DO LLMS TAKE SIDES? ADJUDEX-LABOR EVALUATES POSITIONAL TENDENCIES IN LABOR DISPUTES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluations of legal large language models (LLMs) often emphasize agreement with human adjudicators, overlooking whose interests their disagreements favor. In labor disputes, this omission is consequential: denying an employment relationship can exclude workers from legal protection, and repeated departures may concentrate legal risks on the same party. We introduce Adjudex-Labor, a benchmark of real Chinese labor disputes that examines positional tendencies across arbitration and subsequent judgments, using human decisions as institutional references rather than normative ground truth. Our evaluation distinguishes how often models disagree with human adjudicators from differences in the rates of employer- and employee-favoring departures. Employer-favoring tendencies predominate in independent arbitration. When the same model configuration performs arbitration and judgment, replacing human arbitration outcomes with model-generated outcomes shifts subsequent judgments toward employers. Crucially, closer agreement with human courts can coexist with more employer-favoring departures, while additional reasoning does not consistently improve positional balance. A separate monetary evaluation finds that most configurations award employees less than matched human adjudicators, including some that favor employees or appear balanced in employment-status decisions. These findings show that institutional agreement and positional balance are distinct properties, and that positional tendencies depend on both the adjudicative context and the outcome evaluated. LLM evaluation must therefore examine not only agreement with human decisions, but also how departures distribute protection, risk, and remedies between parties and across decision stages.

1 INTRODUCTION

Bias and fairness in large language models (LLMs) have long been central concerns in both academic research and public debate on AI ethics (Gallegos et al., 2024). However, the positional tendencies of LLMs in institutional decision-making remain understudied. A crucial question is whether AI-assisted decisions, including those produced under the increasingly common LLM-as-a-Judge paradigm, will mitigate or reinforce existing social inequalities when incorporated into professional practice. This risk need not arise from conspicuous failures. Small but stable differences in the positions favored by model errors may be repeatedly reproduced across cases and decision stages, eventually altering how institutional protection and legal risk are distributed.

Labor law provides a particularly consequential setting in which to examine this question. Whether an employment relationship exists determines access to wages, social insurance, compensation, and other labor-law remedies, making worker classification a central concern across jurisdictions. Chinese labor legislation, for example, expressly identifies the protection of employees' lawful rights and the maintenance of harmonious and stable employment relations as core statutory objectives (NPC Standing Committee, 1994; 2007). Crucially, classification errors favoring opposing parties are not institutionally equivalent: erroneously recognizing an employment relationship may extend legal protection, whereas erroneously denying one may exclude a worker from the protective framework altogether. Because such disputes are resolved through specialized arbitral and judicial procedures, repeated errors are not confined to isolated cases. They may affect who receives le-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

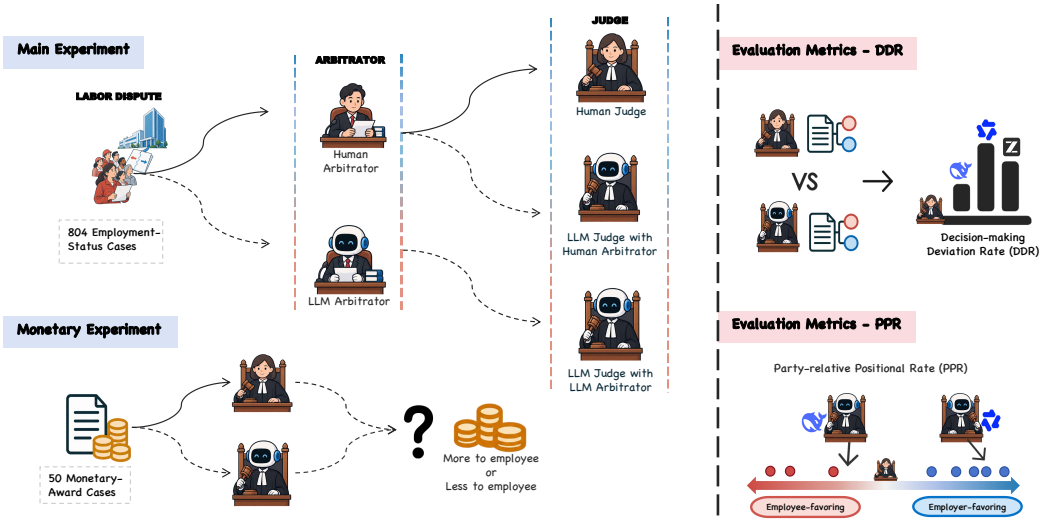


Figure 1: Overview of Adjudex-Labor.

gal protection, who bears responsibility, and which claims obtain meaningful remedies within an employment structure already marked by unequal bargaining power and access to evidence.

Existing research measures bias and legal competence, but leaves the positional consequences of replacing human adjudicators insufficiently examined. Bias benchmarks such as BBQ evaluate stereotyped responses concerning protected groups (Parrish et al., 2022). LegalBench and LawBench assess legal reasoning and knowledge across diverse tasks (Guha et al., 2023; Fei et al., 2024), while CLC-UKET evaluates outcome prediction in employment disputes (Xie et al., 2024). These approaches provide essential foundations, but do not directly establish whether departures from human decisions consistently benefit one party in a concrete dispute, or whether such positional patterns extend from status determination to monetary relief.

We ask three questions. First, how often do LLM decisions depart from arbitral and judicial reference decisions, and which party do those departures favor? Second, how do institutional alignment and positional balance differ between matched LLM-arbitrator and LLM-judge settings, and how does LLM review compare with human judicial review? Third, do similar positional patterns appear in access to relief and the calculation of monetary awards? Together, these questions distinguish professional legal performance from the distributional consequences of model error.

We introduce **Adjudex-Labor**, a benchmark that makes such positional deviation observable in real employment disputes. Built from 804 judicial decisions, it evaluates 14 model configurations in LLM-arbitrator and LLM-judge settings, supports matched analysis on cases with explicit arbitral outcomes, and extends relation determination to monetary judgments in 50 cases. The results reveal a predominant but not universal employer-favoring pattern: most models more often deny institutionally recognized employment relationships than make the reverse error, while several configurations favor workers or remain close to positional balance. Models show lower deviation relative to the corresponding institutional reference in the LLM-judge setting, but positional balance does not improve uniformly. The monetary evaluation further shows that most configurations award employees less, on average, than the corresponding human adjudicators, including some configurations classified as employee-favoring or middle in employment-relationship decisions. Our study therefore connects institutional alignment, positional tendency, and remedial consequences within a single framework, showing that legal LLMs must be evaluated not only by whether they reach professionally plausible decisions, but also by how their errors redistribute legal protection and risk.

2 Related Work

AI fairness research has established that automated decisions can distribute opportunities, resources, and burdens unevenly, but it usually evaluates these distributions across individuals or social groups. Classical work formalizes group disparities through criteria such as error-rate parity (Hardt et al., 2016). In NLP, defining bias also requires specifying the harms at stake and who experiences them (Blodgett et al., 2020). Allocative accounts distinguish fair prediction from distributive justice

(Suresh & Guttag, 2021; Kuppler et al., 2022; Das, 2022). Recent LLM studies extend this concern to hiring, salary recommendations, and economic allocation, where models exhibit demographic disparities and structured distributional preferences (An et al., 2024; Nghiem et al., 2024; Cook et al., 2026). Similar concerns arise in adjudication because legal decisions allocate recognized claims, liabilities, and remedies between parties (Huq, 2026). Yet the relevant comparison in a bilateral dispute is not only whether different people or groups receive different outcomes. When AI replaces a human adjudicator, the human–AI difference itself can shift the allocation between opposing parties. This motivates our party-relative question: when an LLM departs from a human adjudicator, whose legal position does that departure favor?

Legal-AI research increasingly compares model decisions with human or institutional outcomes, but it usually interprets human–AI differences through prediction quality, behavioral sensitivity, or fairness attributes rather than through the interests of opposing parties. First, predictive-justice research asks whether algorithms can improve on human decisions. Work on pretrial release, for example, compares machine predictions with judges' decisions in terms of crime reduction and detention outcomes (Kleinberg et al., 2018). Second, controlled behavioral studies examine whether LLMs and human judges respond differently to legally relevant or irrelevant factors. Posner & Saran (2026), for instance, replicate an experiment with U.S. federal judges and compare model and judicial responses to precedent and sympathy. Third, recent benchmarks evaluate disparities and sensitivity within LLM adjudication. JudiFair measures inconsistency, bias, and imbalanced inaccuracy across demographic, substantive, and procedural attributes (Hu et al., 2026), while labor-law research compares LLM-generated compensation with 10,000 real judgments and examines sensitivity to social-media pressure (Xie et al., 2026). Related work also studies perceived fairness under human, algorithmic, and hybrid adjudication (Barysė & Sarel, 2024; Chen et al., 2026a). Across these approaches, human decisions serve as prediction targets, empirical baselines, or behavioral comparison points. What remains less developed is the distributive meaning of the deviation itself. A model may depart from a human adjudicator by the same apparent magnitude while producing opposite consequences for the parties. *Positional difference* therefore organizes human–AI divergence according to which party's legal position the departure favors.

A positional tendency observed at one adjudicative stage may not characterize a linked decision process. Research on compound decision systems shows that fairness properties of individual stages need not compose into fairness of the overall pipeline (Bower et al., 2017). Dynamic-fairness research further shows how present decisions can alter later opportunities and populations through feedback and behavioral responses (Liu et al., 2018; Heidari et al., 2019; Wen et al., 2021). Within legal evaluation, JurisBench models sequential, interdependent stages of civil adjudication and examines error propagation across the workflow (Chen et al., 2026b). Studies of human–AI adjudication also examine how the placement of algorithmic assistance affects perceived fairness (Barysė & Sarel, 2024; Chen et al., 2026a). These studies establish the importance of stages and dependencies, but leave a distinct distributive question: whose legal position benefits when model decisions replace human decisions across arbitration and judicial review? Adjudex-Labor compares judgments conditioned on human versus model arbitration to examine whether party-relative positional tendencies attenuate, persist, or strengthen across linked adjudicative stages. We further ask whether such tendencies align across employment-status recognition and monetary relief.

3 Adjudex-Labor Benchmark

3.1 Cases and Human References

The main experiment contains 804 real Chinese labor disputes in which employment status was contested and both arbitral and judicial outcomes could be identified.[1] We denote the observed arbitral and court decisions as **Human Arbitrator** and **Human Judge**, respectively. They serve as empirical institutional references for measuring human–AI differences, not as normative ground truth. A separate 50-case sample with positive recorded awards at both human stages is used for the monetary analysis. Output extraction and checks are summarized in Appendix A.

[1] Chinese labor disputes must first undergo labor arbitration before court litigation; final-award rules limit later judicial remedies in certain cases.

Table 1: Case counts and human adjudication outcomes.

	Employment relationship		
Human adjudicator	**Recognized**	**Not recognized**	**Total**
Human Arbitrator	556 (69.2%)	248 (30.8%)	804
Human Judge	416 (51.7%)	388 (48.3%)	804
Monetary-award sample			50

3.2 Main Experiment: Employment-Relationship Adjudication

The main task asks whether an employment relationship exists between the employee and employer, with Yes encoded as 1 and No as 0. At arbitration, **LLM Arbitrator** receives case materials without the recorded human arbitration result and makes an independent determination. At judgment, **LLM Judge with Human Arbitration** receives the recorded human arbitration outcome, whereas **LLM Judge with LLM Arbitration** is conditioned on an LLM-produced arbitration outcome. Human judicial reasoning and the final court disposition are excluded from all LLM Judge inputs.

To isolate the effect of replacing human arbitration with LLM arbitration, the key sequential comparison holds the Judge configuration fixed and varies only the arbitration outcome supplied to the judgment stage. Only the binary arbitration outcome is propagated between stages. Let $h_i \in \{0, 1\}$ denote the Human Arbitrator outcome for case i, and $N_i^{(u)}$ the independent arbitration decision of Arbitrator configuration u. For Judge configuration v, $T_i^{(v)}$ denotes the judgment when the supplied arbitration outcome equals h_i, and $F_i^{(v)}$ the judgment when it is reversed. We reconstruct

$$J_i^{(u,v)} = \begin{cases} T_i^{(v)}, & N_i^{(u)} = h_i, \\ F_i^{(v)}, & N_i^{(u)} \neq h_i. \end{cases} \tag{1}$$

This selects the Judge response consistent with the LLM arbitration outcome while holding the case and Judge configuration fixed. Diagonal pairs ($u = v$) provide the same-configuration comparison, while the 14 Arbitrator configurations crossed with the 14 Judge configurations yield $14 \times 14 = 196$ arbitration–judgment pairings. Because only the binary outcome is propagated, the analysis tests sensitivity to preceding adjudicative outcomes rather than transmission of generated legal reasoning.

3.3 Models, Metrics, and Statistical Analysis

We evaluate 14 configurations from seven model families: Claude Opus 4.8, GPT-5.6-Sol, Gemini 3.6 Flash, Qwen3.8 Flash, Kimi K3, GLM-5.2, and DeepSeek-V4-Pro. Each family is evaluated in Non-Thinking (NT) and Thinking modes.

For a human reference with n_1 positive and n_0 negative cases, let e denote disagreements in which the reference is 1 and the model predicts 0, and w the reverse. We define

$$\mathrm{DDR} = 100\,\frac{e+w}{n_1+n_0} \quad (\%), \qquad \mathrm{PPR} = 100\left(\frac{e}{n_1} - \frac{w}{n_0}\right) \quad (\mathrm{pp}). \tag{2}$$

Decision-making Deviation Rate (DDR) measures the frequency with which a model departs from the reference decision, whereas Party-relative Positional Rate (PPR) captures the direction of those departures by comparing conditional disagreement rates across the two reference outcomes. Positive PPR indicates an employer-favoring imbalance and negative PPR an employee-favoring imbalance, while values near zero indicate little directional asymmetry. Conditioning on the reference outcome prevents differences in label prevalence from mechanically producing positional imbalance. For descriptive interpretation, we classify $\mathrm{PPR} < -5$ pp as employee-favoring, $-5 \leq \mathrm{PPR} \leq 5$ pp as middle, and $\mathrm{PPR} > 5$ pp as employer-favoring; these thresholds are not inferential, and all analyses retain the continuous PPR values. Together, DDR and PPR separate the frequency of model–human divergence from its directional imbalance.

Reference decisions follow the institutional stage being evaluated. LLM Arbitrator is compared with Human Arbitrator, whereas all judgment conditions and sequential chains are compared with Human Judge. We therefore use $\mathrm{DDR}_{\mathrm{arb}}$ and $\mathrm{DDR}_{\mathrm{jud}}$ to denote deviation rates relative to Human

Arbitrator and Human Judge, respectively. For pairwise comparisons, differences in DDR and PPR are reported with 95% percentile bootstrap intervals based on 10,000 case-level resamples, with all outputs associated with the same case resampled jointly.

3.4 Monetary-Award Experiment

The monetary experiment examines whether employment-status tendencies align with monetary relief in 50 cases with positive awards at both human stages. Unlike the main experiment, it propagates generated arbitral documents rather than binary outcomes between stages.

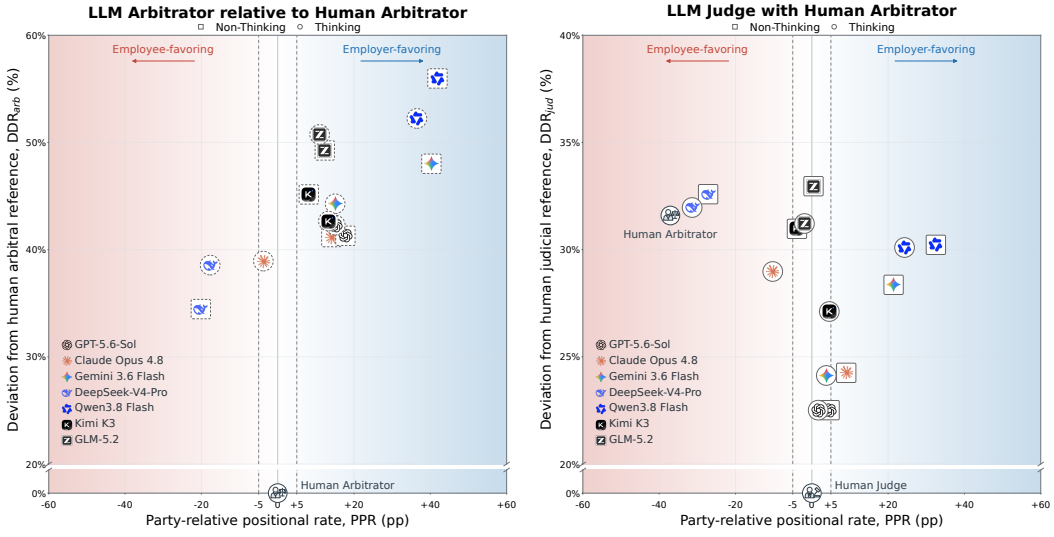
LLM Arbitrator generates an arbitral decision and award without access to the recorded human arbitration result. The same model configuration then produces two first-instance judgments: **LLM Judge with Human Arbitration** receives the human arbitral material, whereas **LLM Judge with LLM Arbitration** receives the generated arbitral document. Both conditions exclude the Human Judge's reasoning and final disposition.

We extract total substantive employee awards, excluding litigation fees and delayed-performance interest, and report the mean case-level difference between each model award and its matched human reference in thousand CNY. LLM Arbitrator is compared with Human Arbitrator, whereas both Judge conditions are compared with Human Judge. Positive gaps indicate higher employee awards than the matched human reference and negative gaps lower awards, opposite to the sign convention for PPR. Two unquantified LLM Arbitrator amounts are excluded only from analyses requiring numerical awards. This experiment is exploratory and assesses cross-task alignment rather than causal effects.

4 Positional Tendencies in Employment-Relationship Adjudication

This section traces positional tendencies across the adjudicative sequence, from single-stage arbitration and judgment outcomes to the effects of replacing Human Arbitration with LLM Arbitration and to the full set of arbitration–judgment pairings. All analyses use the same 804 cases and 14 model configurations.

4.1 Positional Tendencies Across Arbitration and Judgment



(a) LLM Arbitrator **(b) LLM Judge with Human Arbitrator**

Figure 2: (a) LLM Arbitrator relative to Human Arbitrator. (b) LLM Judge with Human Arbitration relative to Human Judge, with Human Arbitrator shown for comparison.

LLM Arbitrators vary substantially in the magnitude of their deviation from human arbitration, but the direction of those deviations is concentrated on the employer-favoring side. Across the 14 configurations, DDR ranges from 34.45% for DeepSeek-V4-Pro Non-Thinking to 55.97% for Qwen3.8 Flash Non-Thinking (Figure 2a). Reasoning mode does not uniformly reduce this deviation: switch-

ing to Thinking lowers DDR for Claude, Gemini, Qwen, and Kimi, but increases it for GPT, GLM, and DeepSeek. By contrast, positional direction is considerably more consistent. Relative to Human Arbitrator, 11 of 14 configurations are classified as employer-favoring, and the overall pattern remains largely unchanged under an alternative human reference. In short, model families differ markedly in how far they depart from human decisions, but those departures are more consistently concentrated on the employer-favoring side.

LLM Judges show lower overall deviation from Human Judge, but their positional distribution is more dispersed across employer-favoring, middle, and employee-favoring outcomes. With Human Arbitration supplied at judgment, DDR ranges from 22.51% to 32.96% across the 14 configurations, with both GPT-5.6-Sol modes at the lower end and GLM-5.2 Non-Thinking at the upper end (Figure 2b). Compared with arbitration, positional outcomes are no longer strongly concentrated: relative to Human Judge, seven configurations fall within the middle interval, four are employer-favoring, and three are employee-favoring. Human Arbitrator itself differs from Human Judge in 31.59% of the 804 cases and lies on the employee-favoring side of the court reference, providing an important institutional baseline for interpreting the judgment-stage distribution. Thus, although LLM judgments are closer to Human Judge in aggregate, positional direction becomes more heterogeneous across model configurations.

Taken together, the transition from arbitration to judgment is associated with lower model–human deviation and a shift from a strongly employer-favoring distribution toward the middle. From the employee perspective, this makes the judgment-stage pattern less adverse, but not predominantly employee-favoring: only three of the 14 Judge configurations fall on the employee-favoring side. Because adjudicative role and information availability change together, this contrast cannot isolate the source of the shift. The next analysis therefore holds the Judge fixed and varies only the arbitration input.

Reasoning mode shows no uniform effect. At arbitration, Thinking reduces DDR for some model families but increases it for others, and no consistent positional direction emerges across either stage. Thinking may improve agreement for particular models, but it does not systematically reduce positional imbalance.

4.2 Replacing Human Arbitration with LLM Arbitration

Replacing Human Arbitration with LLM Arbitration amplifies the deviations already observed at the judgment stage. Holding the Judge configuration fixed, the model-generated arbitration input moves judgments both farther from Human Judge and further toward the employer side (Figure 3). Across configurations, mean DDR increases by 9.53 percentage points, while mean PPR shifts 8.53 percentage points toward employers. The directional effect is especially consistent: every configuration moves rightward on the PPR axis.

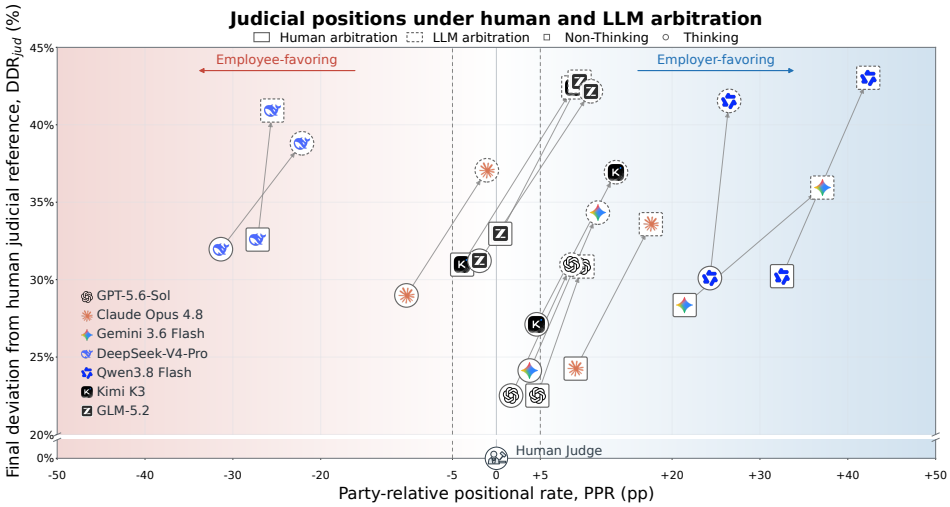


Figure 3: Effect of replacing human arbitration with LLM arbitration on subsequent judgments. Each arrow connects the same Judge configuration under Human Arbitration (solid) and same-configuration LLM Arbitration (dashed). All positions are relative to Human Judge.

The direction of movement should be distinguished from the final positional category. All seven configurations previously classified as middle under Human Arbitration become employer-favoring, while all four configurations already classified as employer-favoring remain so. Yet the shift does not place every configuration on the employer-favoring side: both DeepSeek modes remain employee-favoring, and Claude Thinking remains within the middle interval. LLM Arbitration therefore produces a common employer-ward displacement even though configurations begin from, and ultimately occupy, different positional locations.

A complementary comparison further separates agreement from positional tendency. When independent LLM arbitration and the subsequent LLM judgment are both evaluated against Human Judge, mean DDR decreases by 1.41 percentage points, while mean PPR moves 2.68 percentage points toward employers; 10 configurations exhibit both changes. The judgment stage can therefore move closer to Human Judge in aggregate while the remaining disagreements become more employer-favoring. Agreement with the institutional reference and positional tendency thus capture distinct dimensions of model–human divergence.

These results are based on judgments reconstructed from binary arbitration outcomes rather than transmitted arbitral reasoning. They show that changing the preceding adjudicative outcome can systematically amplify positional deviation in subsequent judgments, but they do not establish self-reinforcement through model-generated legal reasoning. The monetary experiment later examines a separate setting in which generated arbitral documents are transmitted between stages.

4.3 Positional Tendencies Across Arbitration–Judgment Pairings

The employer-favoring shift is not confined to chains in which the same configuration performs both stages. Across all $14 \times 14 = 196$ LLM arbitration–judgment pairings, 147 (75.0%) are employer-favoring, 30 fall within the middle interval, and 19 are employee-favoring (Figure 4). Excluding the 14 same-configuration pairs leaves 182 cross-configuration pairings, of which 136 remain employer-favoring. The pattern is therefore widespread across the evaluated model combinations rather than an artifact of models judging their own arbitration outcomes. These percentages describe model pairings, not the proportion of individual cases decided for employers.

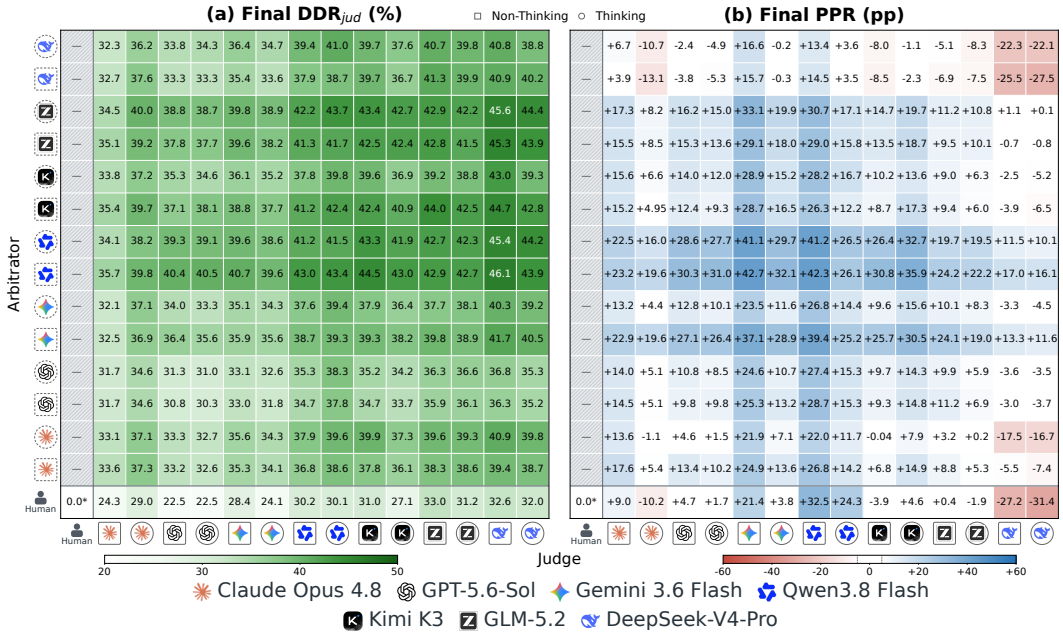


Figure 4: $\mathrm{DDR}_{\mathrm{jud}}$ (a) and PPR (b) for all arbitration–judgment pairings on 804 cases, relative to Human Judge. Rows denote Arbitrators and columns Judges; the Human row gives the Human Arbitration baseline. Hatched cells denote unobserved pairings.

Some configurations exhibit substantial positional stability across partners. As Judges, Gemini Non-Thinking and Qwen3.8 Non-Thinking remain employer-favoring under every one of the 14 LLM arbitration inputs. As Arbitrators, Gemini Non-Thinking and both Qwen3.8 configurations produce

employer-favoring final judgments with every Judge configuration. Changing the model assigned to the other stage is therefore not, by itself, sufficient to neutralize an established positional tendency.

Other configurations show the opposite phenomenon: their position depends sharply on the arbitration input. DeepSeek Thinking, for example, has a PPR of -31.37 pp when judging with human arbitration and remains employee-favoring under DeepSeek Non-Thinking arbitration (-27.46 pp), but becomes employer-favoring under Qwen3.8 Non-Thinking arbitration ($+16.12$ pp). The Judge configuration is unchanged; only the arbitration input differs. Positional tendency is therefore better understood as a property of a model-in-context than as a fixed attribute of the model alone.

The pairing analysis also makes the distinction between disagreement frequency and positional tendency especially clear. DeepSeek Non-Thinking arbitration followed by DeepSeek Thinking judgment produces a $\mathrm{DDR}_{\mathrm{jud}}$ of 40.17% and a PPR of -27.46 pp. Qwen3.8 Non-Thinking arbitration followed by Gemini Non-Thinking judgment has an almost identical $\mathrm{DDR}_{\mathrm{jud}}$ of 40.67%, but a PPR of $+42.74$ pp. The two chains disagree with Human Judge at nearly the same rate while placing those disagreements on opposite sides of the labor dispute. An aggregate disagreement measure would treat them as similar; positional evaluation reveals that they distribute those disagreements very differently.

Taken together, the pairing results reveal both regularity and contingency. Employer-favoring positions dominate the evaluated arbitration–judgment combinations, yet individual configurations can remain stable across partners or switch between employee-favoring and employer-favoring positions under different arbitration inputs. The relevant unit of evaluation is therefore not simply the model or the isolated decision. It is the model within an adjudicative chain, where the source of an earlier decision and the party it favors can materially reshape the positional pattern of the next.

4.4 Analysis of Monetary Awards

The monetary experiment shows that positional tendency is not stable across adjudicative outcomes. Models that appear relatively favorable to employees in employment-status recognition do not necessarily remain so when monetary relief is considered, and the source of the preceding arbitration can further alter award levels. This makes monetary awards an important second dimension of the benchmark rather than a simple extension of the employment-status results.

Holding the Judge configuration fixed, 11 of the 14 configurations produce lower mean awards under LLM Arbitration than under Human Arbitration. Averaged across configurations, the model–human award gap moves from -0.99 to -2.52 thousand CNY, a further decrease of 1.53 thousand CNY. The shift is especially consistent among Thinking configurations: all seven move toward lower awards, compared with four of seven Non-Thinking configurations. DeepSeek-V4-Pro Thinking and GPT-5.6-Sol Thinking even cross zero, moving from mean awards above the Human Judge reference to below it. The same replacement of Human Arbitration with LLM Arbitration therefore moves both employment-status decisions and monetary awards in a less employee-favorable direction.

Lower awards are also common across the three adjudicative conditions. Relative to the matched human reference, 13 of 14 LLM Arbitrator configurations produce lower mean awards; the corresponding counts are 11 of 14 for LLM Judge with Human Arbitration and 13 of 14 for LLM Judge with LLM Arbitration. Reasoning mode, however, does not show a stable pattern. At arbitration, Thinking and Non-Thinking configurations have nearly identical mean gaps (-2.57 versus -2.60 thousand CNY). With Human Arbitration, Thinking configurations are closer to Human Judge (-0.53 versus -1.45 thousand CNY), but this difference narrows substantially under LLM Arbitration (-2.39 versus -2.65 thousand CNY). The monetary effect of reasoning mode therefore varies with the arbitration input.

The clearest cross-task difference appears in configuration–condition pairs that are employee-favoring or middle in employment-status PPR but still produce lower mean awards than the corresponding human reference. Thirteen pairs exhibit this pattern, including seven under LLM Judge with Human Arbitration. DeepSeek-V4-Pro Non-Thinking and Claude Opus 4.8 Thinking show the mismatch under all three adjudicative conditions. Relative favorability in employment recognition therefore does not necessarily extend to monetary relief.

These comparisons are descriptive rather than causal. Employment-status PPR is estimated from the 804-case main experiment, whereas monetary gaps are calculated from a separate 50-case sample using a different generation task. Human awards serve as empirical institutional references rather than normative benchmarks for adequate compensation. Taken together, the two experiments show

Table 2: Mean award gaps in the 50-case sample (thousand CNY). LLM gaps use Human Arbitrator as the reference for LLM Arbitrator and Human Judge for both LLM Judge conditions. †49 matched cases; all other cells use 50. Blue cells indicate employee-favoring or middle employment-relation positions with negative award gaps; employment classifications use the same human references.

Model	LLM Arbitrator	LLM Judge with Human Arbitration	LLM Judge with LLM Arbitration
GPT-5.6-Sol NT	−3.15	−0.18	−3.46
GPT-5.6-Sol Thinking	−1.11	+1.14	−0.93
Claude Opus 4.8 NT	−1.01	−1.14	−0.11
Claude Opus 4.8 Thinking	−1.74†	−0.56	−3.24
Gemini 3.6 Flash NT	−2.93	−2.08	−2.74
Gemini 3.6 Flash Thinking	−2.97	−1.30	−2.78
DeepSeek-V4-Pro NT	−0.19	−2.27	−0.23
DeepSeek-V4-Pro Thinking	−2.04	+0.49	−1.21
Qwen3.8 Flash NT	−5.44	−1.62	−5.25
Qwen3.8 Flash Thinking	−3.05	−0.48	−1.76
Kimi K3 NT	+0.21	+0.28	+0.39
Kimi K3 Thinking	−4.27	−1.39	−4.16
GLM-5.2 NT	−5.67†	−3.14	−7.17
GLM-5.2 Thinking	−2.83	−1.63	−2.64
Human reference comparison			Mean gap
Human Arbitrator − Human Judge			+0.19

that positional tendency depends on the outcome being evaluated: employment-status recognition alone cannot characterize how models allocate monetary relief.

5 Conclusion

LLMs do take sides. On Adjudex-Labor, employer-favoring positions predominate in independent arbitration and judgments with LLM arbitration, while both DeepSeek configurations favor employees. Replacing human arbitration with same-configuration LLM arbitration shifts all 14 Judge configurations toward employers, although their final positions differ. Thinking configurations show no consistent advantage in positional balance over their Non-Thinking counterparts. Additional reasoning therefore does not necessarily produce more balanced adjudication.

Institutional alignment and positional balance should be treated as separate properties of legal decision-making. Arbitration–judgment pairings with nearly identical DDR can have PPR values of opposite signs; closer agreement with human courts can also coexist with movement toward employers. Aggregate agreement alone can therefore conceal substantial differences in how model errors distribute legal benefits and burdens.

These differences matter because their consequences are not distributed evenly. Employment-status decisions determine access to labor-law protection, while monetary decisions determine the amount of relief. An employer-favoring tendency may shift legal risk toward workers; an employee-favoring tendency may extend protection beyond the institutional reference. In the 50-case monetary sample, most configurations award less than the matched human adjudicator, including some classified as employee-favoring or middle in employment recognition. Positional tendencies in status determination and monetary relief therefore need to be evaluated separately.

The practical concern is cumulative rather than limited to dramatic failures in individual cases. If similar models are repeatedly used in legal consultation, dispute screening, settlement assessment, or adjudicative support, even a modest positional tendency may impose costs on the same side across many decisions. These effects could remain obscured by acceptable aggregate accuracy while gradually shifting the effective boundary of labor-law protection. Our experiments do not establish such deployment effects, but motivate reporting positional error rates and monetary outcomes alongside aggregate agreement.

Adjudex-Labor makes these patterns visible within Chinese employment disputes, where human decisions serve as institutional rather than normative references. Future work should test whether positional differences reflect evidentiary parsing or unequal weighting of competing claims. Before deployment, legal LLMs should remain subject to position-sensitive evaluation, human review, and meaningful contestation. Legal AI cannot be positionless once its errors distribute protection and risk; its position must therefore be visible, contestable, and institutionally defensible.

AI USE STATEMENT

Generative AI tools were used for research ideation and to polish the language and presentation of the manuscript. All research questions, experimental designs, analyses, interpretations, and final written content were reviewed and approved by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This study evaluates positional tendencies in LLM decisions on real labor disputes. Human arbitral and judicial decisions are used as empirical institutional references rather than as guarantees of legal correctness, neutrality, or fairness. The benchmark is intended for research and evaluation of AI adjudication, not for determining individual legal entitlements or replacing professional legal judgment. Because the study relies on real judicial materials, any released data or artifacts should also comply with applicable source-access, privacy, and data-use requirements.

REPRODUCIBILITY STATEMENT

The paper specifies the experimental conditions, metric definitions, arbitration–judgment reconstruction procedure, and case-resampling approach. The appendix summarizes output checks and the adjudication workflow. Any released data and code will follow the access, licensing, and privacy constraints of the underlying legal materials.

REFERENCES

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 386–397, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.37. URL https://aclanthology.org/2024.acl-short.37/.

Dovilė Barysė and Roee Sarel. Algorithms in the court: does it matter which part of the judicial decision-making is automated? *Artificial Intelligence and Law*, 32(1):117–146, 2024. doi: 10.1007/s10506-022-09343-6. URL https://doi.org/10.1007/s10506-022-09343-6.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines. arXiv preprint arXiv:1707.00391, 2017. URL https://arxiv.org/abs/1707.00391.

Benjamin M. Chen, Yoan Hermstrüwer, Pascal Langenbach, Alexander Stremitzer, and Kevin Tobia. Mitigating the judicial human–AI fairness gap. *Journal of Legal Analysis*, 18(1):207–242, 2026a. doi: 10.1093/jla/laag007. URL https://academic.oup.com/jla/article/18/1/207/8741554.

Ziang Chen, Guannan Li, Fanlin Ji, Yipeng Kang, Jiaqi Li, Muhan Zhang, Yangtao Zhang, Li Tianjiao, Jiannan Wang, Xin Guo, Song-Chun Zhu, and Bin Ling. JurisBench: A deep benchmark for assessing large language models in professional legal practice. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35994–36018, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1666. URL https://aclanthology.org/2026.acl-long.1666/.

Thomas R. Cook, Sophia Kazinnik, Zach Modig, and Nathan M. Palmer. What do LLMs want? Finance and Economics Discussion Series 2026-006, Board of Governors of the Federal Reserve System, 2026. Working paper.

Sanmay Das. Local justice and the algorithmic allocation of scarce societal resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12250–12255, 2022. doi: 10.1609/aaai.v36i11.21486. URL https://ojs.aaai.org/index.php/AAAI/article/view/21486.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.452. URL https://aclanthology.org/2024.emnlp-main.452/.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi: 10.1162/coli_a_00524. URL https://aclanthology.org/2024.cl-3.8/.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279. Curran Associates, Inc., 2023. doi: 10.52202/075280-1915. URL https://arxiv.org/abs/2308.11462.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html.

Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2692–2701. PMLR, 2019. URL https://proceedings.mlr.press/v97/heidari19a.html.

Yiran Hu, Zongyue Xue, Haitao Li, Siyuan Zheng, Qingjing Chen, Shaochun Wang, Xihan Zhang, Ning Zheng, Yun Liu, Qingyao Ai, Yiqun Liu, Charles L. A. Clarke, and Weixing Shen. LLMs on trial: Evaluating judicial fairness for large language models. In *International Conference on Learning Representations*, 2026.

Aziz Z. Huq. Adjudicative AI and distributive justice. *Computer Law & Security Review*, 61: 106284, 2026. doi: 10.1016/j.clsr.2026.106284. URL https://doi.org/10.1016/j.clsr.2026.106284.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032. URL https://doi.org/10.1093/qje/qjx032.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. From fair predictions to just decisions? conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, 7:883999, 2022. doi: 10.3389/fsoc.2022.883999. URL https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2022.883999/full.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3150–3158. PMLR, 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé III. "you gotta be a doctor, lin" : An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7268–7287, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.413. URL https://aclanthology.org/2024.emnlp-main.413/.

NPC Standing Committee. Labor law of the people's republic of china, 1994. URL https://www.gov.cn/banshi/2005-05/25/content_905.htm. Article 1. Adopted July 5, 1994.

NPC Standing Committee. Labor contract law of the people's republic of china, 2007. URL https://www.npc.gov.cn/zgrdw/englishnpc/Law/2009-02/20/content_1471106.htm. Article 1. Adopted June 29, 2007.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Eric A. Posner and Shivam Saran. Judge AI: A case-study of large language models as judges. *Journal of Law & Empirical Analysis*, 3(1):179–203, 2026. doi: 10.1177/2755323X261433614. URL https://doi.org/10.1177/2755323X261433614.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9. Association for Computing Machinery, 2021. doi: 10.1145/3465416.3483305. URL https://doi.org/10.1145/3465416.3483305.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1144–1152. PMLR, 2021. URL https://proceedings.mlr.press/v130/wen21a.html.

Huiyuan Xie, Felix Steffek, Joana De Faria, Christine Carter, and Jonathan Rutherford. The CLC-UKET dataset: Benchmarking case outcome prediction for the UK employment tribunal. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pp. 81–96, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nllp-1.7. URL https://aclanthology.org/2024.nllp-1.7/.

Yixuan Xie, Yang He, Xiaoyu Yang, Xu Gai, and Pan Hui. LLM safety in judicial AI: A stress test of social media influence on real-world judgments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46):39468–39476, 2026. doi: 10.1609/aaai.v40i46.41297.

A Case Retrieval, Screening, and Validation

A.1 Detailed Case Retrieval and Screening

The benchmark was constructed from an initial pool of approximately 15,000 judicial documents through a four-stage screening procedure. The resulting analysis set contains 804 cases with determinate human arbitration and judicial outcomes.

1. **Initial retrieval and deduplication.** Candidate decisions were retrieved from China Judgments Online and PKULaw using the cause-of-action categories of civil cases, labor disputes, labor contract disputes, and disputes over the confirmation of employment relationships. The initial retrieval yielded approximately 15,000 judicial documents. Because the two databases partially overlap, duplicate records were identified using available case metadata and document content. Repeated records and documents lacking substantive decision text were removed before further screening.
2. **Substantive-dispute screening.** We examined whether the existence of an employment relationship was a substantive issue that the court was required to resolve. A case was retained only when (i) the parties disputed the nature of their relationship and (ii) the court made a determinate finding on whether an employment relationship existed. We excluded cases in which employment status was uncontested, appeared only as background information, or was irrelevant to the disposition. Procedural rulings concerning jurisdiction, withdrawal, retrial, preservation, enforcement, or similar matters were also excluded.
3. **Evidentiary-structure screening.** Each remaining decision was required to contain (i) an identifiable account of the plaintiff's claims and supporting allegations, (ii) an identifiable response from the defendant, and (iii) a separately identifiable section in which the court reviewed the evidence submitted by the parties. A brief reference to evidence or a simple list of submitted materials was insufficient. We excluded decisions that omitted either party's position, provided only a factual summary, or lacked an identifiable judicial review of evidence.
4. **Final manual verification and labeling.** The remaining decisions were manually reviewed to verify (i) the identities of the parties to the alleged employment relationship, (ii) the substantive nature of the dispute, (iii) the completeness of the required document components, (iv) the recorded arbitration outcome, and (v) the court's final relationship determination. Binary human-arbitration and judicial reference labels were assigned only when the corresponding institution explicitly found that an employment relationship either existed or did not exist. Decisions with missing, ambiguous, conditional, or indeterminate outcomes were excluded.

The final benchmark contains 804 cases. Each case includes a determinate Human Arbitrator outcome and Human Judge outcome, enabling both stage-specific evaluation and linked arbitration–judgment comparisons.

A.2 Prompt, Output Extraction, and Validation

For the employment task, the Chinese prompt required exactly one of two outcomes—A, an employment relationship exists, or B, no employment relationship exists—in a JSON object containing `judgment`, `reason`, and three `key_facts`. Only `judgment` entered the quantitative analyses; the other fields were retained for auditing. The Arbitrator condition omitted the recorded human arbitration outcome. The Judge conditions differed only in whether the supplied outcome came from the Human Arbitrator or the LLM Arbitrator; neither included the Human Judge's reasoning or disposition. Outputs were aligned by case, configuration, and condition, and pipeline totals were checked against case-level records.

The exploratory 50-case monetary sample requires positive quantified awards at both human stages and was selected after preliminary analysis. Structured cash totals were checked against payment clauses; substantive employee recoveries were summed, while litigation fees and delayed-performance interest were excluded. Two unquantifiable LLM Arbitrator outputs were omitted only from numerical comparisons. This was not an independent double-human review.

B ADJUDICATION WORKFLOW

Figure 5 distinguishes two comparisons. Replacing human with LLM arbitration holds the downstream Judge configuration fixed and changes its preceding arbitration information. Comparing independent arbitration with subsequent judgment additionally changes the institutional role and available information. These contrasts therefore address different aspects of positional behavior.

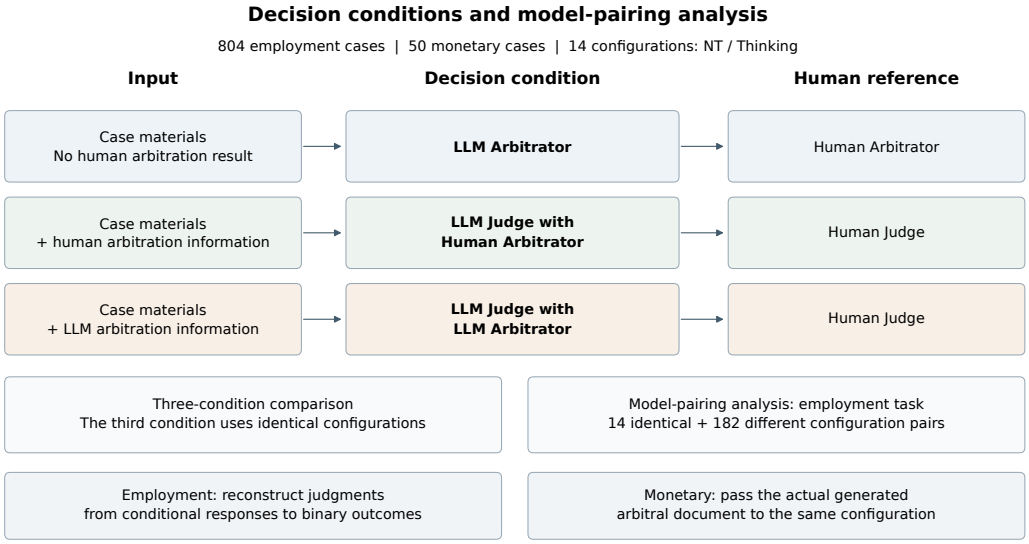


Figure 5: Decision conditions, stage-specific human references, and model pairing. The displayed references apply to the primary comparisons; supplementary employment analyses also evaluate LLM Arbitrator against Human Judge.

For the 804-case employment task, only the binary arbitration outcome links the stages. Each final judgment is selected from the Judge's recorded responses to the observed or reversed human arbitration outcome, according to the independent LLM Arbitrator decision (Section 3.2). The 196 pairings include fourteen same-configuration chains and 182 different-configuration chains. This reconstruction tests responses to preceding outcomes, not the transmission of generated legal arguments.

The 50-case monetary experiment instead passes the generated arbitral document to the same configuration acting as Judge and compares this condition with human arbitral input. Both Judge conditions exclude the Human Judge's reasoning and final disposition. The monetary analysis thus examines a document-based sequence, with awards evaluated against the human decision at the corresponding stage (Section 3.4).

C SELECTED SUPPLEMENTARY RESULTS

C.1 INDEPENDENT ARBITRATION UNDER TWO HUMAN REFERENCES

The same independent LLM Arbitrator outputs can be evaluated against either the recorded Human Arbitrator or the Human Judge (Figure 6). Relative to Human Arbitrator, 11 of the 14 configurations are employer-favoring, one falls within the middle interval, and two are employee-favoring. Relative to Human Judge, the corresponding counts are ten, one, and three. Twelve configurations retain the same category across the two references. Claude Opus 4.8 Thinking changes from middle to employee-favoring, whereas Kimi K3 NT changes from employer-favoring to middle. The employer-favoring majority is therefore not specific to the arbitral reference, although individual category assignments can depend on the institutional comparator.

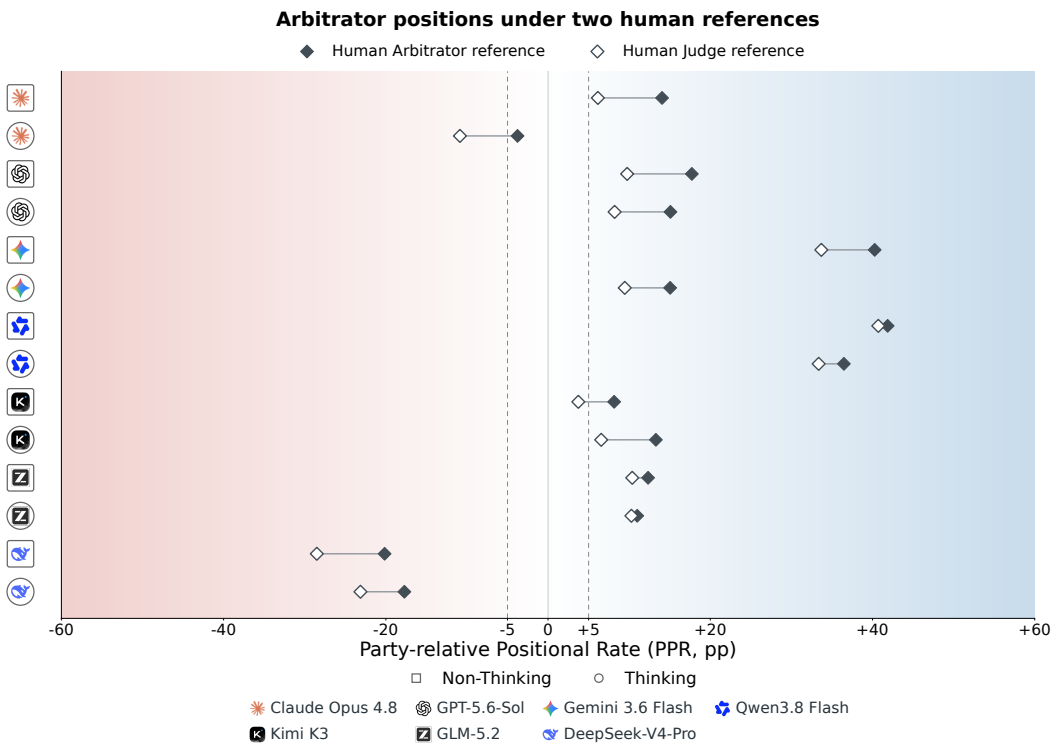


Figure 6: PPR of the same independent LLM Arbitrator outputs under two human references. Connected symbols identify the same model configuration, not a temporal transition. Shading denotes the employee-favoring, middle, and employer-favoring regions; dashed lines mark the ± 5 pp thresholds.

C.2 Configuration-Level Results for Same-Model Chains

Table 3 gives the exact same-configuration results. Relative to LLM Arbitrator, LLM Judge with LLM Arbitrator changes mean DDR by -1.41 pp (95% CI $[-2.55, -0.29]$) and mean PPR by $+2.68$ pp ($[+0.36, +5.01]$). Relative to LLM Judge with Human Arbitrator, the changes are $+9.53$ pp ($[+7.81, +11.27]$) and $+8.53$ pp ($[+5.09, +11.96]$), respectively. The intervals use the joint case-level resampling in Section 3.3. All conditions use Human Judge as reference; categories use unrounded PPR and the ± 5 pp thresholds.

Table 3: Configuration-level DDR (%) and PPR (pp) under the three same-model conditions. All values use Human Judge as the reference on 804 cases.

Model	Mode	LLM Arbitrator		LLM Judge with Human Arbitrator		LLM Judge with LLM Arbitrator	
		DDR	PPR	DDR	PPR	DDR	PPR
Claude Opus 4.8	NT	36.94	+6.14	24.25	+9.02	33.58	+17.58
Claude Opus 4.8	Thinking	37.94	-10.86	28.98	-10.24	37.06	-1.09
GPT-5.6-Sol	NT	31.59	+9.75	22.51	+4.66	30.85	+9.80
GPT-5.6-Sol	Thinking	32.34	+8.21	22.51	+1.67	30.97	+8.55
Gemini 3.6 Flash	NT	38.31	+33.69	28.36	+21.43	35.95	+37.09
Gemini 3.6 Flash	Thinking	35.82	+9.46	24.13	+3.80	34.33	+11.55
Qwen3.8 Flash	NT	44.78	+40.71	30.22	+32.51	43.03	+42.33
Qwen3.8 Flash	Thinking	43.03	+33.36	30.10	+24.30	41.54	+26.49
Kimi K3	NT	42.91	+3.73	30.97	-3.90	42.41	+8.75
Kimi K3	Thinking	38.18	+6.55	27.11	+4.58	36.94	+13.61
GLM-5.2	NT	44.03	+10.38	32.96	+0.44	42.79	+9.47
GLM-5.2	Thinking	45.52	+10.27	31.22	-1.93	42.16	+10.76
DeepSeek-V4-Pro	NT	40.92	-28.51	32.59	-27.18	40.92	-25.52
DeepSeek-V4-Pro	Thinking	38.81	-23.13	31.97	-31.37	38.81	-22.13