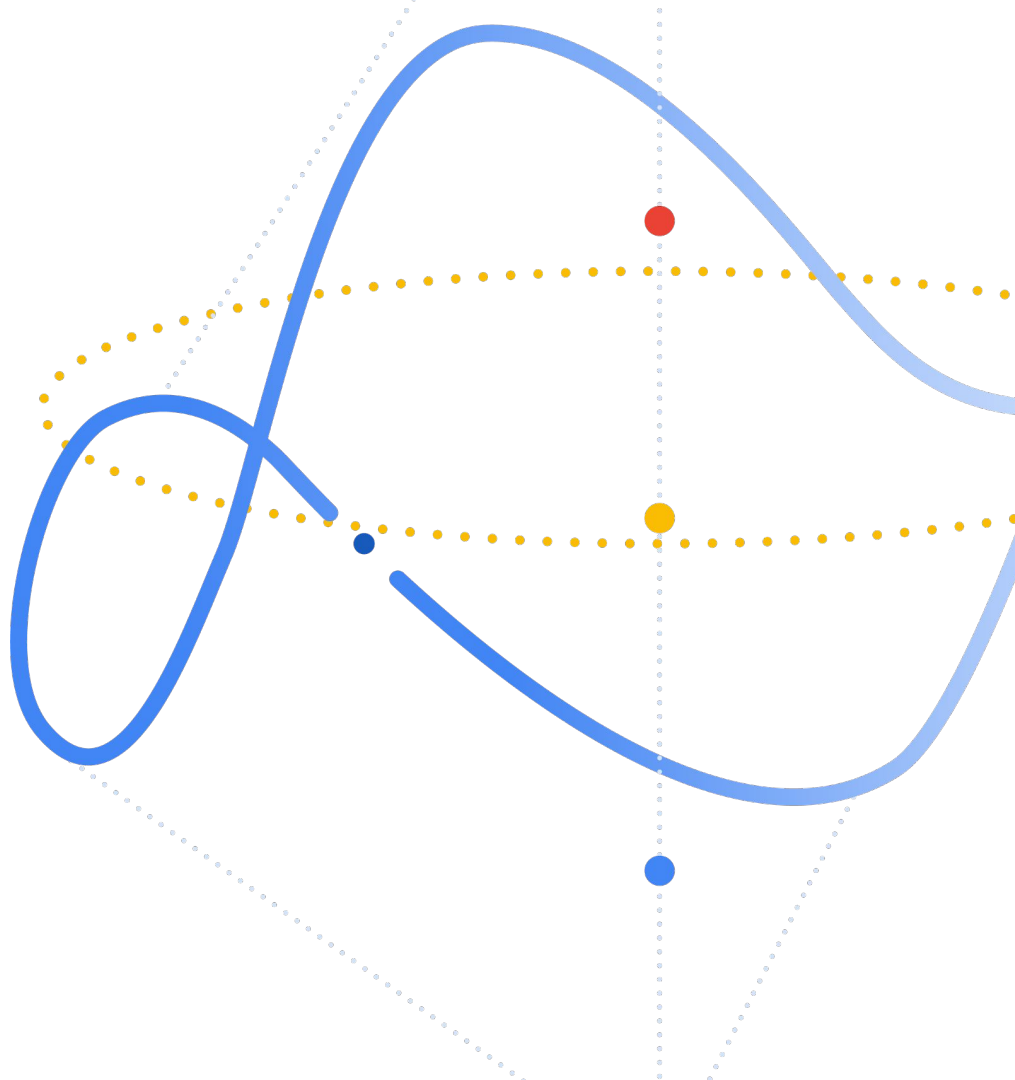


A Case for Better Evaluation Standards in NLG

Sebastian Gehrmann, Elizabeth Clark, Thibault Sellam

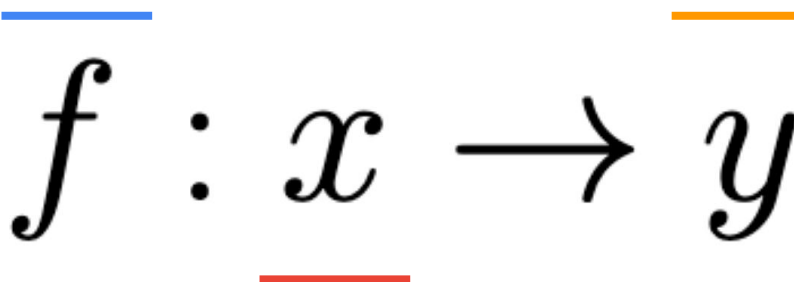
{gehrmann,eaclark,tsellam}@google.com



We all know that ML evaluation has issues.
What makes Natural Language Generation special?

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

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



The diagram shows the mathematical representation of an NLG system as a function $f: x \rightarrow y$. The function symbol f is underlined with a blue line. The input variable x is underlined with a red line. The output variable y is underlined with an orange line.

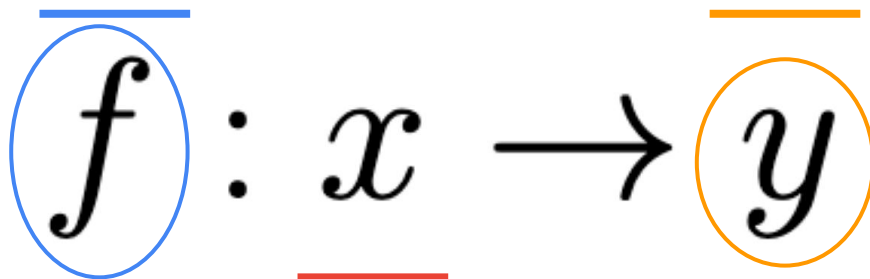
$$f : x \rightarrow y$$

Structured or textual information
that defines the output space

What makes Natural Language Generation special?

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

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

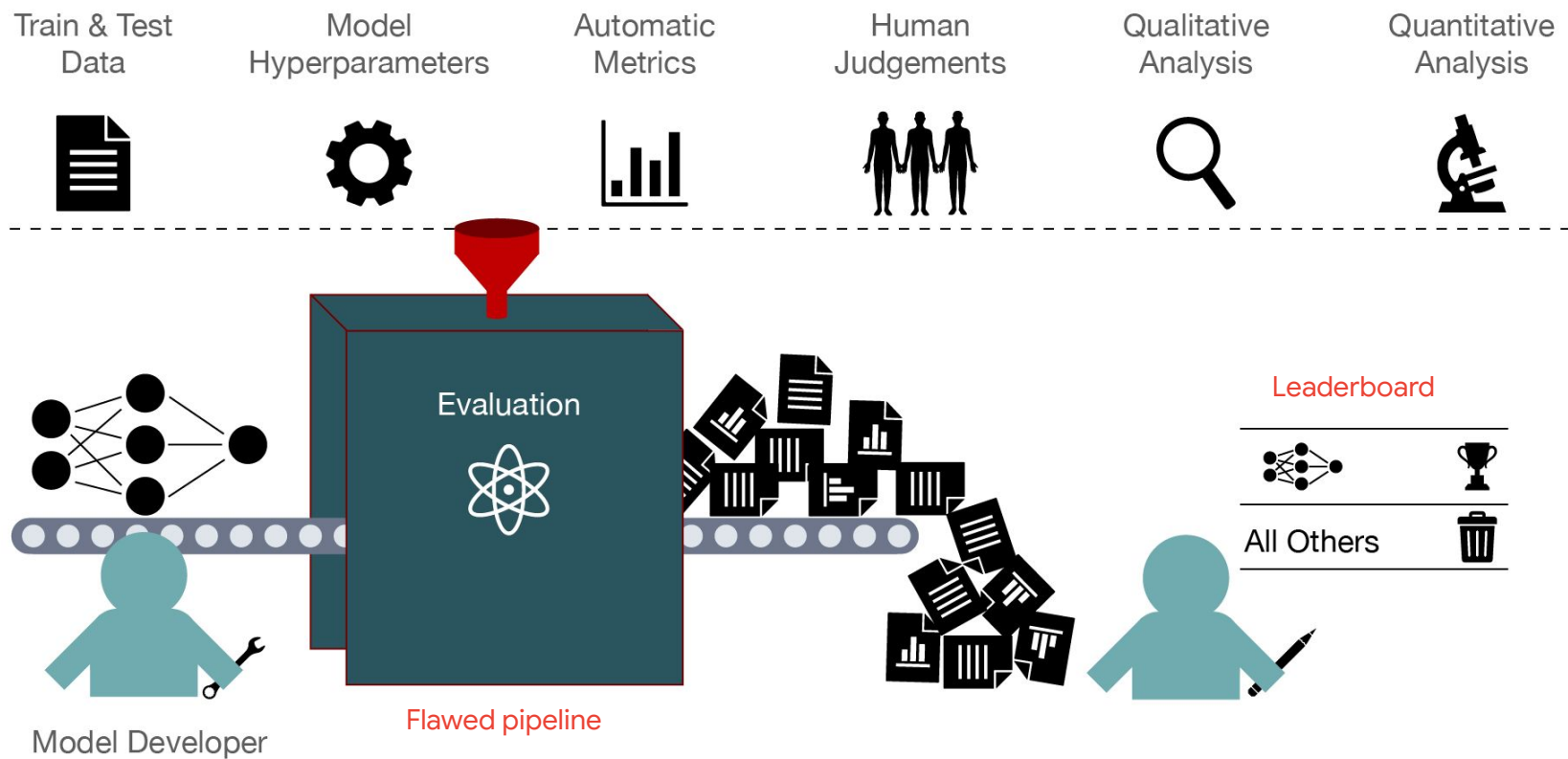


No one size fits all

Structured or textual information
that defines the output space

*Huge output space.
No "accuracy".*

NLG suffers from “leaderboarding” and flawed standardization.



We propose 8 categories of best practices with 29 suggestions.

- 1) They should lead to better evaluations without major additional work (*baby steps...*)
- 2) Whether someone follows them should be easy for reviewers to identify.
- 3) They should be grounded in all the related literature.*

We propose 8 categories of best practices with 29 suggestions.

- 1) They should lead to better evaluations without major additional work (*baby steps...*)
- 2) Whether someone follows them should be easy for reviewers to identify.
- 3) They should be grounded in all the related literature.*

What are they and do people already follow them?

Some highlights of what we found.

Make informed choices and document them.

While 84% of papers evaluate on multiple datasets, only 29% include any non-English ones.

<30% of papers state why they use a particular dataset or metric (*standardization effect!*)

Measure specific effects. Avoid overclaims.

About half of the papers make claims about overall **quality** when this is not what is being measured.

Only 30% of papers discuss limitations.

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

Some highlights of what we found.

Address issues in the data.

Almost 30% point out issues with the data, but we found only 3 papers that something about it.

Only a single paper contributed to data documentation. *We have a long way to go.*

Conduct robust human evaluations.

The median n in human eval is 100, but we know that we need at least 300-500 to get repeatable results!

Only 40% analyzed result validity and 20% discussed whether subjects were qualified for a task.

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

Some highlights of what we found.

Analyze and document the results.

20-25% each reported results for subpopulations, on non-i.i.d. test sets, or conducted error analyses.

Release outputs and annotation.

Almost no model outputs or human annotations were released. This is **crucial** for evaluation research.

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

Takeaways

We can take easy steps to improve our evaluations.

Many of the suggestions are not specific to NLG.

We can only expect better evaluation standards, if we as reviewers hold authors accountable for bad eval practices.

Sebastian Gehrmann, Elizabeth Clark, Thibault Sellam

{gehrmann,eaclark,tsellam}@google.com

