GlobalHRBench explained: does the language of a legal question change an LLM's answer?

25 November 2026 · 7 min read · blogs.iictp.org/globalhrbench-multilingual-employment-law-llm-accuracy-explained

Short answer: GlobalHRBench is a benchmark in an arXiv preprint submitted on 23 November 2026 by Yuding Wang and Yuhang Yao that pairs 240 employment-law problems across eight jurisdictions and eight languages, with facts, applicable law and required conclusions held fixed. Across nine models and 1,920 prompts, the best model reached 76.2% strict passes with a 5.8% relative decline from its best to its worst language, against 42.2% for the weakest listed. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- GlobalHRBench pairs 240 employment-law problems across eight jurisdictions and eight languages, with facts and correct conclusions held fixed.
- The authors report that answers vary with the language of the question, and that some models swap rank between English and Chinese.
- It is a narrow, closed-book preprint, so test legal tools with known-answer questions in each language your clients use.

Read online, listen to it, or change the text size: https://blogs.iictp.org/globalhrbench-multilingual-employment-law-llm-accuracy-explained

What is GlobalHRBench?

LLM agents are deployed across borders, and law has to hold across working languages. The benchmark asks whether changing only the language of a question changes answer accuracy. It pairs 240 employment-law problems across eight jurisdictions and eight languages, giving 1,920 prompts.

How it works

- **Controlled design.** Facts, applicable law and required conclusions are fixed across languages.
- **Closed-book, single-answer trials.** Models answer without looking things up.
- **Three scores.** Legal correctness, completeness and answer-language compliance.
- **Evidence study.** A four-model study on 16 paired families compares closed-book answers with governing evidence, added cross-jurisdiction distractors and removal of a governing passage.

What is new compared with earlier approaches

Most legal benchmarks are in one language. This one isolates the effect of language by holding the legal content constant.

Results reported by the source

- The top model, named in the paper as GPT-6.1 Sol, leads at 76.2% strict passes with a 5.8% relative best-to-worst language decline; DeepSeek V4.1 Flash is reported at 42.2%.
- Kimi K3.1 Pro and Claude Sonnet 5.5 reverse order between English and Chinese.
- Governing evidence corrects legal errors and shortens reasoning on recovered answers, but recognising an evidence gap does not ensure correct application of the remaining rules.

Limitations and open questions

- Employment law across eight jurisdictions is a narrow slice of legal work.
- Closed-book trials do not match how a lawyer works with sources.
- Model rankings change quickly. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to test a legal AI tool in your own language

1. Take ten questions with known answers in your jurisdiction.
2. Ask them in each working language of your clients.
3. Compare correctness, not only fluency.
4. Add the governing statute to the prompt and check whether errors disappear.

What it means for lawyers and legal-tech buyers

A model strong in English may be weaker in another language, and the gap can vary by model. If your practice is multilingual, test in the language the user writes in. Retrieval of the governing text helps but does not guarantee correct application. See SG-LegalClarify and HireWatch.

Frequently asked questions

- **What is GlobalHRBench?** A benchmark of 240 employment-law problems across eight jurisdictions and eight languages.
- **What is the headline result?** The top model scores 76.2% strict passes with a 5.8% relative best-to-worst language decline.
- **Does adding the law help?** It corrects legal errors, but does not guarantee correct application.
- **Is it peer reviewed?** No, it is a preprint.

Related reading

- SG-LegalClarify: legal reasoning and ambiguity
- HireWatch: LLM compliance with US employment law
- AI Practitioner certification

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Wang Y, Yao Y. GlobalHRBench: Does Changing the Question Language Change the Legal Answer? arXiv 2611.29852, 23 Nov 2026 (preprint). arxiv.org/abs/2611.29852

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/globalhrbench-multilingual-employment-law-llm-accuracy-explained. More guides: blogs.iictp.org.