Error Analysis Report: All Splits

Dataset

- Total examples analyzed: 856 (529 train, 66 val, 67 test, 28 gold)
- Models evaluated: 18 (8 baselines, 4 fine-tuned, 6 zero-shot LLMs)
- Note: Train split was only evaluated by 6 LLM models. Val, test, and gold were evaluated by all 18 models.

Hardest Examples

Examples where the most models made errors:

Models Wrong	Split	True Label	Text
16/18	test	LITERAL	"And then and then you clicked it again, she's dressed. She is a business woman, she is walking down the street and oh oh oh she's naked."
15/18	test	LITERAL	"Oh, just put it in your mouth and pop it like a zit."
14/18	test	SARCASTIC	"Uh, Rachel's here, so good luck man, let me know how it works out."
14/18	test	LITERAL	"Lois Lane is falling, accelerating at an initial rate of 32 feet per second per second. Superman swoops down to save her by reaching out two arms of steel. Miss Lane, who is now traveling at approximately 120 miles an hour, hits them and is immediately sliced into three equal pieces."
13/18	test	LITERAL	"Provided he has already read and is familiar with the reestablishment of the DC multiverse."
13/18	gold	LITERAL	"RAJ: Leonard, may I present, live from New Delhi, Dr. and Mrs. V. M. Koothrappali."
13/18	gold	LITERAL	"LEONARD: Don't you think looking for a new city to live in is a bit of an overreaction?"

Easiest Examples

Only 1 example was classified correctly by all 18 models:

Split	True Label	Text
gold	LITERAL	"RACHEL: All right, let's do it."

Error Types in Hardest Examples

Among the 7 hardest examples listed above:

- 5 were False Positives (literal utterances predicted as sarcastic)
- 2 were False Negatives (sarcastic utterances predicted as literal)

Model Performance on Hardest Examples

Hardest example (16/18 wrong):

- Models correct: xlmr_finetuned_context, xlmr_finetuned_no_context
- All 8 baselines: wrong
- All 6 LLMs: wrong
- DeBERTa (both conditions): wrong

Second hardest example (15/18 wrong):

- Models correct: xlmr_finetuned_context, xlmr_finetuned_no_context, deberta_finetuned_no_context
- All 8 baselines: wrong
- All 6 LLMs: wrong

Examples where only LLMs were correct:

For gold example 0 ("Yeah, sure. Why not you?" — SARCASTIC, 12/18 wrong):

- Models correct: gemma2_9b_context, gemma2_9b_no_context, qwen2.5_14b_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context
- All 8 baselines: wrong
- All 4 fine-tuned models: wrong

For test example 36 ("And I love that I work and do all the cleaning, and you're okay with that." — SARCASTIC, 12/18 wrong):

- Models correct: gemma2_9b_context, gemma2_9b_no_context, qwen2.5_14b_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context
- All 8 baselines: wrong

- All 4 fine-tuned models: wrong

Baseline Performance

All 8 TF-IDF baseline models appear in the "wrong" list for every example with 13 or more models wrong.

Context Effect

For gold example 9 ("Don't you think looking for a new city to live in is a bit of an overreaction?" — LITERAL, 13/18 wrong):

- Models correct: xlmr_finetuned_no_context, deberta_finetuned_no_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context
- Models wrong (with context): xlmr_finetuned_context, deberta_finetuned_context, qwen2.5_14b_context

Three models got this wrong with context but correct without context.