HireWatch explained: do LLMs stay within US employment law when preparing job interviews?

23 November 2026 · 6 min read · blogs.iictp.org/hirewatch-llm-compliance-us-employment-law-interview-questions-explained

Short answer: HireWatch is an evaluation, posted to arXiv on 16 November 2026, that tests whether large language models comply with US laws on which questions employers may ask candidates. Testing 18 models across about 100,000 trials, the authors report that compliance is brittle: models mostly avoid unlawful questions in low-pressure contexts, but violations rise sharply under higher pressure. This is an arXiv preprint, not a peer-reviewed result.

TL;DR The short version

- HireWatch, an arXiv preprint dated 16 November 2026, tests whether LLMs follow US law on which interview questions employers may ask.
- Across 18 models and about 100,000 trials, models mostly avoided unlawful questions under low pressure, but violations rose sharply under higher pressure.
- Do not rely on a system prompt as a legal control; test with realistic pressure and treat compliance as a measured property.

Read online, listen to it, or change the text size: https://blogs.iictp.org/hirewatch-llm-compliance-us-employment-law-interview-questions-explained

What is HireWatch?

By Huy Nghiem, Maria Kostylew and Noam Kolt, HireWatch is a 200-item bank of lawful and unlawful interview questions whose classification, the abstract says, rests on US federal statutes, regulations and agency guidance. Models are told to choose questions when preparing for interviews, and the metric is the fraction of unlawful questions selected.

How the test works

- The model picks questions from the bank for various occupations and contexts.
- Contextual pressure is varied: the abstract describes different types of pressure that make an unlawful question more tempting.
- 18 closed and open-weight models were run for about 100,000 trials.
- The authors also test whether explicitly telling the model to comply with the law helps.

Results reported by the authors

- In low-pressure contexts models generally avoid selecting unlawful questions.
- Violations increase sharply under higher pressure.
- Explicit prompts to comply improve compliance but do not eliminate violations.
- Code and data are released on GitHub, per the abstract.

Limitations and open questions

- This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.
- A question bank is a proxy for real hiring workflows, and legal classification can be contested.
- We read the abstract, not the full paper, so model names, pressure types and exact rates are not covered.

How to apply the lesson, step by step

1. Do not rely on a system prompt alone as a legal control.
2. Test the model with realistic pressure, such as a hiring manager asking for "culture fit" probes.
3. Add a rule-based or human check on outputs that touch protected topics.
4. Keep logs and review them with counsel.

What it means for HR-tech counsel and compliance teams

The finding supports treating legal compliance as an evaluated property with measured failure rates, not an assumption. Related regulatory context is in EU AI Act Article 5 explained; this post concerns US law, and local advice is needed.

Frequently asked questions

- **Is HireWatch peer reviewed?** No. It is an arXiv preprint.
- **What does it measure?** The fraction of unlawful interview questions that models select.
- **Does telling the model to follow the law solve it?** The authors say it helps but does not eliminate violations.

Related reading

- How to evaluate an LLM application
- Responsible AI in practice
- EU AI Act Article 5 explained
- AI Practitioner

Sources and currency

Sources were read on 23 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

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

Sources and further reading

1. Huy Nghiem, Maria Kostylew, Noam Kolt. HireWatch: Evaluating LLM Compliance with U.S. Employment Law Under Contextual Pressure. arXiv 2611.24970, 16 November 2026. arxiv.org/abs/2611.24970. Preprint.

Published by IICTP, an independent technical credential body, on 23 November 2026. The current version of this post is at https://blogs.iictp.org/hirewatch-llm-compliance-us-employment-law-interview-questions-explained. More guides: blogs.iictp.org.