentences per document on which all agree.

→ When there is only one reference, we can't use lexical overlap to capture everyone's summarization preferences.

Human vote threshold	Sent. per article considered important	
	Unconstrained	*Constrained*
$= 5$	0.028	0.251
≥ 4	0.213	0.712
≥ 3	0.627	1.392
≥ 2	1.695	2.404
≥ 1	5.413	4.524

Kryscinski et al., 2019

The correlation between human judgements and ROUGE is poor.

For 100 CNN/DM test examples, ask 5 raters to judge:

- **Relevance**: selection of important content from the source
- **Consistency**: factual alignment between the summary and the source
- **Fluency**: quality of individual sentences
- **Coherence**: collective quality of all sentences.

All 5 judgements are averaged.

Then, measure Pearson's correlation coefficients and Kendall rank correlation coefficients between judgements and ROUGE.

	Pearson correlation			Kendall rank correlation		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
Relevance	0.03	0.02	0.02	0.28	0.29	0.27
Consistency	0.02	0.01	0.01	0.28	0.28	0.27
Fluency	0.05	0.03	0.04	0.26	0.28	0.27
Coherence	0.05	0.04	0.05	0.27	0.27	0.27

Kryscinski et al., 2019

Repeating the evaluation at scale does not results in (much) better results

Same dataset + criteria, but 8 annotations per example (5 Mturkers, 3 experts) and 16 systems.

Results are similarly not great.

Some “unconventional” summarization metrics like ROUGE-3 and METEOR perform better than “standard” ROUGE settings.

→ We need better 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
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

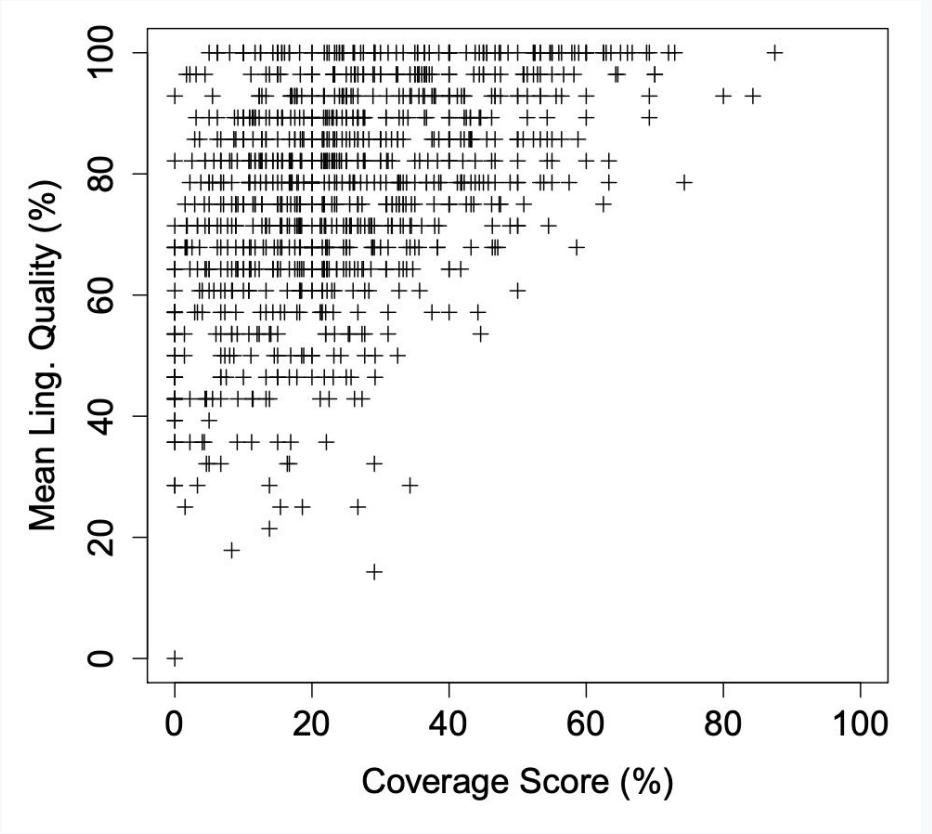
Fabbri et al., 2021

A single metric is not enough.

Human annotations of DUC-2004 ⟶ almost no correlation between linguistic quality and coverage, but coverage is almost never higher than linguistic quality.

This finding is consistent with Pitler et al. (2010) who find correlations between some evaluation categories, but not between linguistic and content quality.

⟶ We cannot rely on a single metric to provide all details.

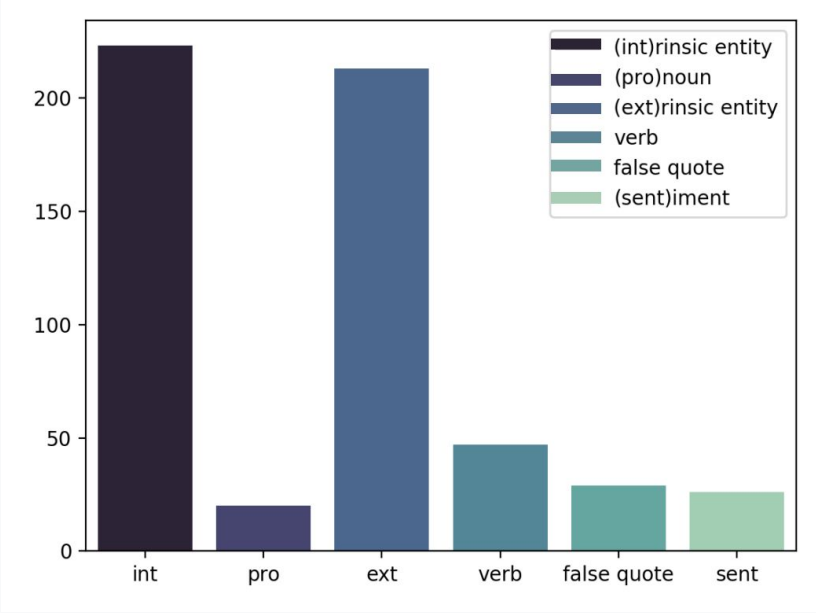


Graham, 2015

We can go deeper. Can we audit metrics?

If we know common hallucinations, we can inject them into the references and test if a metric score decreases.

Scores of a well-calibrated metric should negatively correlate with monotonically increasing number of errors.



[Gabriel et al., 2021]

We can go deeper. Can we audit metrics?

If we know c
the referenc

Scores of a
correlate wit

	STANDARD and CONTEXTUAL				
	R-1	**R-2**	**R-3**	**R-L**	**BERTScore**
Upper Bound	10.61	2.56	0.72	9.32	83.76
Level 1	10.49 / 10.76	2.54 / 2.56	0.70	9.22 / 9.42	83.53 / 83.56
Level 2	10.40 / 10.86	2.51 / 2.54	0.69 / 0.68	9.16 / 9.49	83.36 / 83.38
Level 3	10.33 / 10.92	2.49 / 2.52	0.69 / 0.67	9.10 / 9.55	83.21 / 83.26
Lower Bound	5.44	0.39	0.01	4.94	80.08
Correlation	**-1.00** / **0.98**	-0.97 / **-1.00**	**-0.87** / **-1.00**	**-1.00** / **1.00**	**-1.00**
p-value	0.03* / 0.10	0.16 / 0.05*	0.33 / 0.05*	<0.01** / 0.02*	0.02* / 0.06

Left: entity errors, Right: non-entity errors

→ Most metrics are calibrated, but R-1+R-L fail completely.

Also note that this is system-level correlation, not segment.

Gabriel et al., 2021

Let's look deeper into faithfulness.

A model not faithful if it hallucinates.

Intrinsic: A model misrepresents facts in the input
"Former London mayoral candidate" ⟶ *"Former London **mayor**"*
Extrinsic: A model ignores the input
"mayoral candidate Peter" ⟶ *"**mayor Sara**"*
Factual: A model hallucinates facts that are true
"mayoral candidate Peter" ⟶ *"**2016** mayoral candidate Peter"*

Factual hallucinations may be acceptable.

Semantic or lexical similarity does not help for these fine-grained determinations.

⟶ When assessing a model, entailment-type metrics may be necessary to detect hallucinations.

[Maynez, Narayan, et al., 2020]

Metric	Faithful	Factual
ROUGE-1	0.197	0.125
ROUGE-2	0.162	0.095
ROUGE-L	0.162	0.113
BERTScore	0.190	0.116
QA	0.044	0.027
Entailment	**0.431**	**0.264**

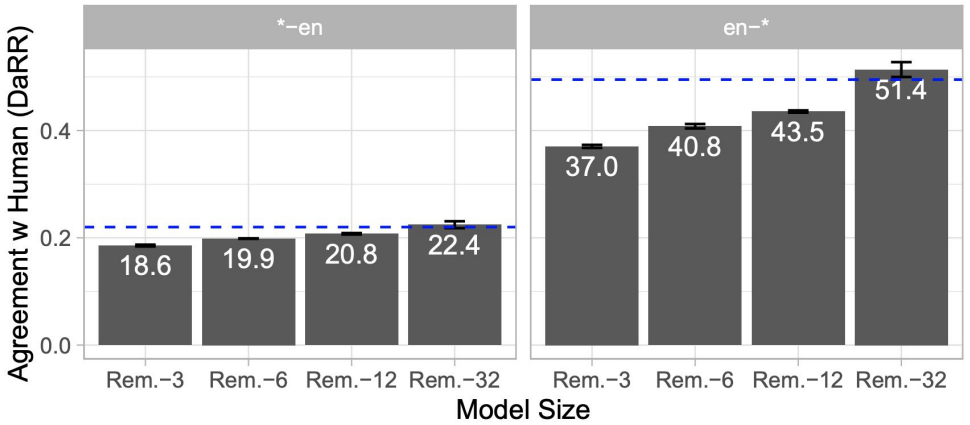
Table 4: Spearman's correlation coefficient ($|r_s|$) of different metrics with faithful and factual annotations.

A glimmer of hope on the horizon

Trained metrics can have much higher correlations.[1]

“Only” requirements:

1) Many high-quality annotations
2) Large pretrained models



DaRR score to- and from- English translations across model sizes.

[Pu et al., 2021]

[1]At least in MT

Let’s build better metrics!

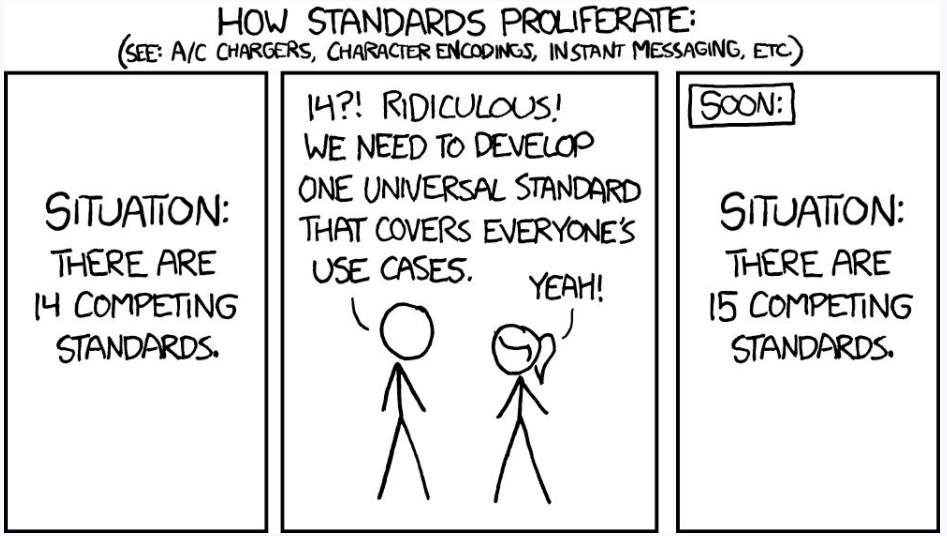
But, how do we get people to adopt it?
It has to be fast, and easy to use, and work for all languages, and all tasks, and ... 🤗

Great, let’s do that!
Not so fast! We need high-quality data first.

How do we get that?
Human evaluation.

Do we know how to do human eval?
No, not really.

But you said that human eval is necessary!
Let’s move to the next section.



Sebastian Gehrmann, Google Research, 2021

03

Human Evaluation is broken

Takeaways so far

One number cannot characterize all performance aspects of a model output
→ We need **multiple specialized metrics**.

None of our metrics correlate well with human judgements
→ **Human evaluation is** a **necessary** component of model evaluation.

Trained metrics can potentially have much better correlations
→ We need **many** high quality human annotations.

Coming back to the survey

9/20 papers used human evaluation.

But what were they assessing? 👉

Wide range of criteria, there is no agreement here.

And the problem runs even deeper.

Informativeness 5
Conciseness/Succinctness 4
Fluency 4
Relevance/Salience 4
Coherence 2
Consistency 2
Coverage 1
Error Classifications 1
Factualness 1
Faithfulness 1
Grammaticality 1
Meaning-Preserving 1

What is being measured?

In a study of 478 INLG papers, the authors found:

- 204 unique names of quality criteria.
- **71 truly different aspects.** 👉

Similar aspects may be considered equal by readers:
Spelling Accuracy vs. Correctness of the Surface Form

Often, details are not provided:

- >50% missing definitions (279/478)
- ~66% missing evaluator prompts/questions (311/478)
- 20% missing criteria names (98/478)

⟶ We need to understand what is measured and not group different annotations into one bucket

Howcroft et al., 2020

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.

How is it being measured?

Form	Count
direct quality estimation	207
relative quality estimation	72
(dis)agreement with quality statement	48
classification	38
task performance measurements	35
qualitative feedback	20
evaluation through post-editing/annotation	18
unclear	15
user-system interaction measurements	10
counting occurrences in text	8
user-text interaction measurements	6
other	1

Howcroft et al., 2020

How is it being measured?

Positive and Negative Framing

How much more fluent is sentence A versus sentence B?

⟶ implicitly prime rater that A is better than B

Demand Characteristics

We consider sentences that end with "." as more formal than sentences that end with "!"

⟶ Biases raters to pay more attention to model artifacts

Anchoring and Adjusting

Select sentences from model A as examples in the instruction

⟶ Biases raters to prefer outputs that look like A over B.

Schoch et al., 2020

Form	Count
direct quality estimation	207
relative quality estimation	72
(dis)agreement with quality statement	48
classification	38
task performance measurements	35
qualitative feedback	20
evaluation through post-editing/annotation	18
unclear	15
user-system interaction measurements	10
counting occurrences in text	8
user-text interaction measurements	6
other	1

Howcroft et al., 2020

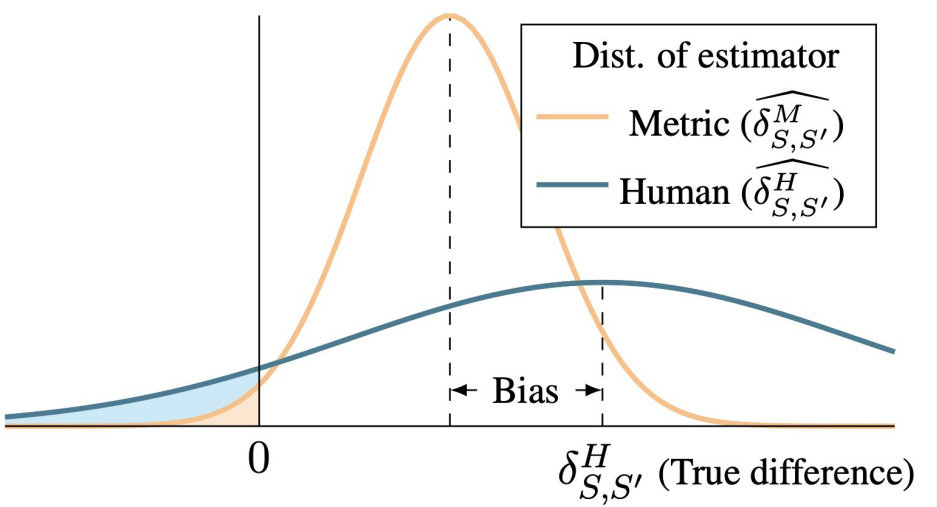
How many annotations do we need?

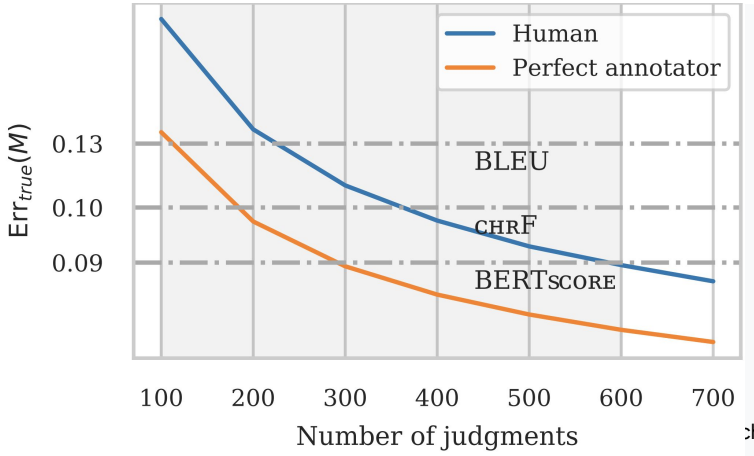
Humans measure the “true” difference between two systems, but have **high variance**. Metrics have lower variance, but are **biased**. Both are sources of errors.

As models get better, the differences between them get smaller. As a result, we need more annotator judgements.

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

Yet, most annotations in my survey had n=100 or smaller.





[Wei + Jia, 2021]

ch, 2021

Who is measuring?
And why may this be a problem?

Some aspects are easier to assess without professional raters (linguistic quality vs. content quality).

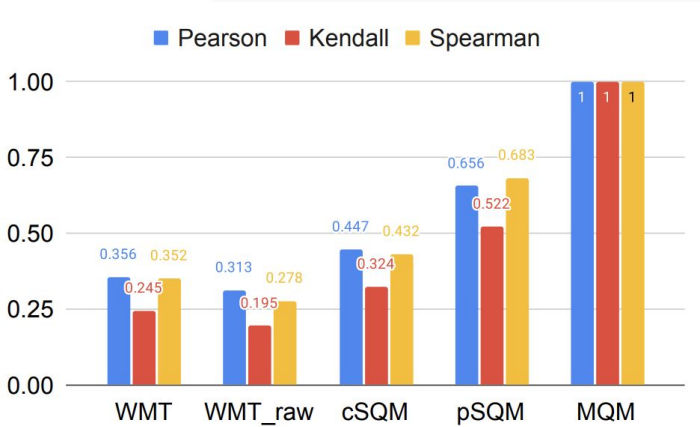
Crowdworkers tend to much have a higher variance than professional raters

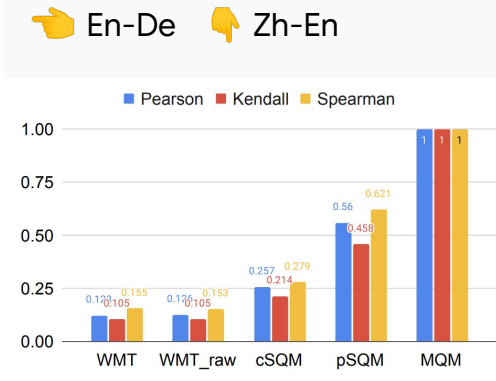
Agreement between ratings produced by linguists and those from crowdworkers can be extremely low.

Eval	Judges	Topics	Systems
TAC	0.28	0.40	0.13
MTurk	0.44	0.13	0.05

Table 4: Linear regression is used to model Overall Quality scores as a function of judges, topics, and systems, respectively, for each data set. The R^2 values, which give the fraction of variance explained by each of the six models, are shown.

MTurk workers also had a much higher correlation between linguistic and overall quality than experts. Gillick + Liu, 2010





Freitag et al., 2021

We need methods to deal with noisy ratings.

Text Analysis Conference Summarization track evaluation:

- Each assessor is assigned to a topic and evaluates all summaries, even duplicate ones
- We can identify within-annotator consistency

CLASSY is a (non-neural!) logistic regression model trained on these ratings

→ Excluding the most inconsistent annotated data can lead to higher correlation.

	NoModels		AllPeers	
	main	update	main	update
Pyramid				
CLASSY1_Pyr	0.956	0.898	0.945	0.936
CLASSY1_Pyr_new (a)	0.950	0.895	0.932	**0.955**
CLASSY1_Pyr_new (b)	**0.960**	**0.900**	0.940	**0.955**
Responsiveness				
CLASSY2_Resp	0.951	0.903	0.948	0.963
CLASSY2_Resp_new	**0.954**	**0.907**	**0.973**	0.950
CLASSY4_Resp	0.951	0.927	0.830	0.949
CLASSY4_Resp_new	0.943	**0.928**	**0.887**	0.946
Readability				
CLASSY3_Read	0.768	0.705	0.844	0.907
CLASSY3_Read_new	**0.793**	**0.721**	**0.858**	0.906

Numbers are correlation between output and measure of the subsection.

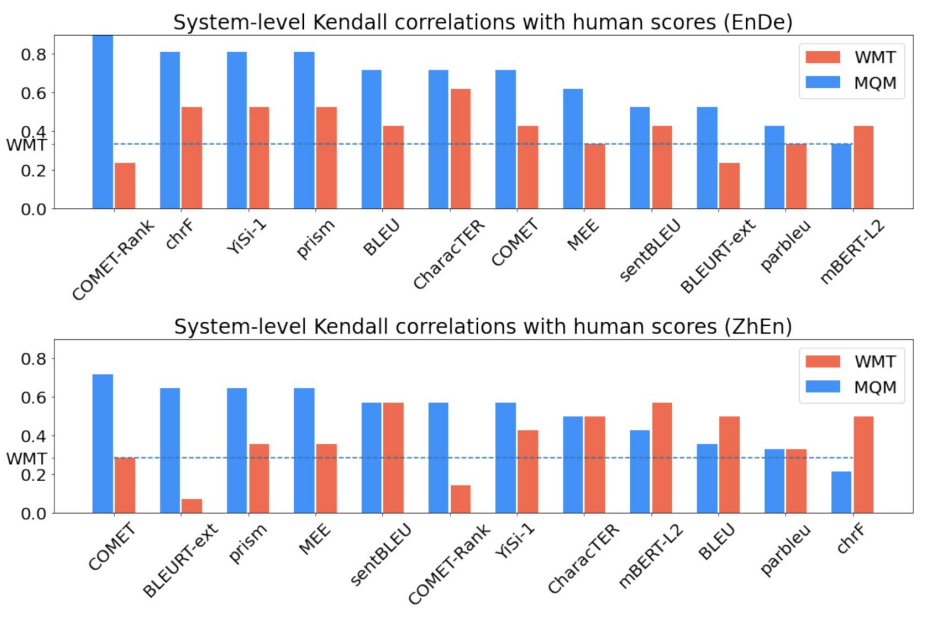
Owczarzak, et al., 2012

What do noisy ratings mean for metrics?

Surprise #1
Metrics agree more with the high-quality annotations than with noisy ones, despite being trained on noisy annotations.

Surprise #2
Metrics have a higher agreement with MQM than WMT has with MQM.

~~**Surprise #3**~~ **not really**
Previous findings about metric quality rankings are wrong.



[Freitag et al., 2021]

Note that this is for MT, not summarization.

What can we do about this?

We don't know what is being measured and how.
→ Write human evaluation datasheets (Shimorina + Belz, 2021)

None of our results are statistically significant.
→ Estimate the effect size before running evaluations and use significance tests to verify results.

Expert raters provide much better results than crowdworkers.
→ Verify crowdsourcing results multiple times and think what qualifications are required for what you want to measure.

→ Don't treat human evaluation as the ultimate answer to all evaluation problems

So far...

- Our metrics don’t measure what we want (at least not well).
- Human evaluation can help evaluate models and develop metrics, but only in theory.

What about our datasets?
What does a better score on dataset X mean?

04

Datasets are broken

Sebastian Gehrmann, Google Research, 2021

So far...

- Our metrics don't measure what we want (at least not well).
- Human evaluation can help evaluate models and develop metrics, but only in theory.

What about our datasets?
What does a better score on dataset X mean?

And what is this dataset X?
Let's look at the survey.

CNN/DM 5
AMI 4
XSum 4
Amazon Review Dataset 2
WikiSum 2

ArXiv / PubMed 1
BIGPATENT 1
CQASumm 1
DUC QFS 1
EmailSum 1
En2ZhSum/Zh2EnSum 1
FacetSum 1
Justice (Chinese) 1
MATINF (google translate to En) 1
Medical 1
MeQSum 1
Multi-News 1
NYT 1
NYT-Comments 1
Reddit 1
SAMSum 1
Spotify Podcast 1
StackExchange 1
TD-QFS 1
Timeline Summarization 1
W3C 1
Yelp Review Dataset 1

What does the survey tell us?

- 27 different datasets in 20 papers
- Only two non-English datasets
- CNN/DM remains the most popular dataset

→ How can we as a field make progress on improving summarization if we don't have a (good) standard benchmark?
→ Also, how can we say we are making progress if we focus on a single language?

CNN/DM 5
AMI 4
XSum 4
Amazon Review Dataset 2
WikiSum 2

ArXiv / PubMed 1
BIGPATENT 1
CQASumm 1
DUC QFS 1
EmailSum 1
En2ZhSum/Zh2EnSum 1
FacetSum 1
Justice (Chinese) 1
MATINF (google translate to En) 1
Medical 1
MeQSum 1
Multi-News 1
NYT 1
NYT-Comments 1
Reddit 1
SAMSum 1
Spotify Podcast 1
StackExchange 1
TD-QFS 1
Timeline Summarization 1
W3C 1
Yelp Review Dataset 1

Problem #1: Noise

Reference summaries often contain extraneous information, such as hyperlinks and click-bait descriptions of other articles

Raters prefer lead-3 over the CNN/DM reference.

→ Can we expect faithful models if our data is not?

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

Doesn’t that make the whole CNN/DM task pointless?

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

Fabbri et al., 2021, Stiennon, Ouyang, Wu, Ziegler et al., 2020, Maynez, Narayan, et al., 2020

Problem #2: Splits

Results look completely different depending on how the test set was constructed.

A good model should do well on all expected data during deployment in a live scenario. Not just i.i.d. data.

		Splits				
Task	**Model**	**Standard**	**Random**	**Heuristic**	**Adversarial**	**New Samples**
HEADLINE GENERATION*	seq2seq	0.073	0.095	0.062	**0.040**	0.069

Søgaard et al., 2021, Ribeiro et al., 2020

Problem #2: Splits

Results look completely different depending on how the test set was constructed.

A good model should do well on all expected data during deployment in a live scenario. Not just i.i.d. data.

But, most datasets only have one test set.
How do we test calibration?

→ We need focused challenge sets to test capabilities.

		Splits				
Task	**Model**	**Standard**	**Random**	**Heuristic**	**Adversarial**	**New Samples**
HEADLINE GENERATION*	seq2seq	0.073	0.095	0.062	**0.040**	0.069

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%			

Sebastian Gehrmann, Google Research, 2021

Søgaard et al., 2021, Ribeiro et al., 2020

Problem #3: New Concepts

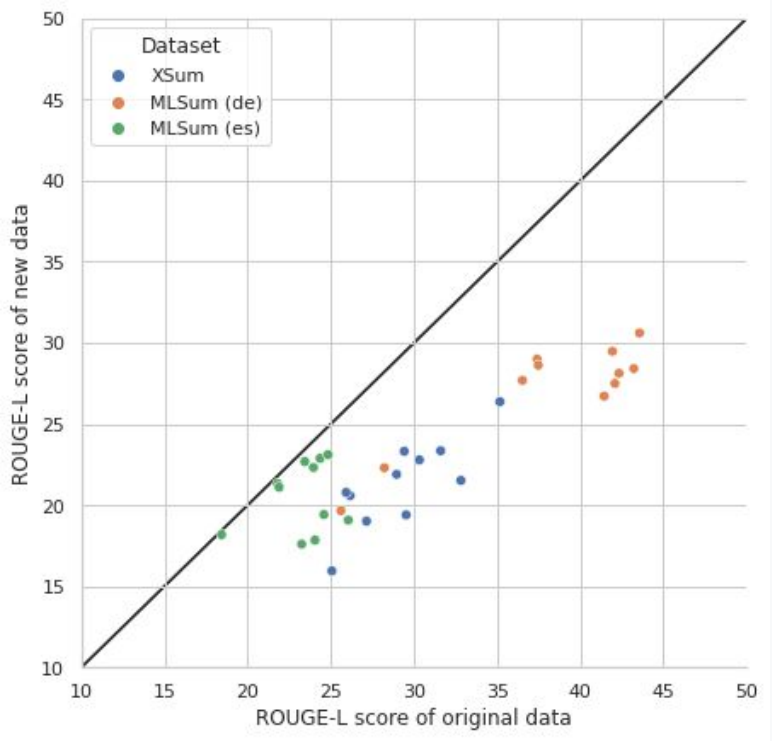
Training sets usually remain static, but real test data does not.

How does a model perform for new concepts?

We created 3 test sets for pre-2020 datasets:

- XSum (En)
- MLSum (De)
- MLSum (Es)

Original collection method, but COVID-19 related articles.



Each dot represents one model.

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

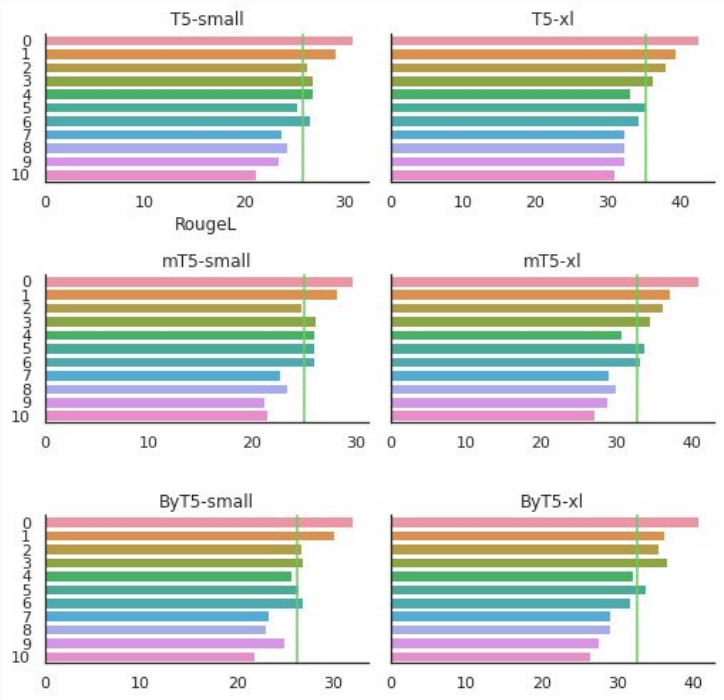
Problem #4: Style

Performance should not depend on the reference style.

We split the XSum test set into 10 buckets depending on reference abstractiveness.

The more abstractive a reference, the lower the score.

Similar finding in MT (Freitag et al., 2020)



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

What can we do about this?

→ Document limitations, issues, and social impact (Gebru et al., 2018, Bender + Friedman, 2018).[1]

→ Create evaluation suites instead of i.i.d. test sets.

→ Evaluate worst-case performance, not only average.

→ Think of a dataset, its splits, and documentation as a “living” object instead of a static entity.

[1]We released an NLG-specific template in McMillan-Major et al., (2021)

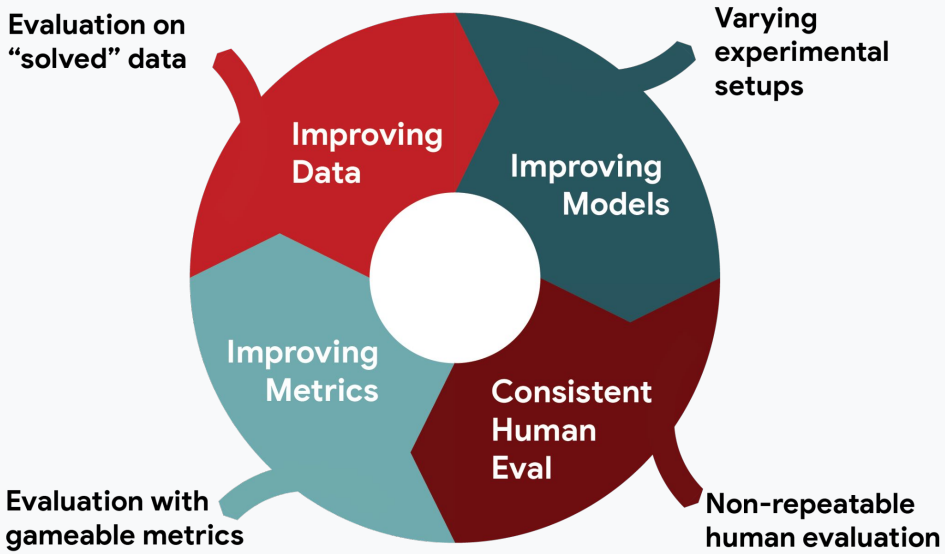
04

So how do we fix things?

We need to break through this circular dependency.

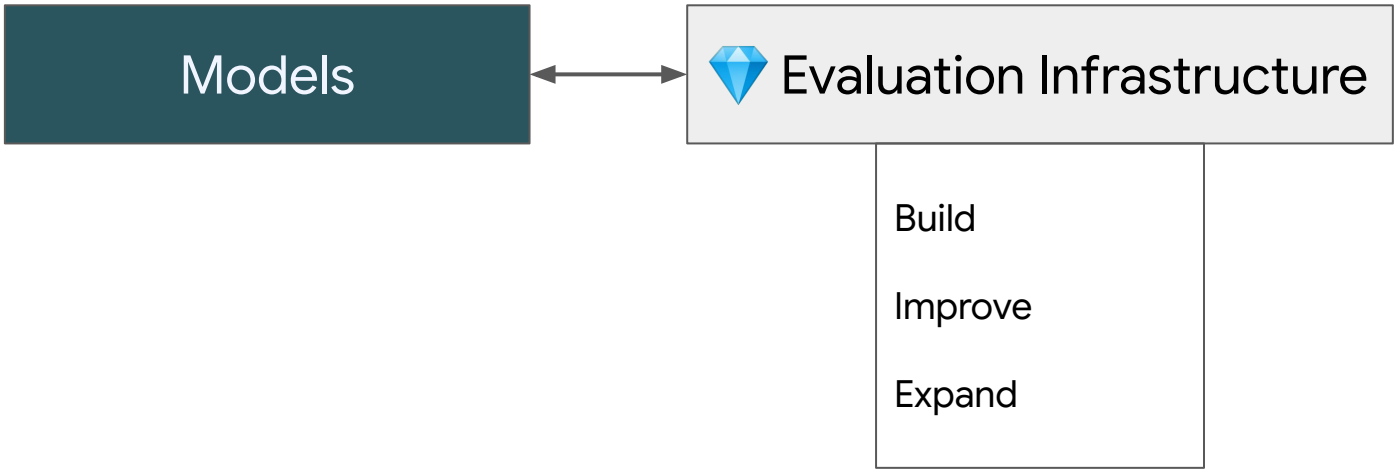
At the moment, we can't identify whether and how our models **fail**, or whether failure is **attributable** to the data, model, or evaluation.

→ A single researcher cannot solve every problem. We thus need easy-to-use infrastructure to stay up to date with the latest developments, combining everyone's strengths.



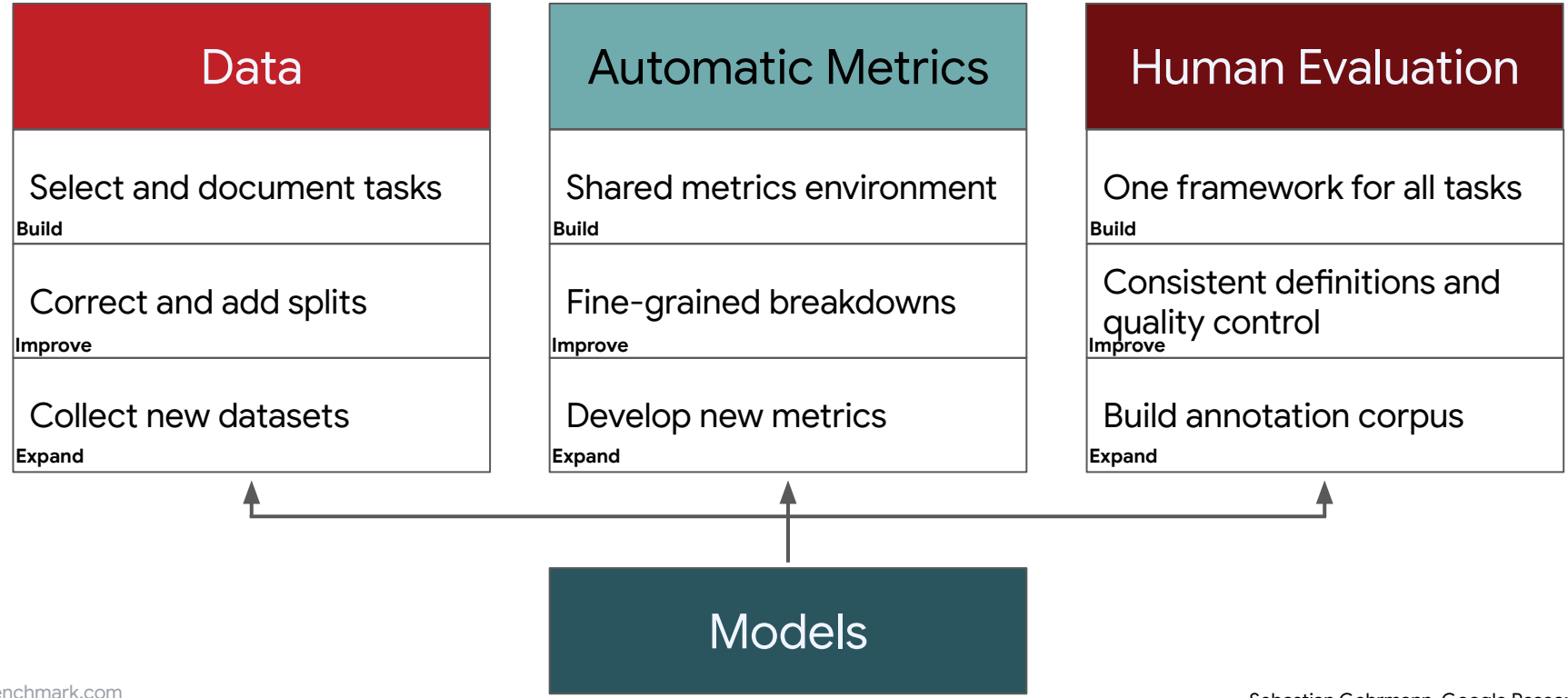
Sebastian Gehrmann, Google Research, 2021

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

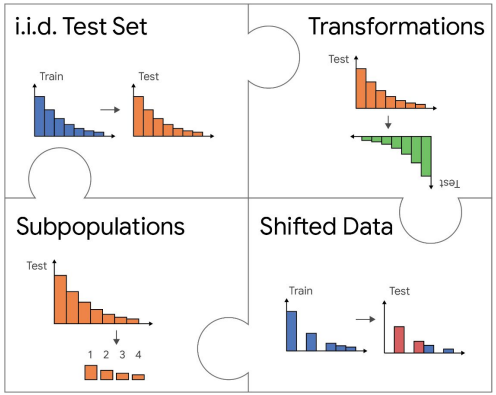


Instead of dictating what should be used, let's make it easy to explore the right way to do things.

What does building infrastructure entail?



Let’s unbreak data



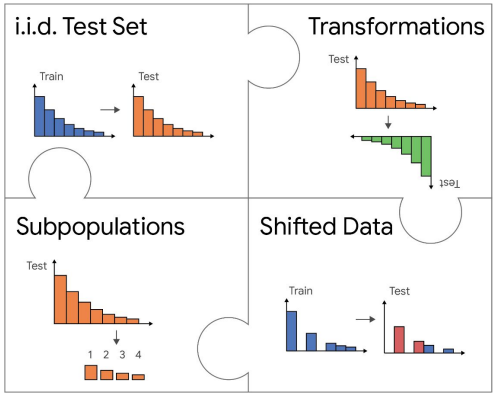
Develop transformations and filters of datasets to test robustness and performance on subpopulations.

Instead of chasing the highest number, try to break models.

More infos at https://gem-benchmark.com/nl_augmenter 🦎 → 🐍
Also, https://robustnessgym.com/ and many others.

Sebastian Gehrmann, Google Research, 2021

Let's unbreak data



Develop transformations and filters of datasets to test robustness and performance on subpopulations.

Instead of chasing the highest number, try to break models.

More infos at https://gem-benchmark.com/nl_augmenter 🦎 → 🐍
Also, https://robustnessgym.com/ and many others.

LIST OF TASKS

The list below links to data statements [1, 2] for each of the datasets that are part of GEM tasks. The template used to produce the statements and a guide on how to write them can be found here: [download template] [view guide].

MLSum summarization
Large-scale multilingual dataset for evaluating abstractive summarization

XSum Summarization
Large scale monolingual dataset for evaluating extreme summarization.

WikiLingua Summarization
Large-scale multilingual dataset for evaluating cross-lingual abstractive summarization

WebNLG Structure-to-text
The WebNLG dataset is a large bi-lingual dataset with crowdsourced reference texts and a rather large variety of knowledge in the inputs. A web-based evaluation platform is already existing.

CommonGen Structure-to-text
A medium sized corpus with a unique reasoning challenge and interesting evaluation possibilities.

E2E Structure-to-Text
One of the largest limited-domain NLG datasets and is frequently used as a data-to-text generation benchmark.

DART Structure-to-Text
Hierarchical, structured format with its open-domain nature

Czech Restaurant Structure-to-Text
One of a few non-English data-to-text datasets in a well-known domain, covering a morphologically rich language.

ToTTo Structure-To-Text
Controlled Table2Text task with non-divergent, annotator-revised text outputs

Wiki-Auto Simplification
Wiki-Auto is the largest open text simplification dataset currently available. For GEM, Wiki-Auto acts as the training set.

TURKCorpus Simplification
TURKCorpus is a high-quality simplification dataset where each source sentence is associated with 8 human-written simplifications.

ASSET Simplification
ASSET is a high quality simplification dataset where each source (not simple) sentence is associated with 10 human-written simplifications.

Schema-Guided Dialog Dialog
Modeling task-oriented dialog.

Currently 13 documented tasks in 18 languages.
Support for loaders in 🤗Datasets and TFDS.
Soon 30+ tasks across 40+ languages.

More at https://gem-benchmark.com/data_cards

Sebastian Gehrmann, Google Research, 2021

Let’s unbreak metrics

We can use multiple metrics instead of only ROUGE.

Our library computes 100+ statistics and metrics for any generation task.
For supported tasks, it provides fine-grained breakdowns

The library has support for caching, runs non-GPU metrics in parallel, and we are adding many more metrics.

We are hoping that it will make the lives of model developers and metrics researchers easier.

```
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}
    }
}
```

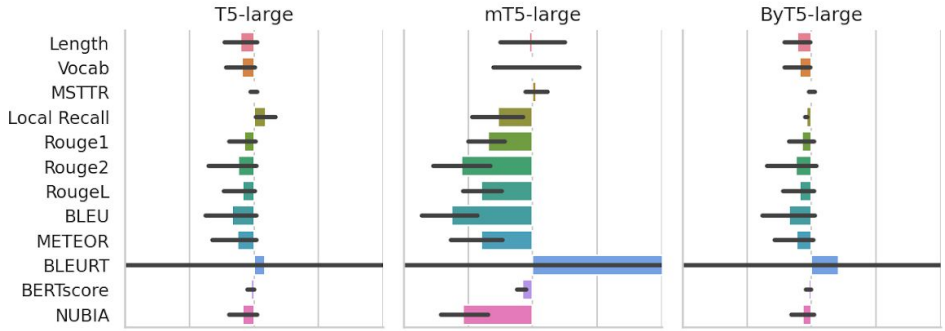
```
python run_metrics.py -s outputs.json  -r targets.json -o predictions.json
```

More at https://github.com/GEM-benchmark/GEM-metrics.
Also, https://github.com/danieldeutsch/repro.

Putting this together - we can develop performance and robustness reports.

Question: How robust is my model to punctuation mistakes?

Answer:



By framing robustness in causal terms and measure multiple response metrics, we can audit models without perfect metric.

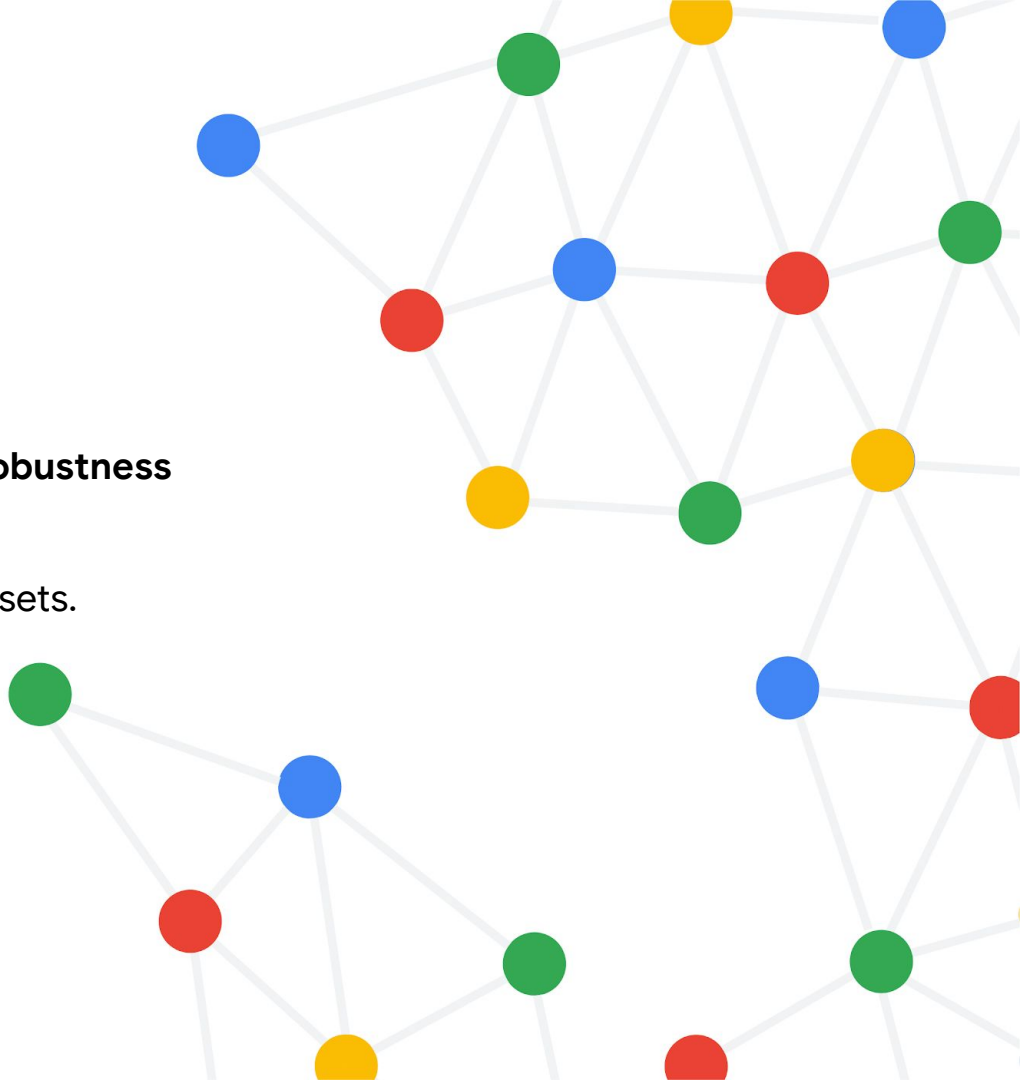
Final Lessons

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 their robustness**

Instead of aiming for higher ROUGE numbers,
let's audit models, evaluation approaches, and datasets.



Sebastian Gehrmann
gehrmann@google.com
@SebGehr
Google Research