

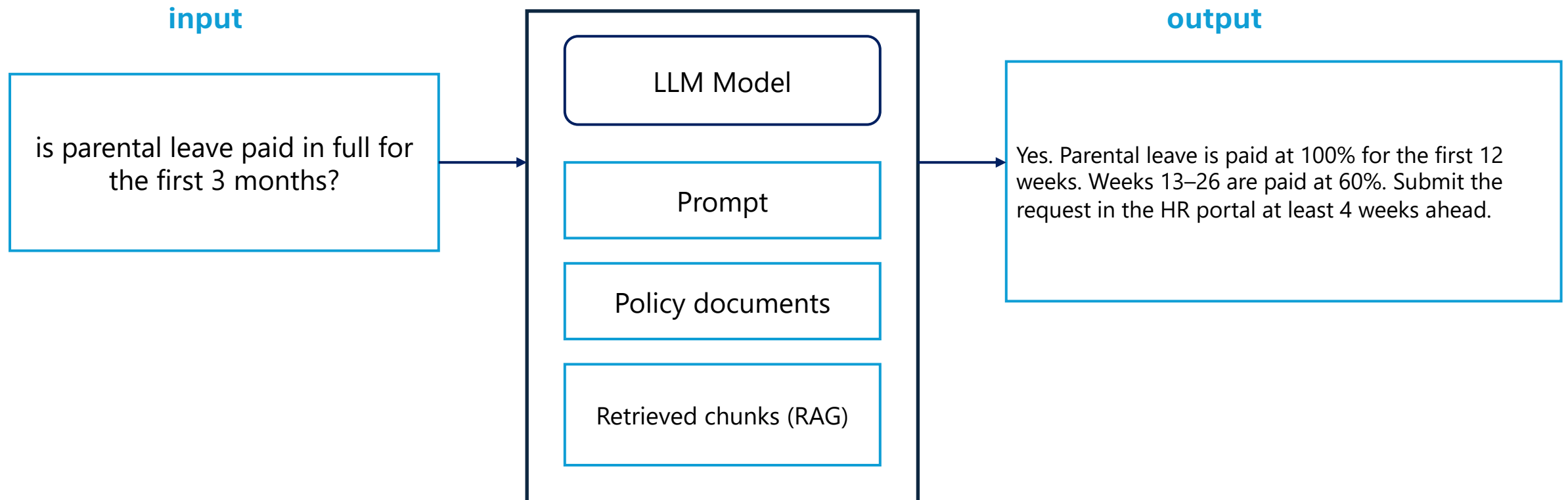
Metrics and measurement – Part II

Assessing GenAI outputs

E4rag: GenAI assistant for HR policy questions

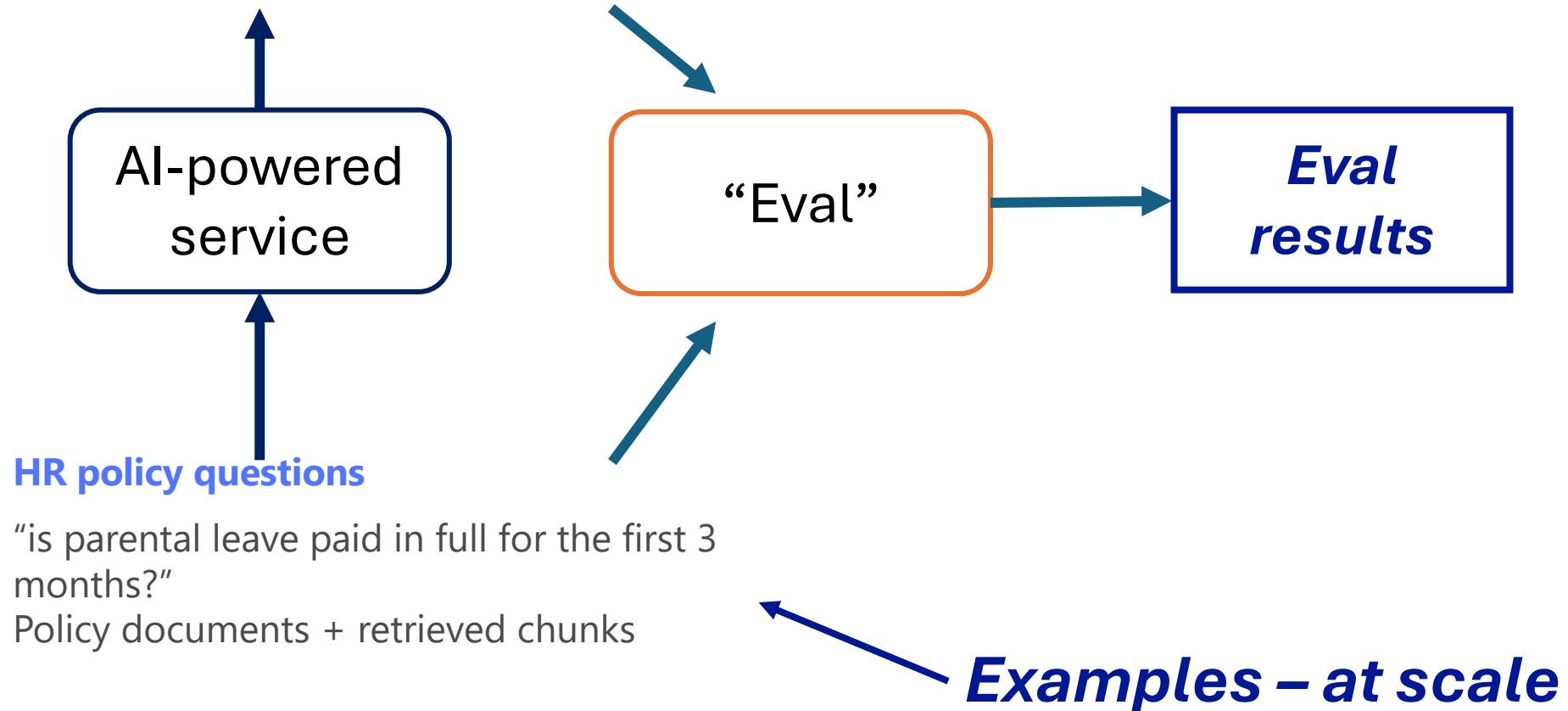
A *generation* task: answer the employee, grounded in the policy documents.

No enumerable set of correct outputs.



Answer

Yes. Parental leave is paid at 100% for the first 12 weeks. Weeks 13–26 are paid at 60%. Submit the request in the HR portal at least 4 weeks ahead.



How do we get "reliable" evals at scale?

1. What quality means / dimensions of quality
2. Who judges quality, and how
3. Whether it is subjective

The definition of “good” is nuanced

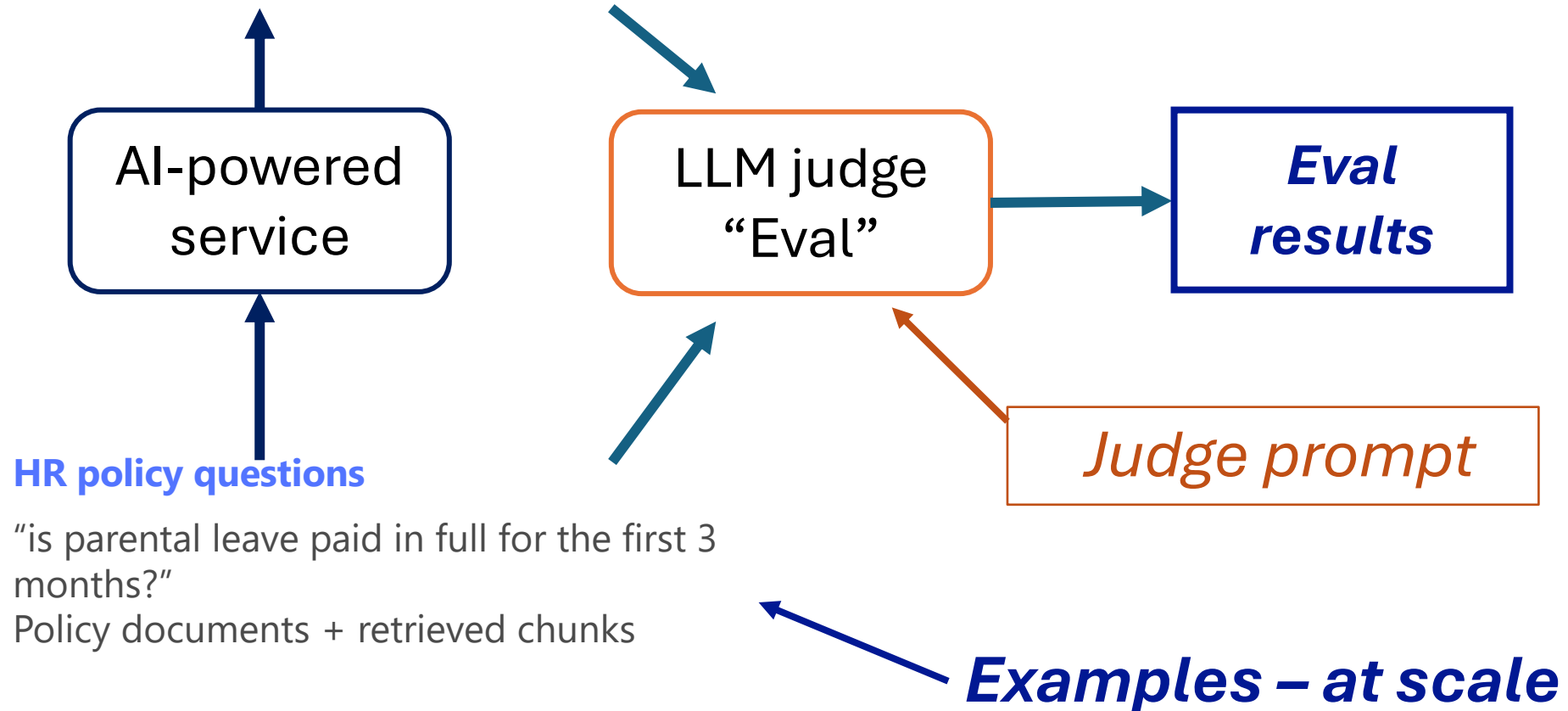
Answer to: “is parental leave paid in full for the first 3 months?”	HR expert says
A. Yes, 100% for the first 12 weeks, then 60% for weeks 13–26. Request it in the HR portal 4 weeks ahead.	Correct and complete
B. Yes, the first 12 weeks are paid in full.	Correct, but misses the 60% phase
C. Yes, parental leave is paid in full for 6 months.	Wrong – 6 months is invented
D. Please check with your HR business partner.	Not "wrong", but not useful
E. Thank you for reaching out about parental leave, which is understandably an important question. Subject to the usual eligibility rules, the first 12 weeks are paid at 100% of base salary, and weeks 13–26 are then paid at 60%, with requests submitted through the HR portal, ideally four weeks in advance ...	Correct and complete, but overly verbose

Example Metrics for LLM summarization or auto-resolution tasks

Metric	What it means
Faithfulness / correctness	Nothing in the output that is not represented in the source
Completeness	All the truth: every key point in the source is there
Conciseness	Nothing but the truth, and briefly
Consistency, latency, cost	Same input, same answer. Plus the operational metrics

Answer

Yes. Parental leave is paid at 100% for the first 12 weeks. Weeks 13–26 are paid at 60%. Submit the request in the HR portal at least 4 weeks ahead.



How: A basic judge prompt

The same rubric, written out as instructions

You are grading an ANSWER to an HR POLICY QUESTION. Judge only against the rubric below; ignore tone, length and writing style.

QUESTION: {{question}} POLICY EXCERPTS: {{chunks}}

ANSWER: {{answer}}

- 1 FAITHFULNESS / CORRECTNESS: pass/fail. Fail if the answer states anything not substantiated by the policy excerpts.
- 2 COMPLETENESS: 1 to 5. 5 = answers the question with every relevant condition. 3 = answers part. 1 = a small part only.
- 3 CONCISENESS: 1 to 5. 5 = nothing can be cut without losing a required element. 1 = half the text is restatement.

Quote the policy span that justifies each score, then JSON only:

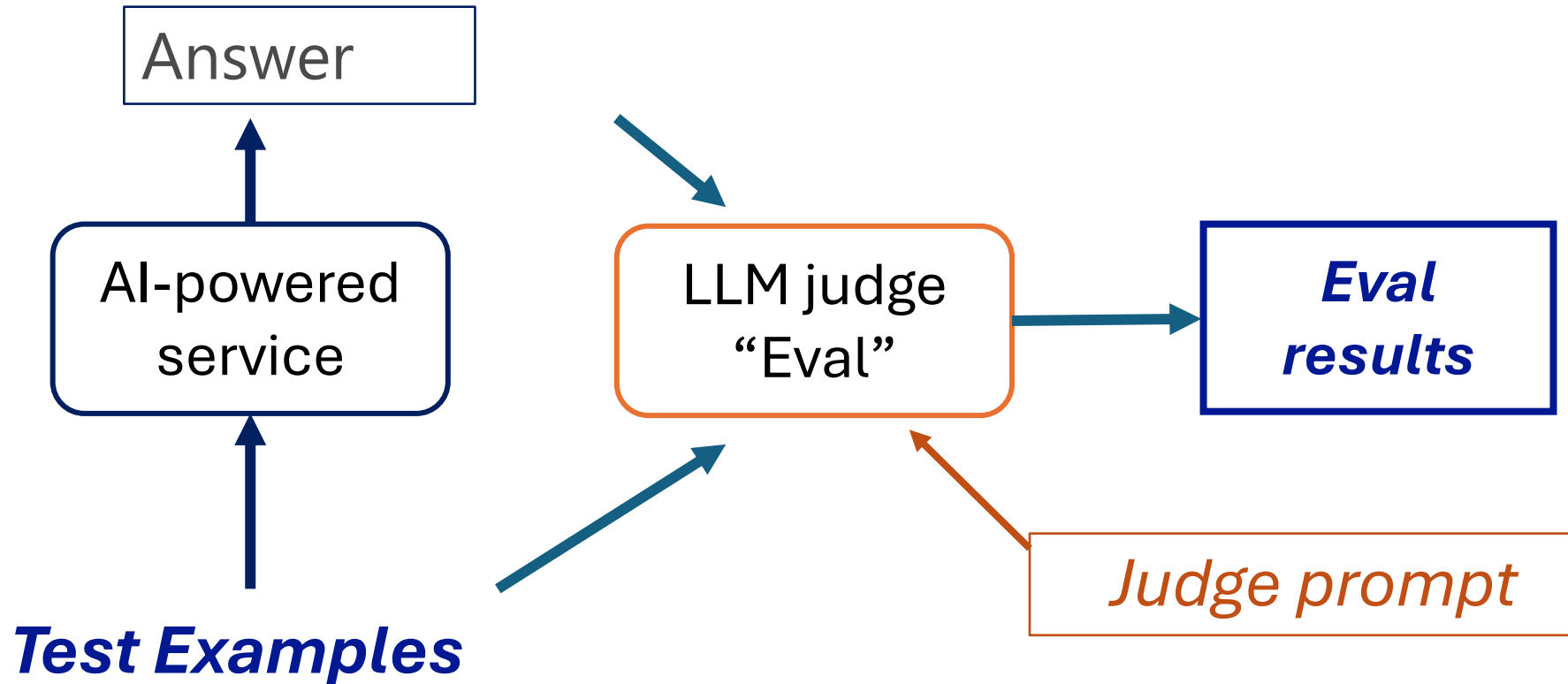
```
{"faithfulness":"pass|fail","completeness":n,"conciseness":n}
```

Precise instructions and examples often help to achieve a consistent output aligned with human judgments

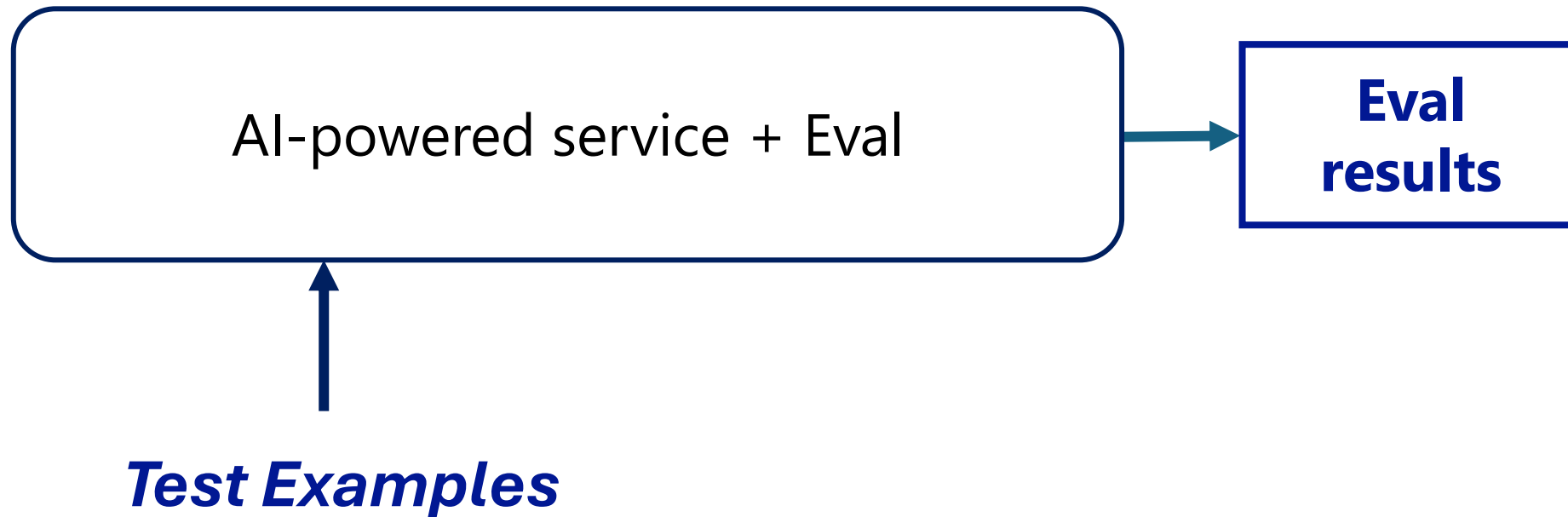
Where to get inspiration for good judge prompts

- **DeepEval: judge techniques and metric prompts**
deepeval.com/blog/llm-as-a-judge
- **Future AGI: judge prompt engineering, versioning, kappa sweeps**
futureagi.com/blog/llm-judge-prompt-engineering-guide-2026
- **Openlayer: bias mitigation and human-baseline validation**
openlayer.com/blog/llm-as-judge-evaluation-guide
- **MT-Bench: the original judge-agreement study**
arxiv.org/abs/2306.05685
- **Replacing Judges with Juries: panels beat single judges**
arxiv.org/abs/2404.18796

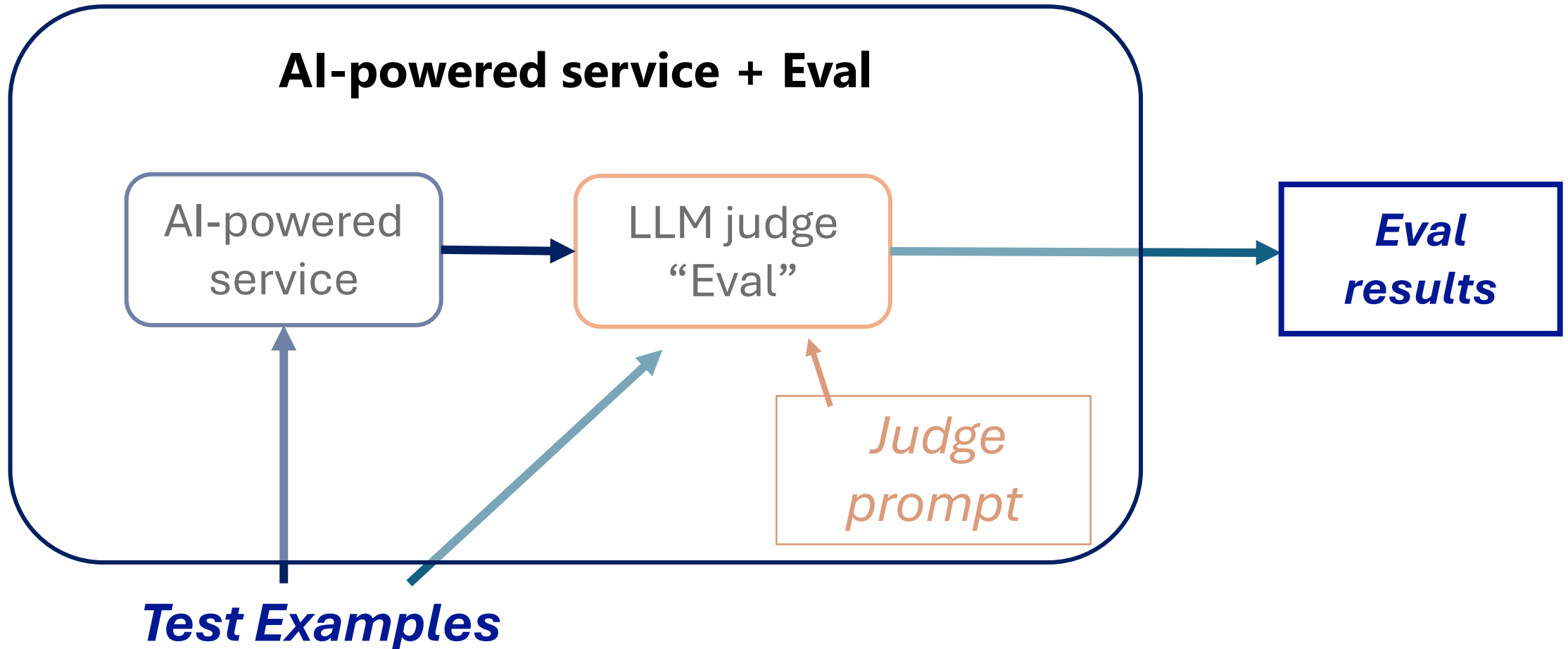
The judge is part of the system



The judge is part of the system



The judge is part of the system



The judge is part of the system

- New judge prompt, new scores. New judge model, new scores
- Lower scores may be good
- Needs the same versioning, scrutiny, discipline, RACI
- But, also, new human (or new day), new scores – especially if eval specs are fuzzy

Why an LLM judge is worth the trouble

1. It forces the rubric to be explicit

2. The judge is inspectable:

- You can run it at scale
- the prompt is a text file and every score comes with its reasoning, so you can read why it graded that way

3. You can iterate on the judge intentionally

4. Humans stay in the loop to any arbitrary degree

Still:

A. we need to “calibrate” the judgment to align it with what we want and

B. we need to estimate subjectivity / disagreement

Comparing systems. Seems easy, but it's not

Share of answers by rating

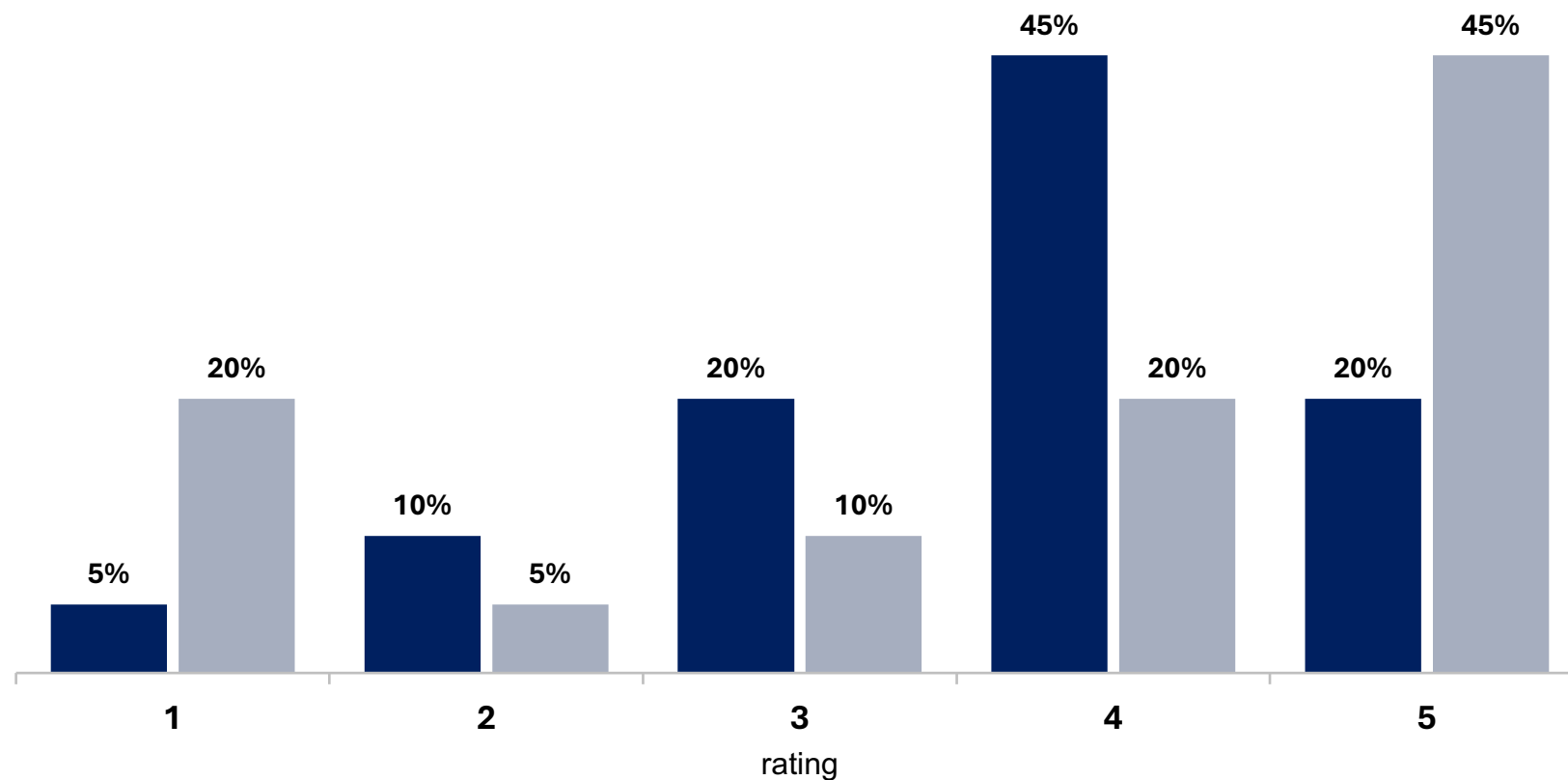
■ Model A ■ Model B

Model A

mean rating: **3.65**

Model B

mean rating: **3.65**



Value functions, again

- You may want to map scores to a single number. Sometimes this is the mean or the median
- Other times, people map it to 0-1 or 0-100

IF YOU DO THAT - NEVER REPORT IT WITH A % SIGN – IT'S A SCORE, NOT A PERCENTAGE – its neither a proportion nor a probability

Two ways to map a 1–5 faithfulness score f to 0–100

Linear: $\text{score} = (f-1)/4 \times 100$

1 -> 0 2 -> 25 3 -> 50 4 -> 75 5 -> 100

Penalizing low scores: $\text{score} = ((f-1)/4)^2 \times 100$

1 -> 0 2 -> 6 3 -> 25 4 -> 56 5 -> 100

Judge Auditor: who is judging?

The same ten answers from an HR-policy assistant, graded by five judge prompts × three models, five times each. 750 real judgments, precomputed. The HR expert's verdict is fixed in advance; the question is how far each judge lands from it, and whether it lands in the same place twice.

UNCERTAINTY DIMENSION: JUDGE BIAS, JUDGE NOISE, AND THE RUBRIC ITSELF. COMPANION TO *METRICS AND MEASUREMENT, PART II*.

The setup

The system under test is **E4rag**: an assistant that answers employee questions from a (fictional) HR handbook. Its outputs are free text, so there is no label to match. Instead each answer gets a three-row scorecard: **faithfulness** (pass/fail: nothing in the answer that the policy does not support), **completeness** (1–5) and **conciseness** (1–5).

Ten answers were generated and then curated so that each one exercises a specific way a judge can go wrong: a fabricated number, a correct-but-incomplete answer, a correct refusal, a verbose correct answer, a true answer to the wrong question. An HR expert graded each one before any LLM judge was run: that is the **oracle**.

The policy excerpt every judge (with excerpts) sees

https://reliableai.github.io/playground/judge_audit/index.html

Thanks :)

[LinkedIn](#) · [X](#) · [YouTube](#)

