# Researching Responsible and Trustworthy Natural Language Processing

Error Analysis, a.k.a. actually looking at our output

Emily Allaway

2 February 2026

School of Informatics
University of Edinburgh
eallaway@ed.ac.uk

# Overview

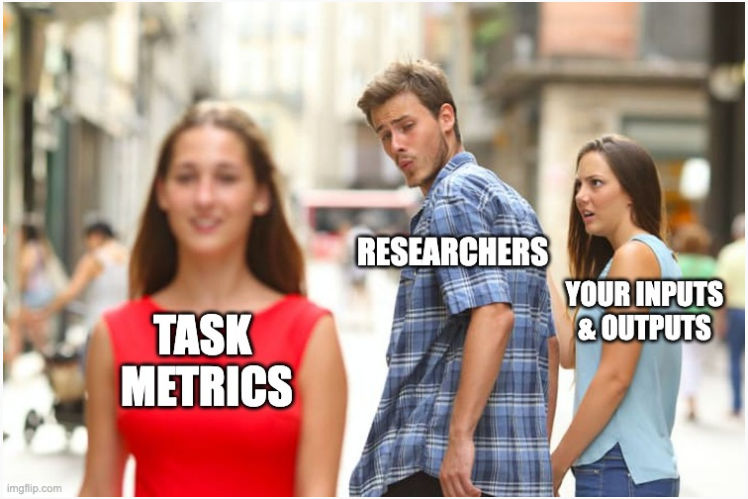
Types of Error Analysis

Exercise

# Overview

Types of Error Analysis

Exercise


TASK
METRICS
RESEARCHERS
YOUR INPUTS
& OUTPUTS
imgflip.com

# What is the point?

# In the beginning (in the 2010s) ...

Features of the model (feature ablation)

| **Classifier** | Atheism | Climate Change | Feminist Movemt. | Hillary Clinton | Legal. of Abortion | F-macroT | F-microT |
|---|---|---|---|---|---|---|---|
| *I. Benchmarks* | | | | | | | |
| a. Random | 31.1 | 27.8 | 29.1 | 33.5 | 31.1 | 30.5 | 33.3 |
| b. Majority | 42.1 | 42.1 | 39.1 | 36.8 | 40.3 | 40.1 | 65.2 |
| c. First in shared task | 61.4 | 41.6 | 62.1 | 57.7 | 57.3 | 56.0 | 67.8 |
| d. Oracle Sentiment | 65.8 | 34.3 | 61.7 | 62.2 | 41.3 | 53.1 | 57.2 |
| e. Oracle Sentiment and Target | 66.2 | 36.2 | 63.7 | 72.5 | 41.8 | 56.1 | 59.6 |
| *II. Our SVM classifier* | | | | | | | |
| a. $n$-grams | 65.2 | **42.4** | 57.5 | 58.6 | **66.4** | 58.0 | 69.0 |
| b. a. + POS | **65.8** | 41.8 | **58.7** | 57.6 | 62.6 | 57.3 | 68.3 |
| c. a. + encodings | 65.7 | 42.1 | 57.6 | 58.4 | 64.5 | 57.6 | 68.6 |
| d. a. + target | 65.2 | 42.2 | 57.7 | 60.2 | 66.1 | **58.3** | **69.1** |
| e. a. + sentiment | 65.2 | 40.1 | 54.5 | **60.6** | 61.7 | 56.4 | 66.8 |

From Mohammad et al. (2017).

# Some thoughts

- If you don't know where to start, look at some examples!
- Often: less good results $\rightarrow$ more error analysis
  - This doesn't need to be the case!
- Process is often not straightforward

## Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types
   → evaluate by type

# Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types → evaluate by type
2. (6) **Think** stance & sentiment are related → **look** at examples → **observe** patterns → evaluate with swapping

# Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types → evaluate by type
2. (6) **Think** stance & sentiment are related → **look** at examples → **observe** patterns → evaluate with swapping
3. (2) **Think** not satisfied with explanation of results → **think** about difference → **look** at difference (clusters)

# Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types → evaluate by type
2. (6) **Think** stance & sentiment are related → **look** at examples → **observe** patterns → evaluate with swapping
3. (2) **Think** not satisfied with explanation of results → **think** about difference → **look** at difference (clusters)
4. (3) **Observe** big clusters of topics → evaluate by cluster size

# Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types → evaluate by type
2. (6) **Think** stance & sentiment are related → **look** at examples → **observe** patterns → evaluate with swapping
3. (2) **Think** not satisfied with explanation of results → **think** about difference → **look** at difference (clusters)
4. (3) **Observe** big clusters of topics → evaluate by cluster size
5. (4) **Think** some topics have few examples → evaluate by number of examples

# Example: my recollection of an error analysis process

Chronological order (paper order) for Allaway and McKeown (2020)

1. (5) **Look** manually at outputs→ **observe** error types → label types → evaluate by type
2. (6) **Think** stance & sentiment are related → **look** at examples → **observe** patterns → evaluate with swapping
3. (2) **Think** not satisfied with explanation of results → **think** about difference → **look** at difference (clusters)
4. (3) **Observe** big clusters of topics → evaluate by cluster size
5. (4) **Think** some topics have few examples → evaluate by number of examples
6. (1) **Reviewer**: what if topics are lexically similar → extra analysis

# Easiest: dataset based

Statistics or metric across partitions of your dataset. For example

- Labels
- Language
- Source domain
- Length

Important when imbalanced data

# Example 1: labels

| | F1 All | | | F1 Zero-Shot | | | F1 Few-Shot | | |
|---|---|---|---|---|---|---|---|---|---|
| | pro | con | all | pro | con | all | pro | con | all |
| CMaj | .382 | .441 | .274 | .389 | .469 | .286 | .375 | .413 | .263 |
| BoWV | .457 | .402 | .372 | .429 | .409 | .349 | .486 | .395 | .393 |
| C-FFNN | .410 | .434 | .300 | .408 | .463 | .417 | .413 | .405 | .282 |
| BiCond | .469 | .470 | .415 | .446 | .474 | .428 | .489 | .466 | .400 |
| Cross-Net | .486 | .471 | .455 | .462 | .434 | .434 | .508 | .505 | .474 |
| BERT-sep | .4734 | .522 | .5014 | .414 | .506 | .454 | .524 | .539 | .544 |
| BERT-joint | .545 | **.591** | .653 | .546 | .584 | .661 | .544 | **.597** | .646 |
| TGA Net | **.573*** | .590 | **.665** | **.554** | **.585** | **.666** | **.589*** | .595 | **.663** |

Table 5: Macro-averaged F1 on the test set. * indicates significance of TGA Net over BERT-joint, $p < 0.05$.

From Allaway and McKeown (2020).

# Example 2: languages

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| NYT-base | 32.94 | 10.36 | 22.51 |
| Abs-so* | 37.72 | 15.39 | 26.56 |
| Abs-mix* | **38.07** | **15.76** | **26.82** |

(a) Performance on Somali NYT.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| NYT-base | 35.28 | 12.96 | 25.64 |
| Abs-sw* | 39.24 | 17.01 | 29.88 |
| Abs-mix* | **39.96** | **17.56** | **30.24** |

(b) Performance on Swahili NYT.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| NYT-base | 37.17 | 14.67 | 27.27 |
| Abs-tl* | **40.96** | 18.72 | 31.06 |
| Abs-mix* | 40.87 | **18.91** | **31.14** |

(c) Performance on Tagalog NYT.

Table 3: Abs-so, -sw, and -tl are the Somali, Swahili, and Tagalog systems, respectively. * indicates significant improvement over NYT-base ($p < 1.16 \times 10^{-19}$).

From Ouyang et al. (2019).

# Still pretty easy: model ablations

How do various components impact the model? For example

- Features
- Base LM family
- Base LM size

# Example 3: components of the model

| | Entity | | Event | |
|---|---|---|---|---|
| | F1 | Δ | F1 | Δ |
| Our Model | 75.3 | | 80.8 | |
| − Coref feat (§3.6) | - | - | 79.6 | -1.2 |
| − Args (§3.2) | 74.8 | -0.9 | 78.7 | -2.1 |
| − Arg comp (§3.2) | 74.6 | -0.7 | 78.3 | -2.5 |
| − CLS (Eq. 1) | 74.5 | -0.8 | 78.9 | -1.9 |
| − MP cosine (§3.4) | 74.5 | -0.8 | 79.1 | -1.7 |
| + GloVE | 70.1 | -5.2 | 76.7 | -4.1 |
| + RoBERTa | 71.2 | -4.1 | 78.1 | -2.7 |

Table 4: Feature ablation results (CoNLL F1) on the ECB+ test set. For entity coreference arguments (Args) are events, for event coreference they are entities.

From Allaway et al. (2021).

# Example 4: base model

| Method | Model | Recall | Precision | Full |
|---|---|---|---|---|
| | *Answers only* | | | |
| 0-shot | 4B Instruct | 27.4 | 51.8 | 2.4 |
| | 4B Thinking | 55.5 | **71.2** | 32.6 |
| | 235B MoE Instruct | 44.0 | 72.0 | 10.8 |
| | 235B MoE Thinking | 53.1 | 55.1 | 37.2 |
| SFT | 4B Instruct | 40.6 | 50.8 | 21.9 |
| | *Interpretations and Answers* | | | |
| CoT | 4B Instruct | 20.5 | 27.2 | 9.3 |
| | 4B Thinking | 20.7 | 25.1 | 12.2 |
| | 235B MoE Instruct | 60.5 | 63.2 | 38.2 |
| | 235B MoE Thinking | 51.8 | 43.2 | 38.1 |
| SFT | 4B Instruct | 32.9 | 51.4 | 9.1 |
| IntentRL | 4B Instruct | **66.9** | 58.2 | **49.1** |

Table 7: Recall, Precision, and Full Coverage (%) on AmbiQT (SFT/IntentRL trained on Ambrosia).

From Saparina and Lapata (2025).

# A tiny bit harder: sensitivity analysis

How sensitive is the model to changes. For example

- Hyperparameters
- Training data
- Random seed
- Prompt formulation

# Example 5: number of prompt examples

| Model \ Shots | Avg. Acc. (%) | | | | SD (across seeds) | | | | SD (across prompts) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 8 | 64 | 512 | 0 | 8 | 64 | 512 | 0 | 8 | 64 | 512 |
| GPT-5 | 19.8 | 73.3 | 83.3 | 87.2 | 1.7 | 1.4 | 1.1 | 1.3 | 0.0 | 0.4 | 0.9 | 1.4 |
| GPT-5-nano | 18.9 | 61.4 | 64.3 | 65.2 | 1.1 | 1.8 | 2.1 | 2.0 | 0.4 | 0.5 | 0.9 | 1.9 |
| GPT-4o-mini | 19.8 | 40.3 | 44.3 | 47.2 | 1.7 | 2.1 | 1.8 | 2.1 | 0.0 | 0.2 | 1.7 | 4.4 |
| Llama3 8B | 18.4 | 28.5 | 32.7 | - | 1.5 | 2.0 | 1.2 | - | 9.0 | 13.5 | 14.3 | - |
| Qwen2.5 0.5B | 3.9 | 5.0 | 8.0 | 1.8 | 1.3 | 0.9 | 0.9 | 0.2 | 2.6 | 5.8 | 1.9 | 1.2 |
| Qwen2.5 1.5B | 17.4 | 19.4 | 19.4 | 16.9 | 1.0 | 0.9 | 1.3 | 1.7 | 1.0 | 0.4 | 0.4 | 2.9 |
| Qwen2.5 3B | 19.0 | 29.4 | 31.6 | 31.0 | 1.3 | 1.7 | 2.0 | 0.8 | 1.9 | 3.6 | 3.8 | 4.4 |
| Qwen2.5 7B | 19.0 | 29.1 | 34.7 | 38.7 | 1.6 | 2.9 | 1.8 | 2.9 | 0.1 | 3.0 | 2.5 | 3.0 |
| Qwen2.5 14B | 19.7 | 33.0 | 45.2 | 49.3 | 1.7 | 1.3 | 1.3 | 2.1 | 0.1 | 1.8 | 4.2 | 4.4 |
| Qwen2.5 32B | 19.7 | 42.3 | 51.2 | 57.6 | 1.7 | 2.0 | 1.3 | 2.8 | 0.0 | 3.5 | 1.1 | 1.0 |
| Qwen2.5 72B | 19.8 | 40.8 | 46.8 | 51.1 | 1.7 | 2.5 | 1.3 | 0.8 | 0.0 | 1.9 | 3.3 | 3.9 |

Table 2: Inductive performance (avg. acc.) and standard deviation (SD) across seeds or prompt variants, under 0, 8, 64, and 512 shots. Model performance is stable even when the test sets are dynamically constructed.

From O'Brien et al. (2025).

# Example 6: subcomponents

| | Entity | | Event | |
|---|---|---|---|---|
| | F1 | Δ | F1 | Δ |
| Our Model | 75.3 | | 80.8 | |
| HeSRL - C | 75.3 | -0.0 | 80.4 | -0.4 |
| HeSRL + BhR + C | 74.9 | -0.4 | 79.2 | -1.6 |
| Swirl + BhR + C | 75.4 | +0.1 | 80.0 | -0.8 |
| Swirl + BhR | 75.4 | +0.1 | 78.7 | -2.1 |

Table 5: Ablation results (CoNLL F1) on methods for identifying event structures on ECB+ test set. HeSRL is He et al. (2018), BhR is additional rules for aligning the SRL and annotations from (Barhom et al., 2019), C is entity type constraint (see §4.2).

From Allaway et al. (2021).

# Example 7: prompt

| | |
|---|---|
| **(a)** | *Please answer with only "more likely", "less likely", or "it has no impact".*<br>Consider the following information: $[\mathcal{P}^i]$<br>From this information we can draw a conclusion about $[K^{\mathcal{H}}]$.<br>Conclusion: $[\mathcal{H}]$.<br>Now suppose we are given additional information.<br>Additional information: $[\mathcal{P}^x]$<br>Given the additional information,<br>how likely are you to believe the conclusion? |
| **(b)** | *Please answer with only "strengthens", "weakens" or "it has no impact".*<br>Consider the following premises: $[\mathcal{P}^i]$<br>This entails the conclusion that $[\mathcal{H}]$.<br>Additional information: $[\mathcal{P}^x]$<br>How does the given additional information<br>impact the conclusion?? |
| **(c)** | *Please answer with only "more likely", "less likely", or "it has no impact".*<br>Let's think step by step.<br>First, consider the following information: $[\mathcal{P}^i]$<br>From this information we can draw a conclusion about $[K^{\mathcal{H}}]$.<br>Conclusion: $[\mathcal{H}]$.<br>Now suppose we are given additional information.<br>Additional information: $[\mathcal{P}^x]$<br>Given the additional information,<br>how likely are you to believe the conclusion? |

Table 8: The three prompts used in our experiments. The system instruction is in *italics* about the dashed line.

From Allaway and McKeown (2025).

# Hard: human evaluation

Humans judge the output based on some criteria

- Some tasks have standard criteria, e.g., fluency
- Recent trend in using LLMs to do this instead

# Example 8: evaluation criteria

| Somali Weblogs | | |
|---|---|---|
| Model | Content | Fluency |
| NYT-base | 1.66 | 1.62 |
| Abs-so | 1.92 | 1.90 |
| Abs-sw | 1.94 | 1.88 |
| Abs-tl | 1.86 | 1.82 |
| Abs-mix | **2.08** | **2.04** |

| Swahili Weblogs | | |
|---|---|---|
| Model | Content | Fluency |
| NYT-base | 1.88 | 1.76 |
| Abs-so | 2.14 | 1.90 |
| Abs-sw | 2.22 | **2.08** |
| Abs-tl | 2.18 | 1.86 |
| Abs-mix | **2.36** | **2.08** |

| Tagalog Weblogs | | |
|---|---|---|
| Model | Content | Fluency |
| NYT-base | 1.72 | 1.76 |
| Abs-so | 1.76 | 1.88 |
| Abs-sw | 1.94 | 1.92 |
| Abs-tl | 1.80 | 2.08 |
| Abs-mix | **2.08** | **2.16** |

Table 5: Average human-rated content and fluency scores on Somali, Swahili, and Tagalog weblog entries.

From Ouyang et al. (2019).

# Scary*: manual data analysis

- Look through the inputs $\rightarrow$ patterns that are causing errors
- Look through model outputs $\rightarrow$ patterns in behavior

* This is a bit sarcastic, but manual data analysis is definitely hard.

# Example 9: manual analysis

**Error Analysis** To better understand model failures, we further analyze the 30 examples sampled for evaluating interpretation alignment (Section 6). On Abg-CoQA, we identify three types of error. First, the model ignores important context from previous dialogue turns. Second, it sometimes produces valid, well-specified interpretations but then gives factually incorrect answers, such as referencing a different person than the one implied in the interpretation. Third, it predicts generic interpretations like "The question is ambiguous." In some cases, this is paired with a specific answer, without clarifying which interpretation the answer corresponds to. In other cases, it is paired with the answer "Unknown" which appears in some training examples, but the model has learned to produce it more often than necessary.

We also observe on Ambrosia that interpretations can be vague, sometimes paraphrasing the question without resolving the ambiguity. Moreover, the model occasionally produces correct interpretations but predicts queries that do not follow them (e.g., adding ID fields not mentioned in the question) and generates non-executable queries, even noting the error in SQL comments (e.g., "this column does not exist"). It also repeats the same interpretation multiple times, although repetitions can be filtered via execution results.

From Saparina and Lapata (2025).

# Example 10: visualising differences

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.
I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(a) Ouyang et al. gold standard annotation.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.
I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(b) Pointer-aligner alignment.

I saved my friend's life from a heroin overdose, and she repaid me by hooking up with my boyfriend.
I saved my (at the time) best friend from a heroin overdose by sticking suboxone under her tongue, which I miraculously had with me at the time. About a month later her and my boyfriend who I had been living with for two years hooked up.

(c) Jacana alignment.

Figure 7: Ouyang et al. alignments. Due to length restrictions, we show only the best-performing baseline, Jacana.

From Ouyang and McKeown (2019).

# Overview

Types of Error Analysis

Exercise

# Discuss

- What kinds of analyses would you do?
- How would you do these?

# References i


Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2025. Evaluating defeasible reasoning in LLMs with DEFREASING. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10540–10558, Albuquerque, New Mexico. Association for Computational Linguistics.

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Dayyán O'Brien, Barry Haddow, Emily Allaway, and Pinzhen Chen. 2025. Mathemagic: Generating dynamic mathematics benchmarks robust to memorization. *arXiv preprint arXiv:2510.05962*.

# References ii

Jessica Ouyang and Kathy McKeown. 2019. Neural network alignment for sentential paraphrases. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4724–4735, Florence, Italy. Association for Computational Linguistics.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Irina Saparina and Mirella Lapata. 2025. Reasoning about intent for ambiguous requests. *arXiv preprint arXiv:2511.10453*.