The Ruler and the Judge

Benchmark-Conditional Evaluation of LLM-as-a-Judge

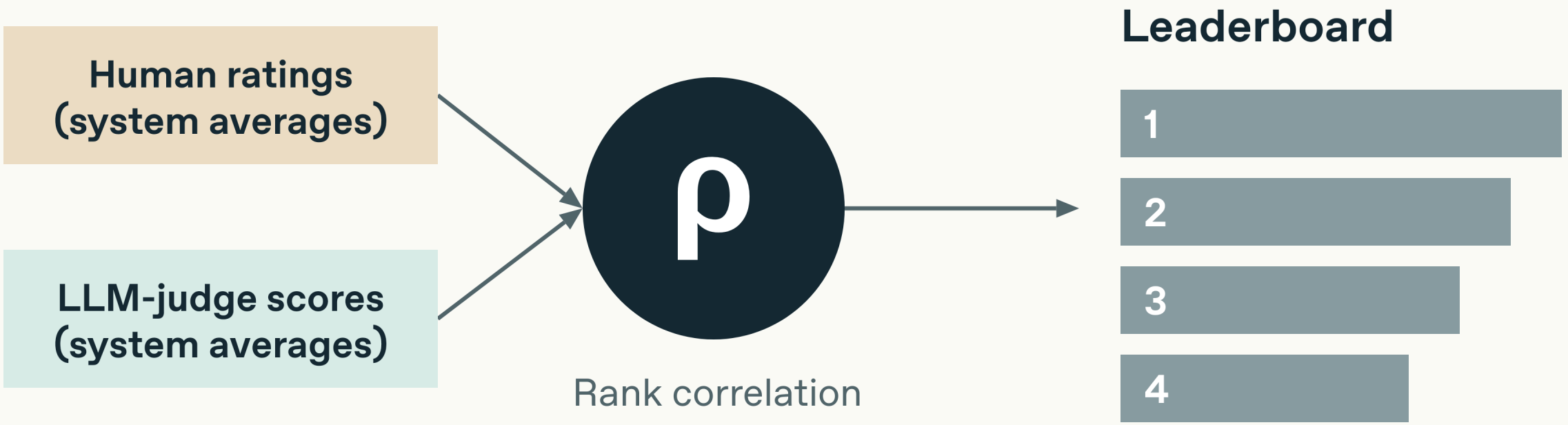
What can the human benchmark support?
Which judges are suitable for the intended use?

Adrián Ghajari · Alejandro Benito-Santos · Víctor Fresno

NLP&IR Group, UNED · NeurIPS 2026 Evaluations & Datasets Track

We trust LLM judges on one number

A language model grades other models' outputs

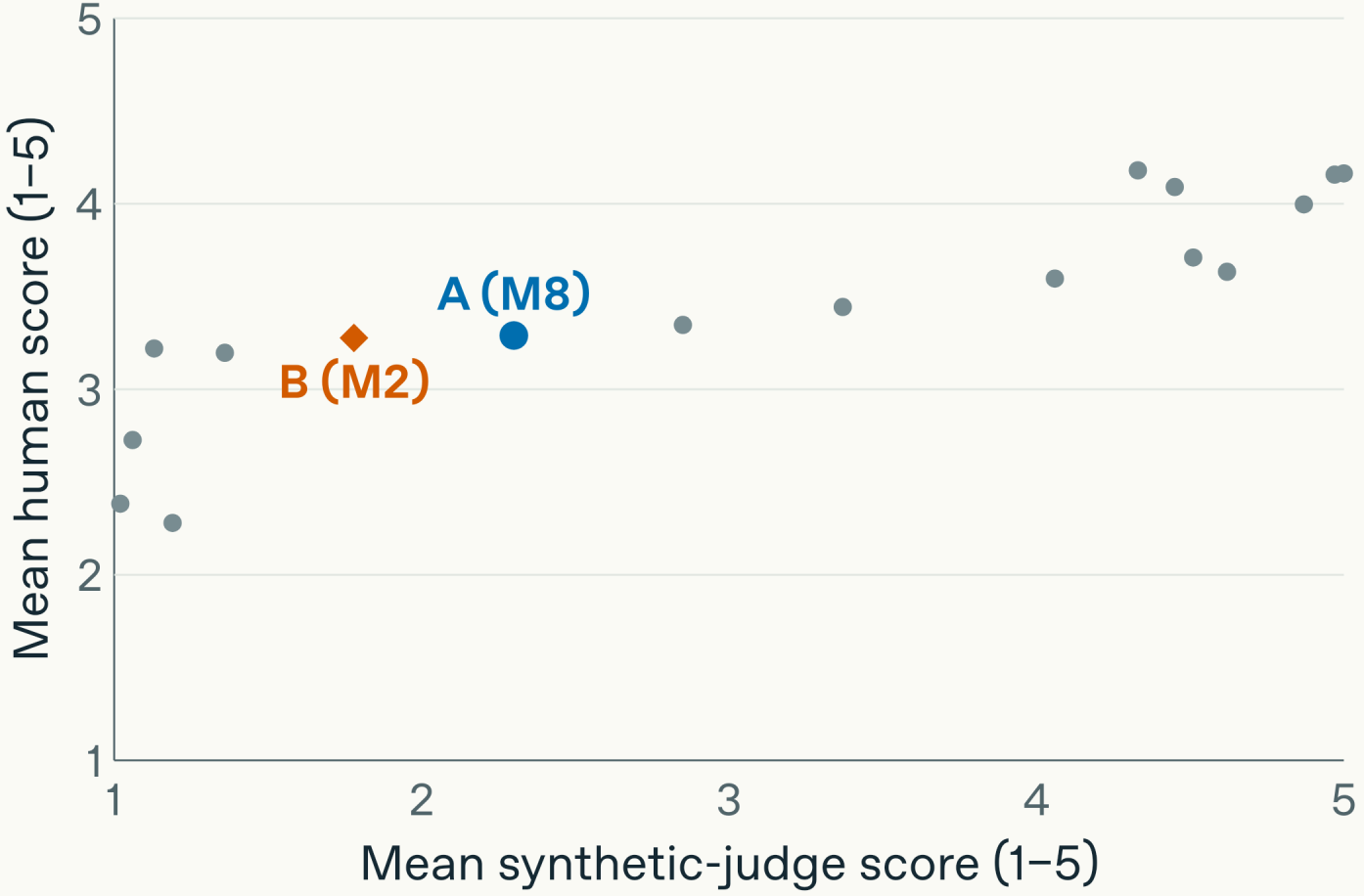


High correlation is taken as a reason to trust the judge.

Schematic validation workflow. The judge's scores, not the correlation itself, determine the ranking.

Raw scores show strong rank agreement

Each dot is one of the 17 systems. A and B are highlighted.



Raw-score Spearman

0.912

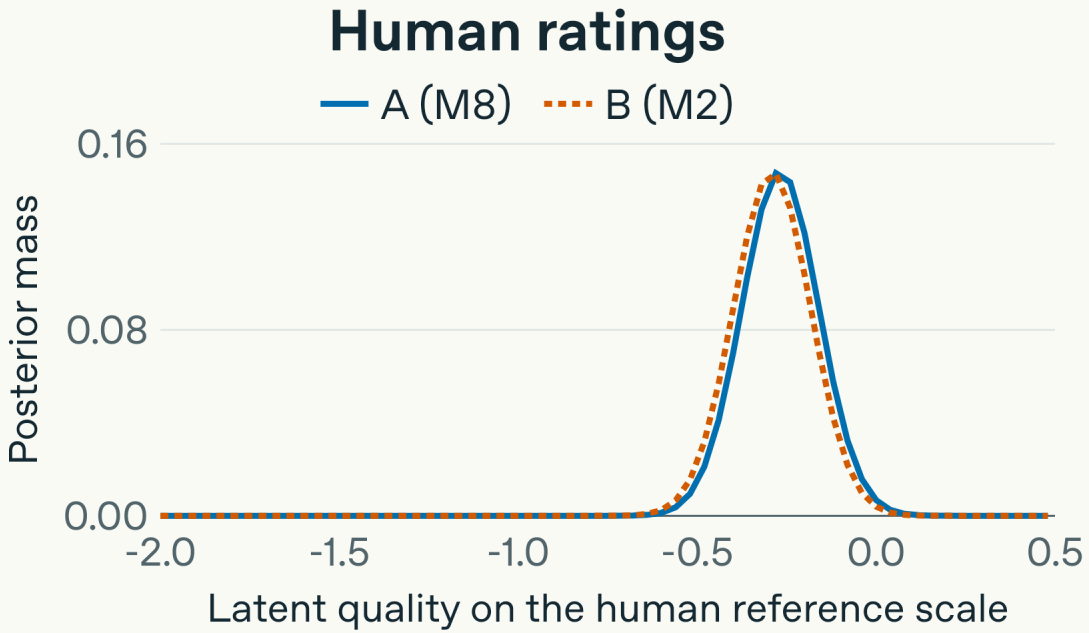
Human means
A: 3.290 B: 3.277

Does this justify ranking A above B?

Synthetic false-confidence control (Table 1). A controlled stress test, not a prevalence estimate.

Human ratings leave this comparison unsupported

SummEval coherence, systems A (M8) and B (M2)



56.0% support for A > B

Below the declared 90% threshold

Item Response Theory (IRT)

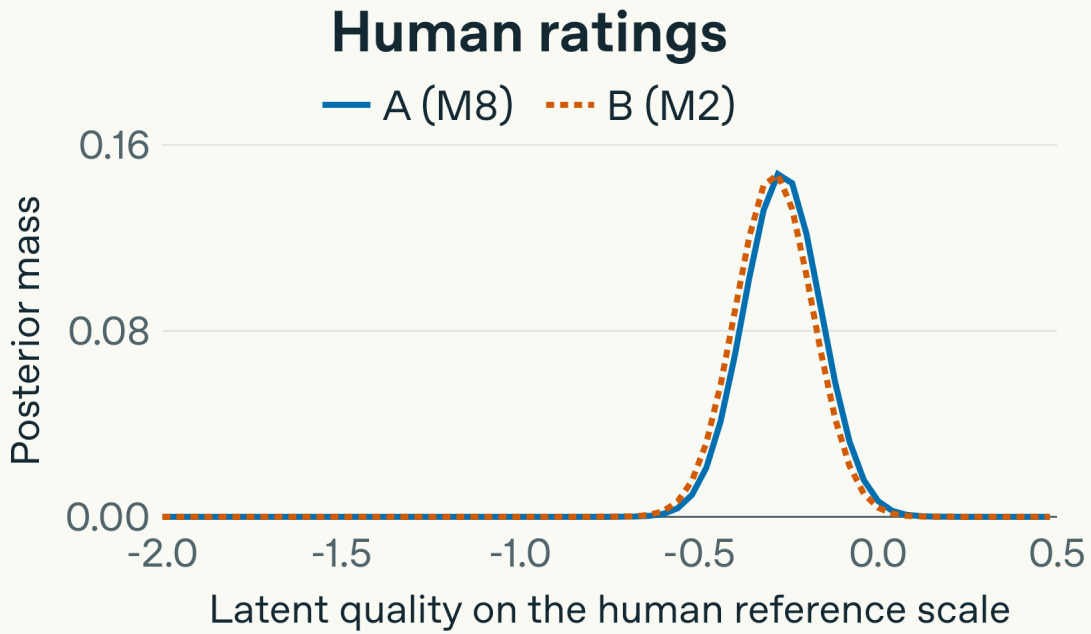
Models rating probabilities from system quality, item difficulty and rater severity.

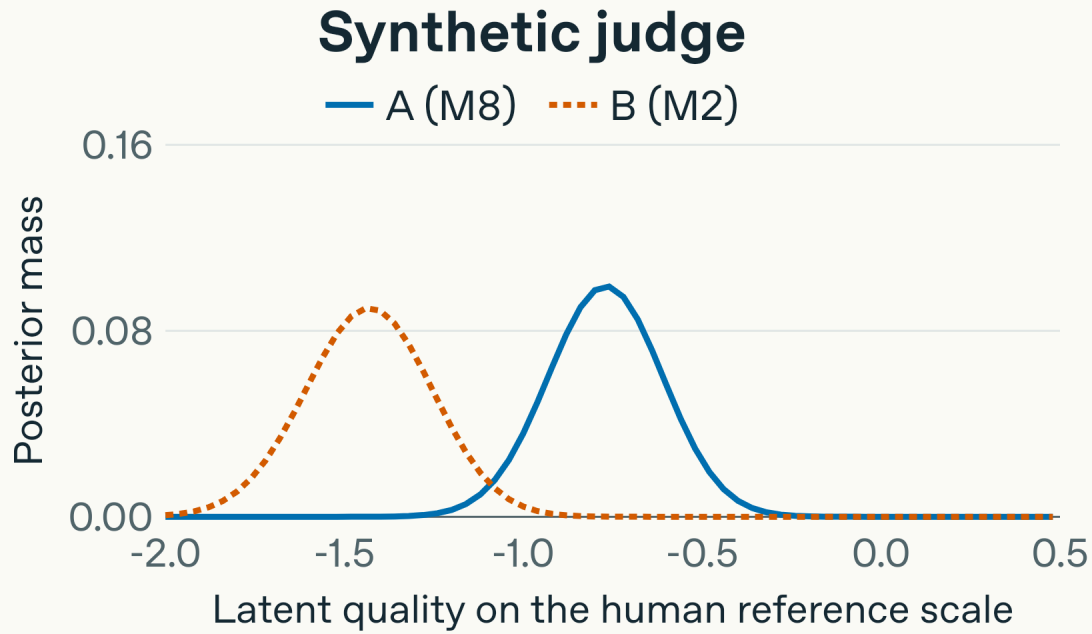
The curves show uncertainty about that latent quality.

Held-out quality posteriors, not raw score histograms. Shared axes; lines join saved grid masses.

The synthetic judge asserts a human-unsupported ordering

SummEval coherence, systems A (M8) and B (M2)





56.0% support for A > B

Below the declared 90% threshold

99.7% support for A > B

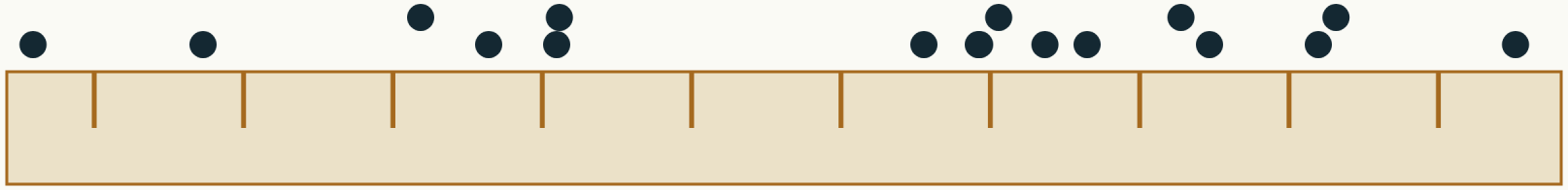
Supported under the judge model

Held-out quality posteriors, not raw score histograms. Shared axes; lines join saved grid masses.

The human benchmark is a ruler

Tick spacing shows approximate pairwise resolution.

SummEval coherence



87.5%

of pairs have supported orderings

Can the benchmark resolve the differences we want to claim?

Table 2. Each ruler spans its benchmark's central 80% of system quality. Dot positions are illustrative. Ticks: Δ_{95} / Spread_{80}. Supported pairs: joint human posterior at $t_H = 0.8$, calculated separately.

Some rulers are too coarse to tell

Tick spacing shows approximate pairwise resolution.

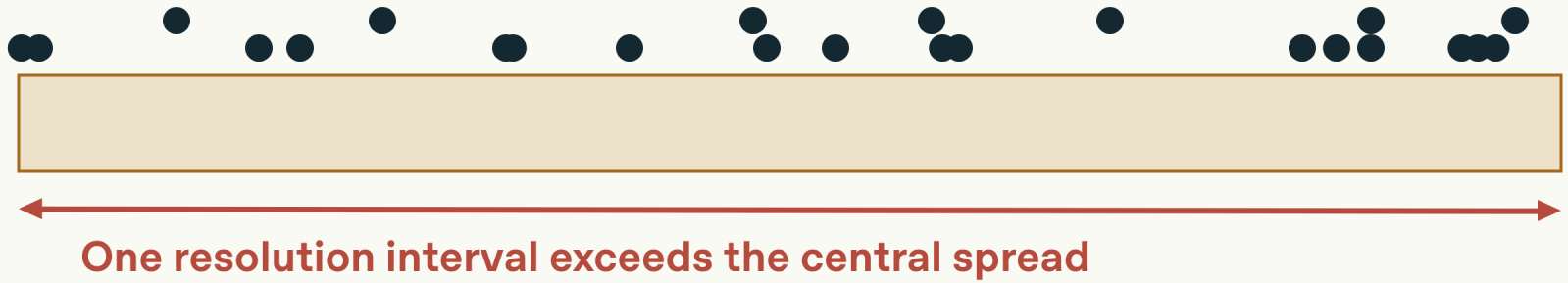
SummEval coherence



87.5%

of pairs have supported orderings

HiTZ-BASSE fluency



48.6%

of pairs have supported orderings

Table 2. Each ruler spans its benchmark's central 80% of system quality. Dot positions are illustrative. Ticks: Δ_{95} / Spread_{80}. Supported pairs: joint human posterior at $_{H}$ = 0.8, calculated separately.

Stage 1: audit the human ruler

Bayesian many-facet, Rasch-style model

Human ratings



Fit

Do ratings fit the model?

Infit / outfit MNSQ

Precision

How fine is the ruler?

Information / CSEM
Δ_{95} relative to system spread

Support

Which pairs can it order?

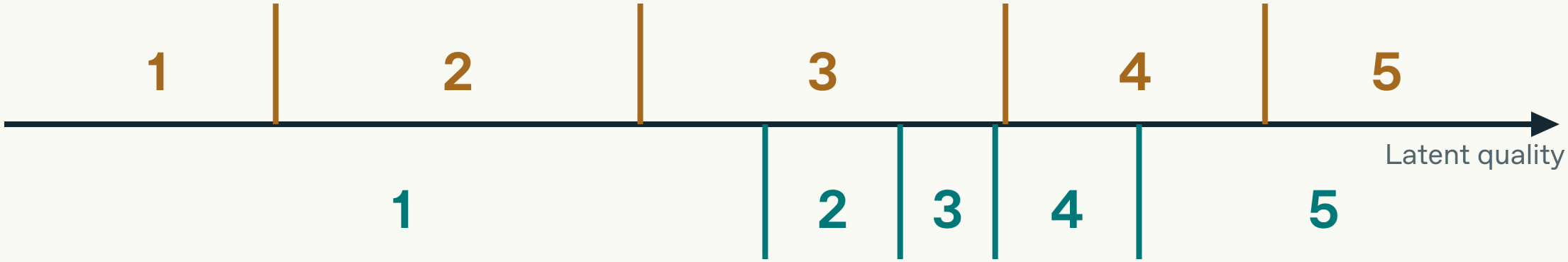
Supported resolution

Cutpoints govern rating probabilities, not fixed bins. Shown for a fixed item and rater.

Stage 2: put the judge on the same ruler

Anchor to the human model; learn the judge's category cutpoints.

Human ratings



Synthetic judge

Do the ratings fit the calibrated model?

Infit 0.69 **Outfit 0.68**

Declared interval: 0.5–1.5

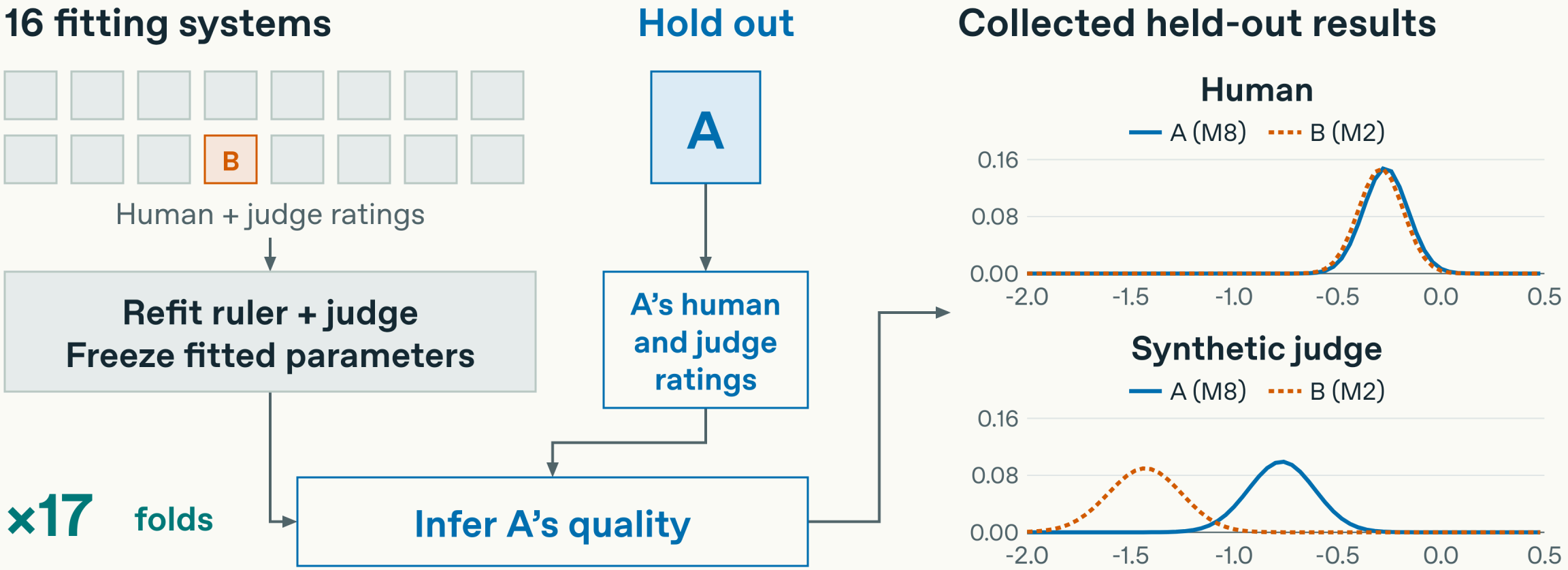
PASS

Both checks within range

Cutpoints shown for a fixed item/rater. Fit statistics from the full-data calibration.

Stage 3: test calibration on a held-out system

Does the calibration transfer beyond the systems used to fit it?



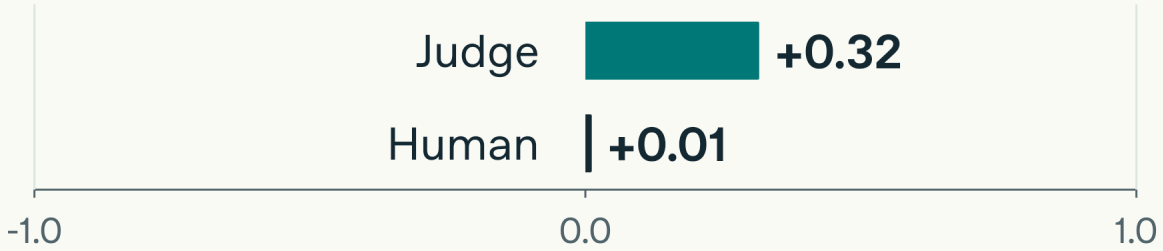
Separate human and judge distributions for each system

Held-out posteriors. Fold fitting excludes the held-out system; alignment uses the full-data human reference.

Stage 3: two readings of the same comparisons

Preserve the human comparison and respect its uncertainty.

Margin: direction and strength



Positive values favour A over B

Support: a decisive ordering?

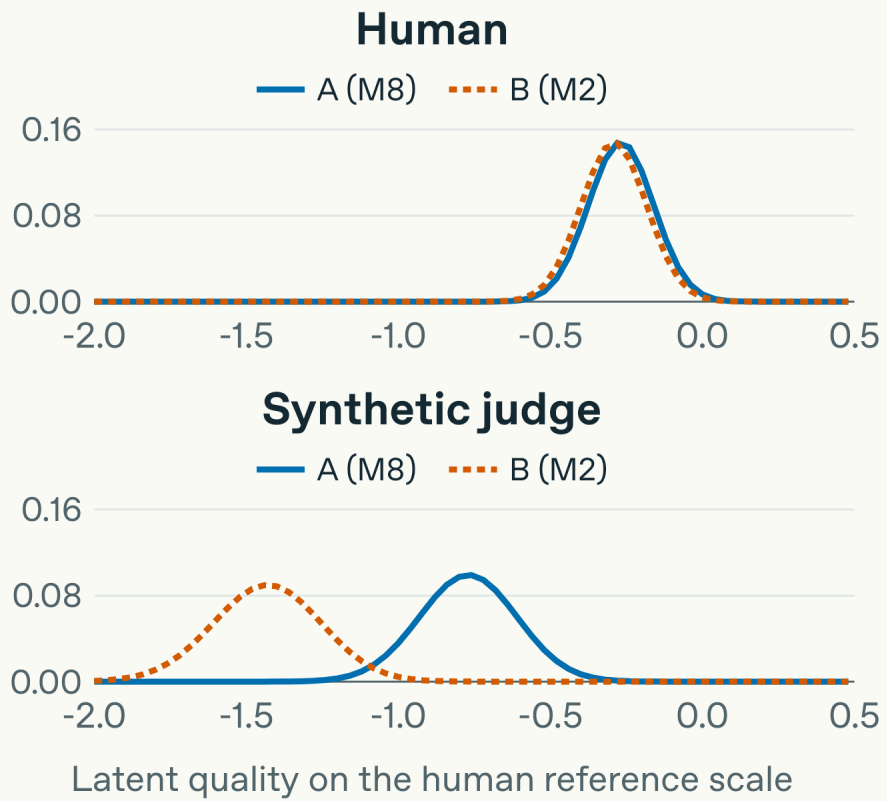
Human: 56.0% — unsupported

Judge: 99.7% — supported

136 pairs evaluated **nSMF = 0.41** **FSR = 17/17**

Risk-sensitive selection maximizes nSMF subject to the declared FSR limit.

The same A/B curves



Latent quality on the human reference scale

The intended claim changes the selected judge

SummEval coherence: fine ruler; all 9 judges pass Stage 2

	Qwen3-14B	GPT-OSS-20B
Calibrated Spearman	**0.836**	0.775
Pairwise fidelity (nSMF)	0.684	**0.689**
False separation rate	0.412	**0.118**

Rank-only: Qwen3-14B

Pairwise or risk-sensitive (FSR ≤ 0.30): GPT-OSS-20B

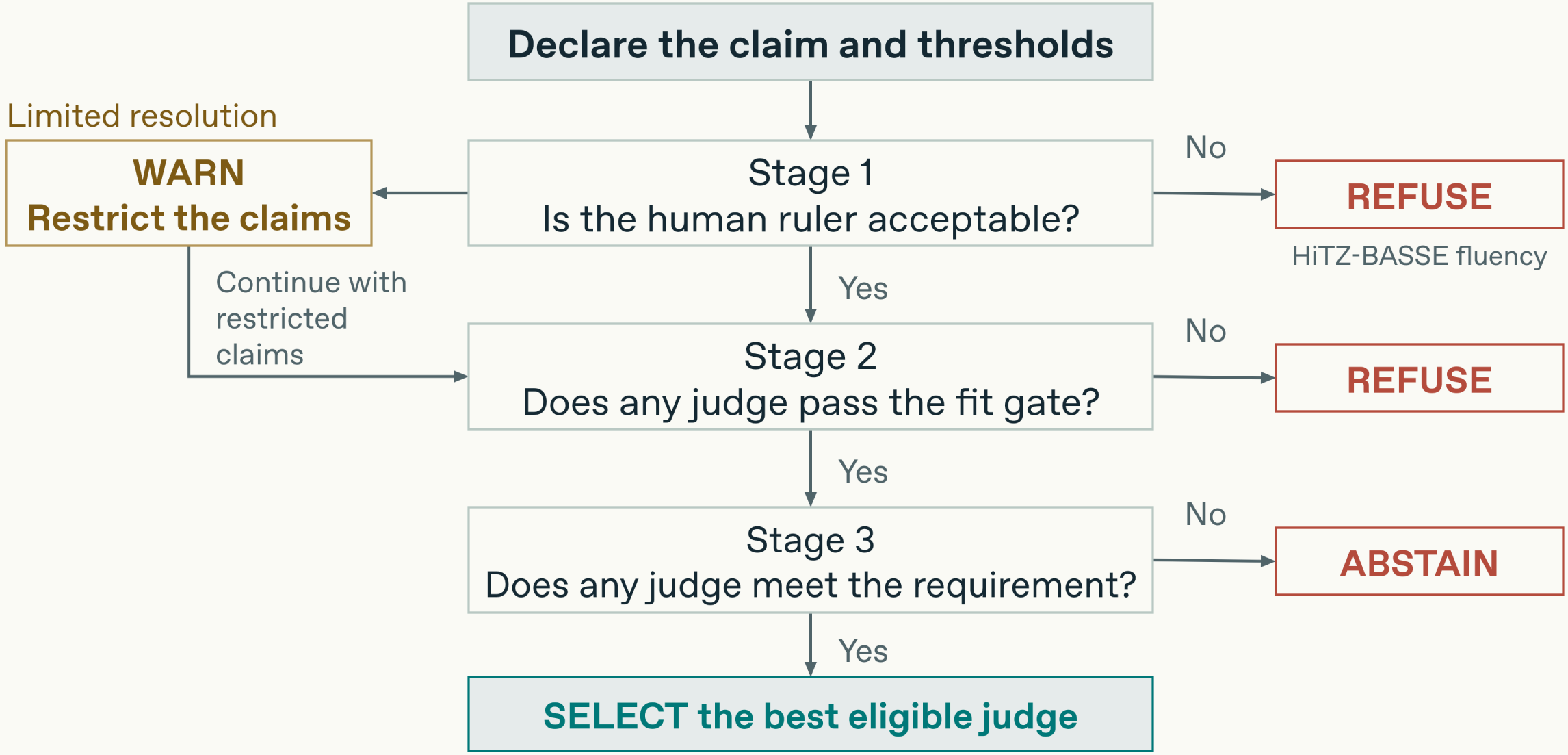
Calibration lowers EMD in 111/135 cases

Earth mover's distance to human rating distributions (lower is better)

135 judge–dimension cases across 9 judges and 3 benchmarks.

The audits lead to an explicit decision

All conclusions remain relative to the benchmark and its annotation protocol.



Judge validation begins with the human ruler

High correlation can conceal unsupported comparisons.

The framework selects a judge only for claims the audits support.

Paper & code

github.com/adriangh-ai/TheRulerAndTheJudge

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

Cofinanciado por
la Unión Europea

AGENCIA
ESTATAL DE
INVESTIGACIÓN

ANNOTATE