Perelyn

LIGHTNING TALK

An advanced validation pipeline for summary evaluation in high-stakes environments

Dr. Liliya Imasheva · Dr. Michael Banf

Perelyn GmbH

Evaluating generated summaries · LLM as a Judge

01 · THE CHALLENGE

Evaluating summary quality in high-stakes environments

Goal

Detect whether summary quality stays stable — or quietly degrades — when we change any of the inputs: a new model, a new dataset, or scaling the task up.

Why it matters

In a high-stakes environment a single flawed summary carries real consequences. We need consistency we can actually measure.

Key questions

Is performance improving with each change? Are summaries getting closer to human-level quality?

02 · WHAT WE MEASURE

Five custom evaluation dimensions

The criteria from the expert evaluation we were given for this task — fully customised, not off-the-shelf metrics.

Completeness	Does the summary include all the critical information?
Correctness	Does it accurately reflect the source, without introducing errors?
Conciseness	Does it deliver the essentials without unnecessary detail?
Harmlessness	Does it avoid anything that could mislead or cause harm?
Readability	Is it clear, well-structured and easy to read?

Experts rated every summary on each dimension

1	2	3	4	5

On a scale from 1 to 5 — the very same criteria we later hand to the LLM judge.

Two ways to evaluate a summary

Traditional metrics

Reference-based scoring

ROUGE | BLEU | BERTScore | Cosine similarity

Compare the summary to a reference text by word overlap and semantic similarity.

Output → one continuous score

LLM as a Judge

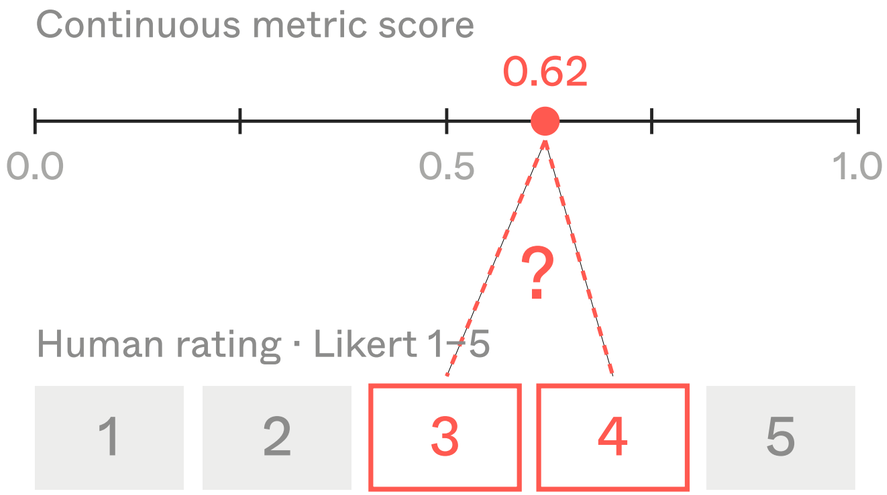
Criteria-based scoring

An LLM reads the summary and scores it on our five custom dimensions — and explains the reasoning behind each score.

Output → a 1–5 score per dimension, with reasoning

The problem with traditional metrics

- They simply performed poorly on our task.
- Metric scores are continuous; human ratings are categorical (Likert 1–5) — the two don't share a scale.
- They reward surface overlap with a reference — not whether a summary is correct or safe.



Where does 0.62 land — a 3 or a 4?

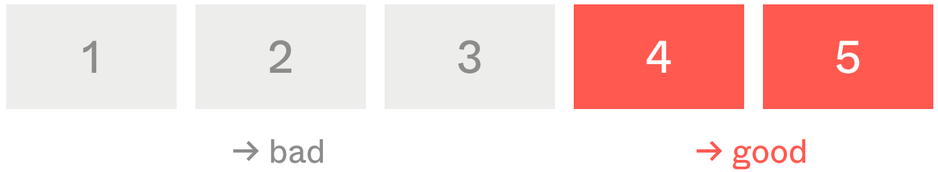
The two scales don't correspond — comparing them directly says little.

05 · THE METHOD

Measuring alignment with human feedback

We don't expect an exact match. The pipeline only has to separate good summaries from bad ones.

So we binarise both sides:



Then we measure agreement — how often the pipeline's verdict matches the expert's.

WORKED EXAMPLE

Summary	Expert	Pipeline	Match
A	5 → good	4 → good	✓
B	2 → bad	1 → bad	✓
C	4 → good	5 → good	✓
D	3 → bad	2 → bad	✓
E	5 → good	3 → bad	✗

Agreement = 4 / 5 = 80%

Inside the LLM-as-Judge pipeline

INPUTS

- LLM model
- Hyperparameters
- Prompt + dimension definitions
- Few-shot examples · positive & negative
- Negative controls
- Validation dataset

LLM as a Judge

scores each summary with reasoning

Quality verdict

Five 1–5 scores per summary

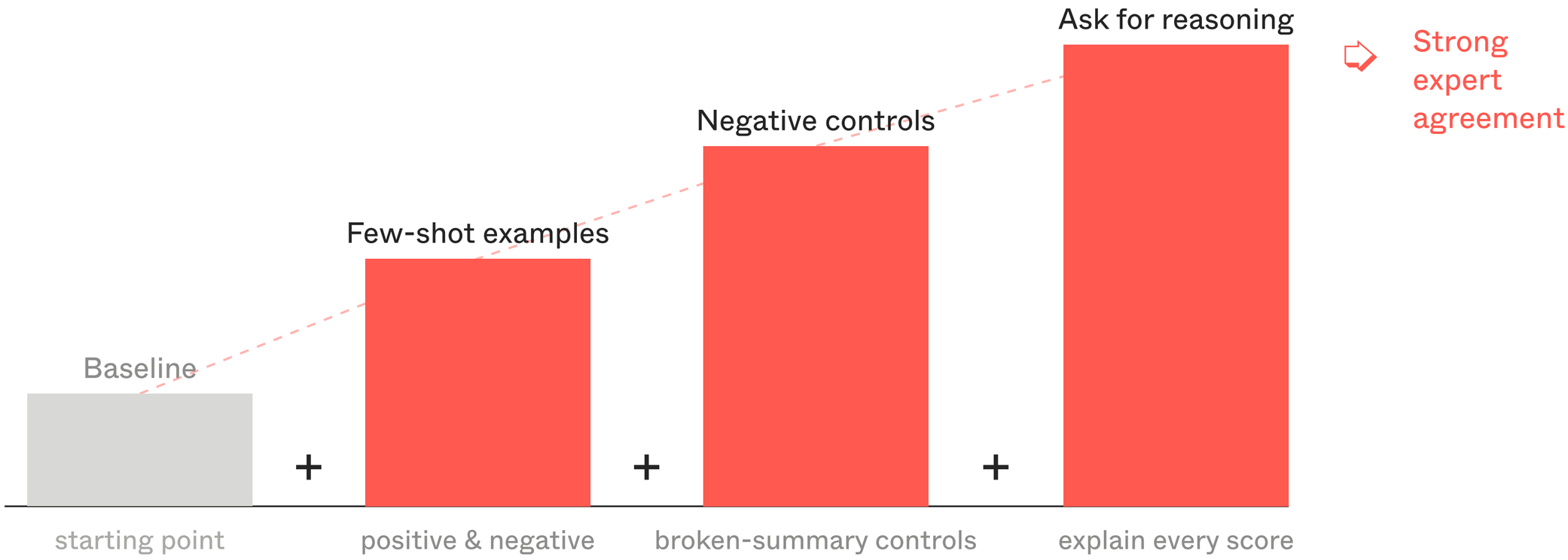
↕ compared with

Human feedback

= agreement score

Improving the validation pipeline

Three changes did most of the work — stacked together, they lifted alignment from a weak baseline to strong agreement with experts.



We also tried prompt engineering, parameter tuning and model upgrades — marginal gains by comparison.

The evaluation dataset

Synthesised summaries

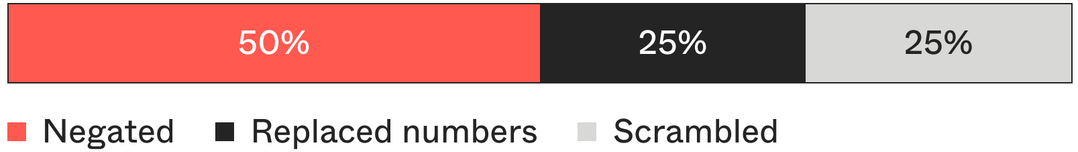
Evaluated by experts — this is our ground truth for what good and bad actually mean.

Expert-rated · the reference standard

Negative controls

≈ a third of the test set

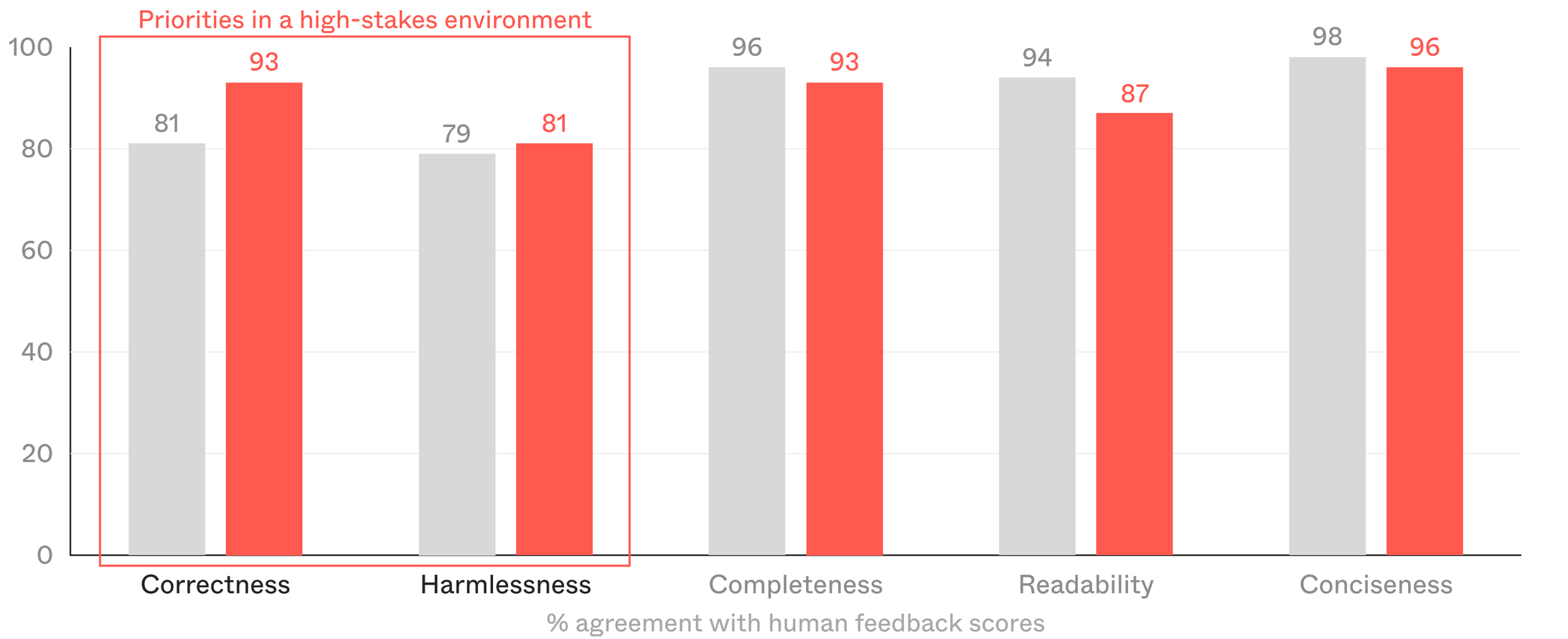
Deliberately broken summaries a good judge must catch. Three kinds:



09 · RESULTS

Results: alignment with human feedback

We focused on the dimensions that matter most when the stakes are high — correctness and harmlessness — where alignment rose the most.



10 · THE ANCHOR

Why expert comparison matters

- The pipeline is only trustworthy because it is validated against expert judgement.
- The expert set can shrink over time — but never to zero. You still need a clear spec of what "good" means, and basic sanity checks always running.
- A reliable baseline like this lets you improve the real task — summarisation — in an automated, repeatable way.

Same five dimensions. Same scoring. Human and judge — measured exactly the same way.

That shared definition is what makes "agreement" mean something.

11 · IN SUMMARY

Key takeaways

01 LLM as a Judge can be a genuinely useful evaluation tool — if you set it up properly and understand its limitations.

02 Customise your metrics to the task — generic overlap scores won't carry a high-stakes evaluation.

03 Score humans and the judge on exactly the same dimensions, so agreement is meaningful.

04 Validate against experts and keep sanity checks — then you have an automated baseline that makes improving the real task reliable.

Dr. Liliya Imasheva · Dr. Michael Banf · Perelyn GmbH