AMCIS 2026 · RENO, NEVADA

# Structural Biases in LLM-as-a-Judge Systems: Implications for reliable IS Adoption

Jonathan Piaget, Ulysse Rosselet and Cédric Gaspoz

August 22, 2026

BACKGROUND

# What is LLM-as-a-Judge

human evaluation

costly

slow

doesn't scale

**Use an LLM as the judge**

the "LLM-as-a-Judge" paradigm

already deployed today

Applicant screening

Legal jury assistance

Medical jury assistance

BACKGROUND

# Issues

- sensitivity to prompts and language
- variability depending on the task
- the existence of systematic biases, such as:

01
### Self-preference bias
the judge favors the content it produced itself, or that of closely related models

02
### AI-AI bias
the judge prefers machine-written content over human-written content

03
### Position bias
the order in which the two candidates are shown changes the verdict

*... and others*

WHY IT MATTERS

# One number to keep in mind:

# 23–60%

more likely to be shortlisted when the candidate used the **same LLM as the recruiter**

Simulated hiring pipeline - Xu et al., 2025

METHOD

# Dictionaries as a Human Gold Standard

On subjective benchmarks, human answers vary too much to separate **judge bias** from **content quality**.

**Professional**

curated reference sources with verified non-LLM provenance

**Comparable**

same task, same target format, so sources sit side by side

**Objective**

stable references, widely accepted as correct

If strong biases appear even in this **narrow, controlled setting**, there is no reason to expect them to disappear in fuzzy organizational tasks.

RESEARCH QUESTION

To what extent do intrinsic biases - such as AI-AI bias, self-preference and position - pose risks to reliable acceptance of the LLM-as-a-Judge paradigm in organizational context?

METHOD

# Corpus

**100 English terms**
COCA frequency list

**5 dictionaries** → 500 human definitions
Merriam-Webster, Cambridge, Oxford ALD, Collins, Wiktionary

**4 LLMs** → 400 synthetic definitions
GPT-5, Mistral Large, DeepSeek R1, Llama 3 8B (local)

**Generation prompt** - deliberately minimal:

*"You are a dictionary editor. For each word received, produce a dictionary definition entry. Include: part(s) of speech, word senses with definitions, and usage examples where helpful."*

METHOD

# Evaluation Protocol

- **Blind pairwise comparison**: definition A vs B
- **Anonymized**: sources hidden from the judge
- **Randomized order**: A first or B first, 50/50

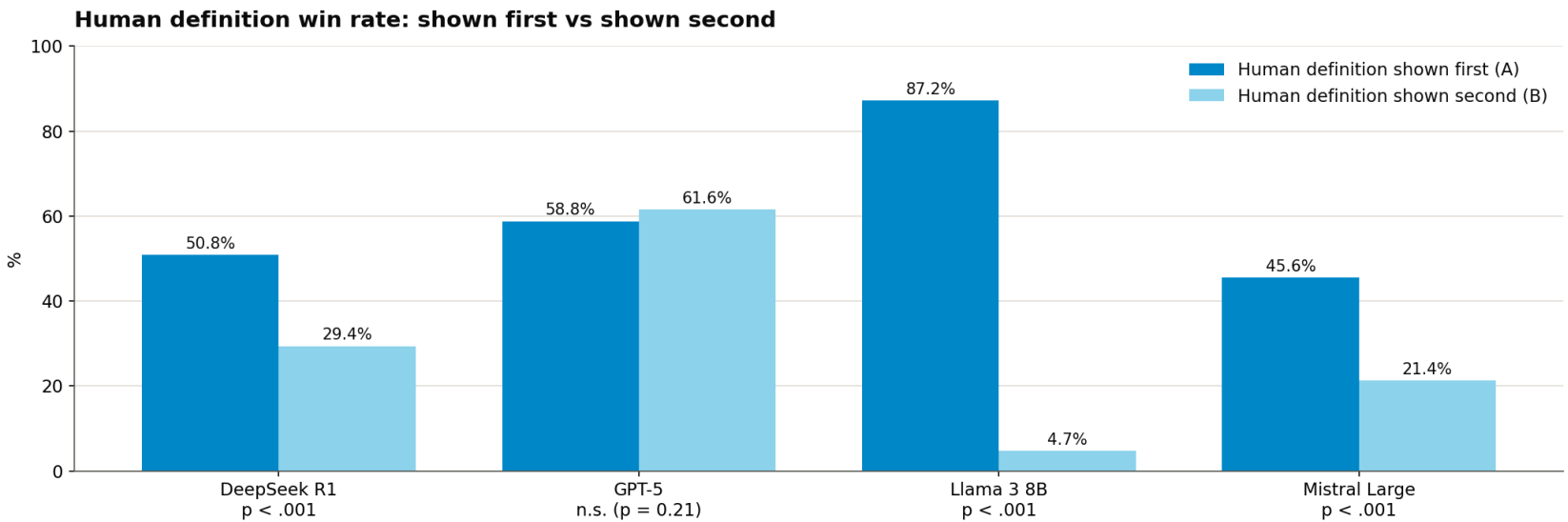
| | | |
|---|---|---|
| 4 LLMs × 5 dicts | → | 20 pairs/term |
| 20 × 100 terms | → | 2,000 pairs |
| 2,000 × 4 judges | → | **8,000 judgments** |

**Judge prompt** - role, no criteria:

> *"You are the editor of an English language dictionary. Two writers have proposed definitions for a given word. It is up to you to decide which definition to include in the dictionary."*

RESULT 1 · POSITION BIAS

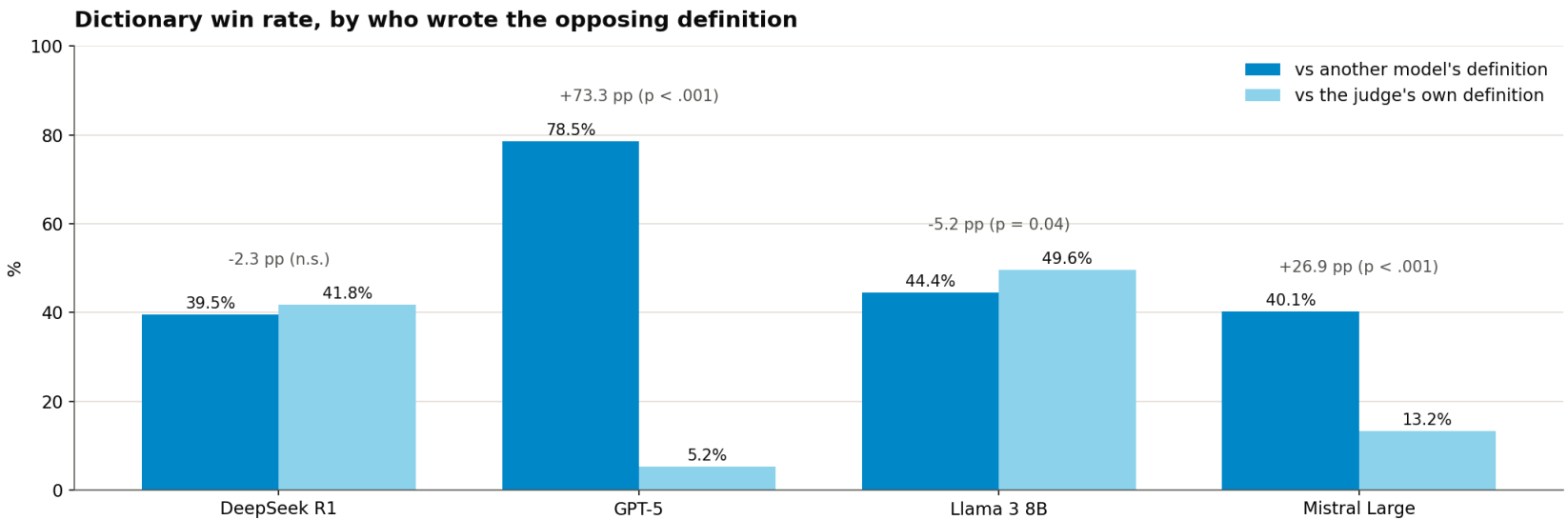
# The definition shown first wins 65.7% of the time



RESULT 2 · SELF-PREFERENCE

# Two of four judges favor their own definitions

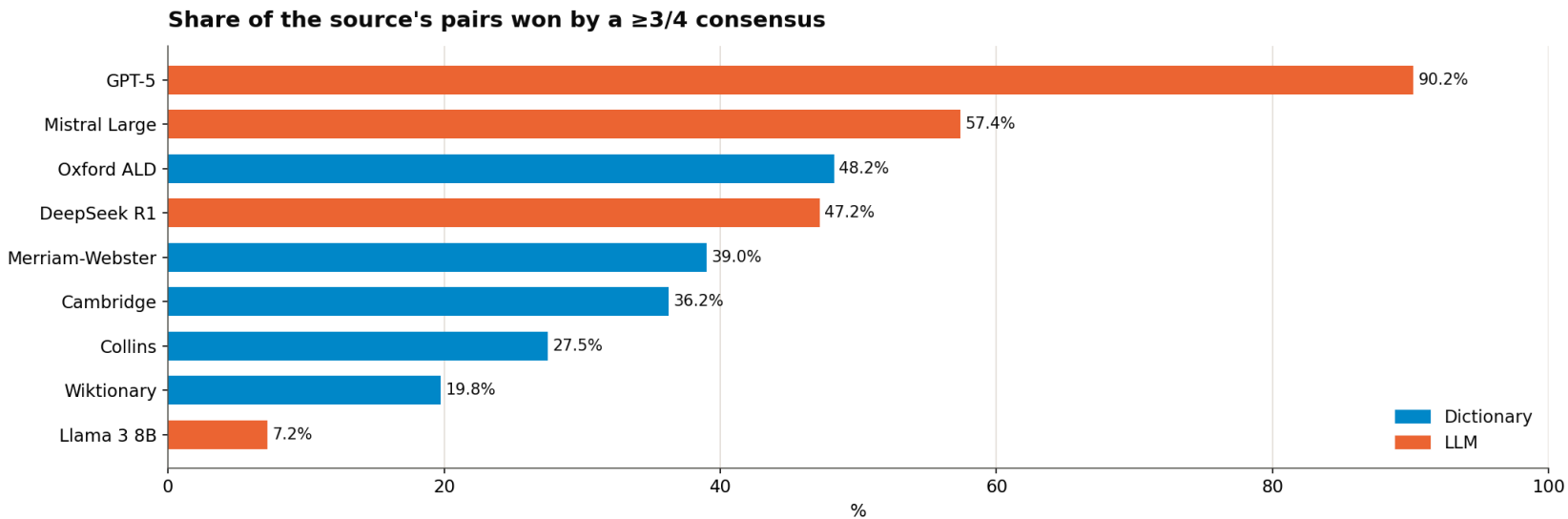
Dictionary wins 50.6% vs other models' output · 27.5% vs the judge's own output



RESULT 3 · CONSENSUS

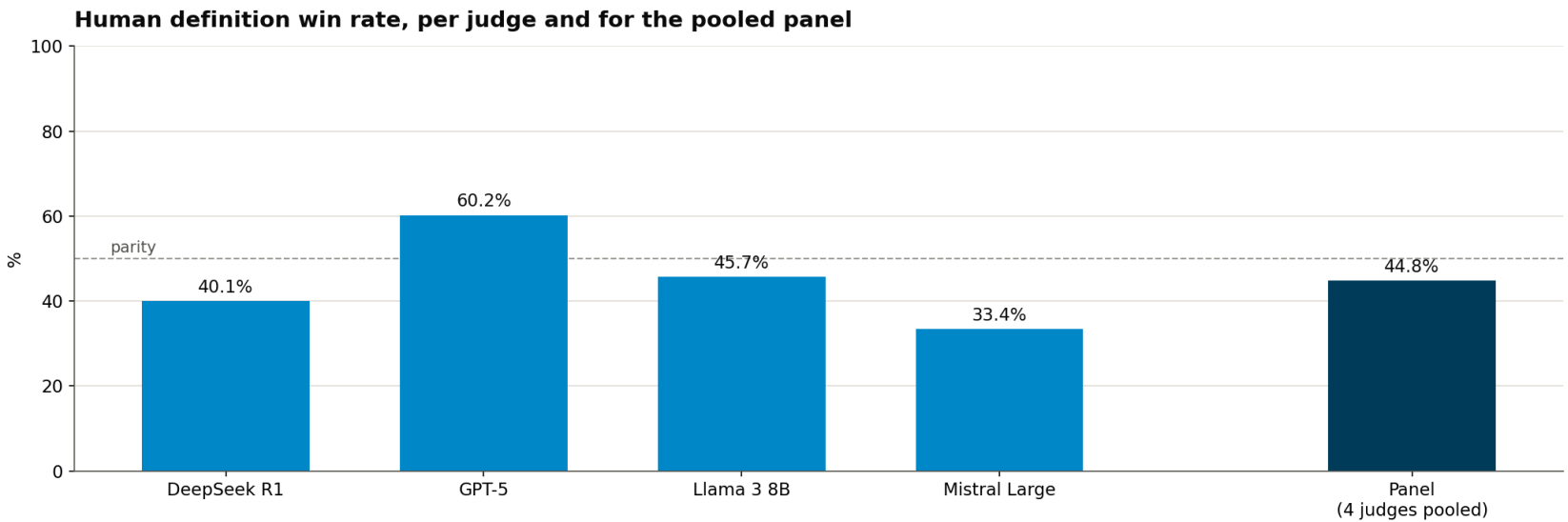
# At committee level, LLM definitions win half of all pairs

Committee verdict: LLM definition wins 50.5% · dictionary wins 34.2% · no consensus (2–2 split) 15.3%



RESULT 4 · WIN RATES

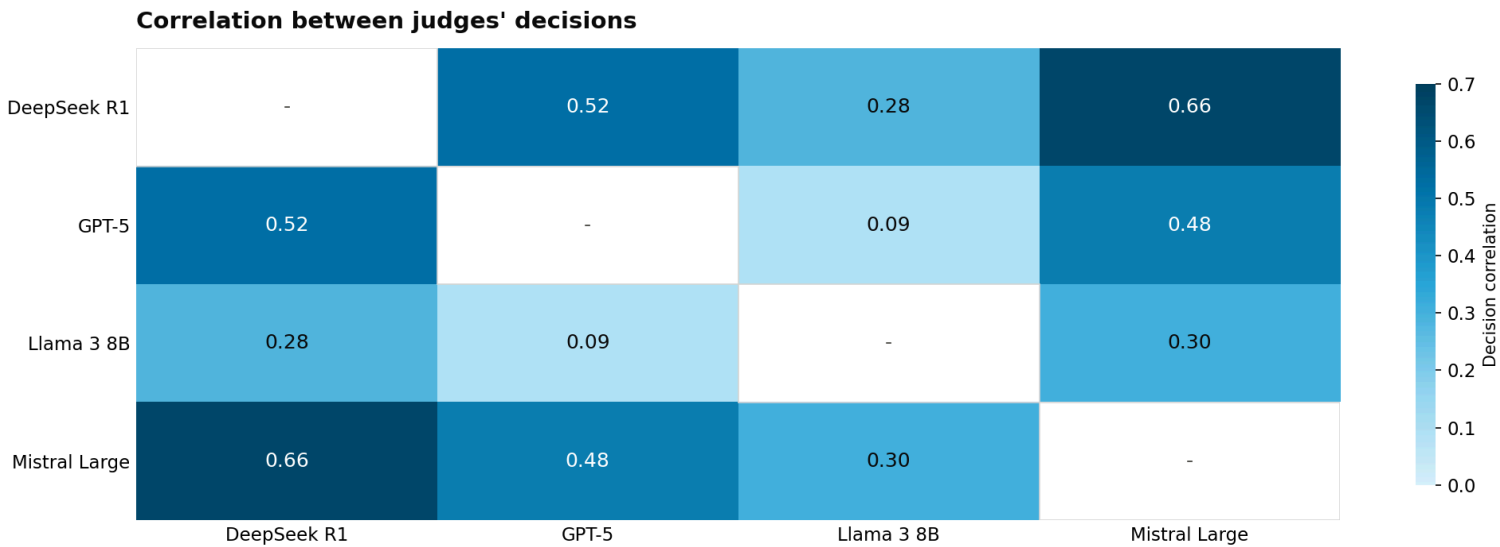
# Opposite individual biases average out in the panel



RESULT 5 · INTER-JUDGE AGREEMENT

# Judges barely agree with each other

Fleiss' κ = 0.357 (moderate) · pairwise decision correlation from 0.09 to 0.66



# Conclusion

Narrow, controlled setting: curated human references, blind protocol

### Biases still massive

**65.7%** position - the definition shown first wins (3/4 judges)

**+73.3 pp** self-preference, at worst

**55.2%** AI-AI - the panel's preference for synthetic text

Widely adopted, yet not neutral even here. The trust we grant these judges must be **calibrated, not assumed**

**01**
Consider **measuring** the biases of your judges

**02**
Favor **blind, diverse panels**

**03**
Keep **calibrating** against human ground truth

AMCIS 2026
RENO • NEVADA
AUGUST 20-22, 2026

# Thank You!

## Questions?

jonathan.piaget@he-arc.ch

piagetjonathan.ch/amcis

Scan for the slides, the article, and my contact

haute école
neuchâtel berne jura
arc gestion
neuchâtel delémont