Evaluating LLMs as Poker Players: An Item-Response Theory and Q-Matrix Analysis of PokerBench

Abhinav Sattiraju
Stanford University
Stanford, CA 94305
asattira@stanford.edu

Abstract

PokerBench [Zhuang et al., 2025], a recent static benchmark for evaluating large language models on no-limit Texas Hold'em poker decision-making, reports a single accuracy score per model and implicitly treats "poker skill" as a unidimensional construct. We examine this assumption using standard item-response and factor-analysis techniques. We evaluate a panel of seven LLMs spanning capability tiers, model families, and reasoning vs. base-instruct variants on PokerBench, and fit the resulting binary response matrix with unidimensional IRT and a multidimensional K-factor model. Exploratory factor analysis does not recover stable extra dimensions: model selection prefers $K{=}1$. We then turn to a confirmatory Q-matrix family that imposes externally-specified item structure (GTO action class, item phase, a rule-based skill heuristic, and an LLM-labeled skill taxonomy), and find that every Q-matrix variant explains held-out response cells substantially better than the single-ability baseline; a joint Action$\times$Skill partition is the strongest. Permutation tests support the claim that the labels carry signal beyond what random partitions of the same shape achieve, and an LLM-free heuristic Q-matrix fits at least as well as the LLM-labeled one. The models rank differently across these axes, which means a single PokerBench accuracy score hides meaningful differences in how they succeed and fail.

1 Introduction

Poker is an important domain for AI: it is a clean example of decision-making under uncertainty, hidden information, and game-theoretic reasoning — capabilities relevant to negotiation and high-stakes-strategy applications of large language models (LLMs). A growing body of recent work applies LLMs to poker, ranging from simulated head-to-head play against solver baselines [Lin et al., 2026, Provost et al., 2026] to static decision benchmarks. PokerBench [Zhuang et al., 2025] pairs 11,000 no-limit hold'em scenarios with game-theoretic-optimal target actions and operationalizes poker skill as a single aggregate accuracy per model. The implicit assumption is that all 11,000 scenarios probe one underlying construct, so that a model with 60% accuracy is uniformly stronger than a model with 55%. The same assumption underwrites aggregate accuracy on MMLU, GSM8K, and most modern LLM leaderboards.

Poker decision-making, however, involves distinct sub-skills: pot-odds calculation, opponent-range reasoning, bet sizing, position-aware play, and multi-street planning. If LLM performance on these sub-skills differs, a single aggregate score conflates the differences and discards information that benchmark designers, model developers, and downstream users of LLM agents need.

This paper asks whether PokerBench's aggregate score is reliable and valid as a unidimensional measure of poker skill. If not, we want to characterize the latent structure and connect it to interpretable poker features. We construct a model-by-item response matrix by evaluating a seven-LLM panel

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

on PokerBench, then fit a unidimensional 2PL item-response model as a baseline (a single ability per LLM, with uncertainty), and a multidimensional K-factor logistic model that lets each LLM carry a K-dimensional ability profile. We select K with the Bayesian Information Criterion and held-out cross-validated log-likelihood (CV LL). If $K{=}1$ wins, the unidimensional view is supported; if $K > 1$, each LLM gets a skill profile rather than a single score.

We find that unconstrained K-factor exploration prefers $K{=}1$ on CV LL. The data do not autonomously reveal stable additional dimensions in this small-subject regime. Rather than concluding that PokerBench is unidimensional, we turn to a confirmatory Q-matrix family that imposes candidate multidimensional structure from external sources (GTO action class, item phase, a rule-based skill heuristic, and an LLM-derived skill taxonomy) and tests whether the data prefer it to the single-ability fit. Every Q-matrix variant explains the data substantially better than $K{=}1$ on held-out log-likelihood, including the two variants that do not involve an LLM at all. A joint Action$\times$Skill partition is the strongest. The models rank differently across these axes, which means a single PokerBench accuracy score hides meaningful differences in how they succeed and fail.

Our contributions are: (i) A unidimensional IRT and multidimensional K-factor analysis of PokerBench on a seven-LLM panel, with held-out model selection. (ii) A constrained Q-matrix family that imposes externally-known item structure (action class, phase, rule-based skill, LLM-labeled skill, and their joint refinement), with held-out comparison to the K-factor baseline. (iii) Permutation-based validity tests — unrestricted and within-stratum — and an LLM-free heuristic Q-matrix as convergent evidence: the strongest fits come from labels that do not use an LLM at all. We interpret the results as follows: we do not view the 8-skill taxonomy as the definitive decomposition of poker skill. The claim is narrower — that pre-specified item structure, including the most theory-free choice (the GTO action class), captures variation in model behavior that a scalar accuracy score averages away.

2 Related work

Prior relevant work falls into three groups: measurement models applied to LLM evaluation, emerging LLM poker benchmarks, and construct-validity critiques of aggregate scores.

Measurement models for LLM evaluation. Item-response theory has been applied to LLM evaluation to separate model ability from item-level differences. Polo et al. [2024] use IRT to construct compressed benchmarks (*tinyMMLU*, *tinyAlpaca*) that reproduce benchmark-level conclusions with a fraction of the items. Lalor and Rodriguez [2023] provide the `py-irt` library used in preliminary fits here. Zhou et al. [2026] extend the framework to a richer item-parameter setting and analyze eleven mainstream LLM benchmarks, finding "varied shortcomings in measurement quality" pervasive across the field. Truong et al. [2025] integrate a Rasch-style model into HELM and report stable item and ability estimates across more than 170 models. These projects establish IRT-style approaches as an active and increasingly mature methodology for LLM evaluation; what remains underexplored is their application to strategic, game-theoretic benchmarks.

LLM poker benchmarks. Zhuang et al. [2025] introduce PokerBench, the static item-level benchmark used here: 11,000 GTO-labeled poker scenarios, scored against an LLM's predicted action. Their headline result is that an LLM's aggregate score on PokerBench correlates with its head-to-head win rate, establishing partial external validity but leaving construct validity untested. Lin et al. [2026] re-examine LLM poker via simulated play in Leduc and Limit Hold'em and identify three named failure modes: heuristic reasoning, factual misunderstanding, and a knowing-doing gap where stated reasoning diverges from the action. Provost et al. [2026] pit poker agents (including frontier LLMs) against the GTO Wizard AI in heads-up no-limit play with AIVAT variance reduction and find all evaluated LLMs far below the baseline. None of these poker works applies IRT or multidimensional latent-trait modeling to audit the measurement properties of their response data.

Construct validity and benchmark critique. Construct-validity concerns have been raised across reasoning benchmarks more broadly. Mirzadeh et al. [2024] show that GSM8K performance can reflect brittle pattern matching rather than robust mathematical reasoning, with substantial accuracy drops under symbolic and numeric perturbations. The measurement-science perspective [Stanford CS321M, 2026] emphasizes that aggregate scores often conflate distinct sub-skills and that validity must be argued rather than assumed.

Q-matrix cognitive diagnostic models. The constrained multidimensional fits we use draw on the cognitive-diagnostic-modeling tradition. Tatsuoka [1983] introduced the Q-matrix as a binary $N \times K$ indicator of which skills each item requires; DINA and related CDMs [de la Torre, 2009, Rupp et al., 2010] treat the Q-matrix as a constraint on the latent structure rather than a free hyperparameter. We adopt the constraint logic of CDMs but stay inside the logistic-link family, so the Q-matrix fits are directly comparable to the unconstrained factor baseline on the same held-out criterion.

3 Methods

3.1 Data, panel, scoring

We use the full PokerBench test split: N=11,000 scenarios spanning preflop and postflop play with GTO solver actions as ground truth. The model panel (Table 1) covers seven LLMs spanning capability tiers, model families, and reasoning vs. base-instruct training. Each model is queried zero-shot at temperature 0 with the PokerBench-supplied instruction used verbatim as the user message (no system prompt, no chain-of-thought scaffold beyond what each model's API exposes natively). We report two scoring rules: *action accuracy* (AA, agreement with the GTO action class regardless of sizing) and *exact match* (EM, agreement on both action class and sizing bucket). AA is the primary response signal for the IRT and factor fits; EM appears in Table 1 as a complementary measure that penalizes sizing mistakes on postflop bets.

Table 1: Model panel and aggregate accuracies on the 11,000-item PokerBench test split. **Reasoning** indicates an explicit chain-of-thought or extended-thinking pathway at inference time.

Model	Provider	Reasoning?	AA	EM
gpt-5-mini (medium effort)	OpenAI	Yes	0.599	0.526
Claude Haiku 4.5	Anthropic	No	0.593	0.539
Llama-3.3-70B-Instruct	Together	No	0.515	0.467
Claude Sonnet 4.6 (2000-tok budget)	Anthropic	Yes	0.491	0.372
Qwen3-235B-A22B-Instruct	Together	No	0.478	0.396
Gemini 2.5 Flash-Lite	Google	No	0.438	0.382
Qwen2.5-7B-Instruct	Together	No	0.270	0.265

3.2 Unidimensional IRT and the K-factor logistic baseline

For LLM i and item j, the 2PL item-response model is

$$P(Y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) \;=\; \sigma(\alpha_j(\theta_i - \beta_j)), \tag{1}$$

where θ_i is model ability, β_j is item difficulty, and α_j is item discrimination. The K-factor multidimensional generalization [Reckase, 2009] replaces scalar ability with a K-dimensional profile $\mathbf{u}_i \in \mathbb{R}^K$ and item loading $\mathbf{v}_j \in \mathbb{R}^K$:

$$P(Y_{ij} = 1 \mid \mathbf{u}_i, \mathbf{v}_j, z_j) \;=\; \sigma(\mathbf{u}_i^\top \mathbf{v}_j - z_j). \tag{2}$$

We fit Equation 2 for $K \in \{1, 2, \ldots, 15\}$ and select K by 5-fold cross-validated held-out log-likelihood over response cells. The 2PL baseline corresponds to a special case of K=1.

3.3 Constrained Q-matrix model

Given an externally-known item-to-class map $k : \{1, \ldots, N\} \to \{1, \ldots, K\}$ encoded as a single-loading Q-matrix, we constrain Equation 2 so that item j depends only on ability column $k(j)$:

$$P(Y_{ij}{=}1) \;=\; \sigma\big(\, u_{i,\,k(j)} \cdot v_j \;-\; z_j \,\big), \tag{3}$$

with $u_{i,\cdot} \in \mathbb{R}^K$ a per-subject ability vector, $v_j \in \mathbb{R}$ a scalar item loading, and $z_j \in \mathbb{R}$ an item difficulty offset. The fit family and held-out criterion match Section 3.2. We consider five Q-matrix variants of increasing item structure:

Phase (K=2)**.**
Preflop vs. postflop split. Coarsest possible partition, derived from the PokerBench item structure.

Action GTO (K=4).
$k(j) \in \{\text{fold, check, call, bet}\}$ from the PokerBench-supplied GTO label.

Heuristic skill (K=8).
A priority-ordered nine-rule decision list over (game phase, GTO action, hand-strength class, draw indicators) produces eight poker-skill categories: `preflop_open_3bet_4bet`, `value_betting`, `bluff_betting`, `bluff_catching`, `pot_odds_drawing`, `folding_discipline`, `pot_control_check`, and `position_aware_continuation`. Fully deterministic and derived from the PokerBench item structure.

LLM-derived skill (K=8).
Same eight categories assigned by `gemini-2.5-flash-lite` with a zero-shot prompt that classifies each scenario into one of the categories.

Joint Action $\times$ Skill (K=11).
Item-wise cross of Action GTO and Heuristic skill, yielding eleven non-empty cells; tests whether skill labels carry signal beyond action class alone.

Fitting. We fit all models with stochastic variational inference using a Delta variational guide. With a Delta guide, the variational posterior collapses to a point estimate, so the fit is equivalent to MAP estimation under weakly informative priors on abilities, item difficulties, and item loadings/discriminations. This gives a scalable way to fit the large IRT-style response matrix, but it does not provide posterior uncertainty by itself. We therefore estimate uncertainty with item-bootstrap resampling (Section 3.4).

3.4 Validity checks

Permutation tests. We draw N=20 random permutations of the skill labels preserving per-class frequencies, refit, and record the distribution of BIC-based fit gains under the shuffled labels. We also run a stricter restricted permutation that shuffles labels within (GTO action, phase) strata, isolating within-stratum skill signal from action and phase information.

Bootstrap, reliability, and uncertainty. We quantify per-cell uncertainty in the recovered Q-matrix ability profiles with N=80 item-bootstrap resamples, computing 95% percentile intervals per (subject, class) cell. We also check ranking stability by comparing per-model AA on a 5K subset to the full 11K set ($\rho = 0.943$, $p = 0.005$). Because the seven-model panel is small, we treat individual skill-axis rankings as uncertain and emphasize aggregate patterns across axes.

4 Results

4.1 Exploratory K-factor analysis prefers K=1

We first fit 1PL/2PL IRT and the unconstrained K-factor logistic model for $K \in \{1, \ldots, 15\}$ on the $7 \times 11{,}000$ response matrix and select K by 5-fold cross-validated held-out log-likelihood. The single-factor model wins by a wide margin: adding free dimensions sharply hurts held-out performance and the K-factor model overfits as K grows. Exploratory MIRT alone would support a unidimensional interpretation of PokerBench.

4.2 Confirmatory Q-matrix variants win, and joint Action$\times$Skill is strongest

Rather than concluding that PokerBench is unidimensional, we ask whether *externally specified* item structure can recover multidimensional signal that the data cannot discover unaided. We test the five Q-matrix variants from Section 3.3 on the same held-out CV LL criterion used for K-factor selection. Results in Table 2 and Figure 1.

Every Q-matrix variant explains held-out response cells substantially better than the K=1 baseline, with held-out log-likelihood gains spanning $+0.09$ to $+0.34$ per cell. The two strongest single-axis variants are LLM-free: Heuristic skill and Action GTO. The LLM-derived Gemini skill labels are the weakest non-Phase variant, beating K=1 but losing to both LLM-free alternatives.

Joint Action×Skill. A skeptical reading of the single-axis results is that the heuristic skill taxonomy is built on top of GTO action, so we may have rediscovered the action class with a different name. To address this we fit a *joint Action×Skill* Q-matrix that assigns each item to its (action, skill) combination, producing 11 non-empty cells: seven preserve the heuristic skill 1:1 with action, and four refine the preflop `preflop_open_3bet_4bet` class by action. The joint variant is the strongest fit of any tested (Table 2, bottom row): held-out gain $+0.336$ nats/cell vs. K=1. Compared to the Action GTO baseline, the postflop value/bluff refinement adds $+0.053$ per cell: the skill labels carry held-out signal beyond action class. Compared to the Heuristic-skill baseline, splitting preflop by action adds $+0.027$ per cell: action carries signal beyond the skill taxonomy. Neither single axis exhausts the structure.

Table 2: Q-matrix multidimensional variants vs. the unconstrained $K{=}1$ baseline. **CV LL/cell** is mean 5-fold cross-validated held-out log-likelihood per response cell; higher is better. **ΔCV LL** reports the gain over $K{=}1$. No-LLM variants use only the GTO action label and / or deterministic rules over PokerBench features.

Variant	K	CV LL/cell	ΔCV LL	Labels
$K{=}1$ free factor	1	-0.8834	—	baseline
Q-matrix Phase	2	-0.7914	$+0.092$	no LLM
Q-matrix Skill Gemini	8	-0.6059	$+0.278$	LLM-labeled
Q-matrix Action GTO	4	-0.6005	$+0.283$	no LLM
Q-matrix Skill Heuristic	8	-0.5737	$+0.310$	no LLM
Q-matrix Joint Action×Skill	**11**	$\mathbf{-0.5472}$	$\mathbf{+0.336}$	no LLM

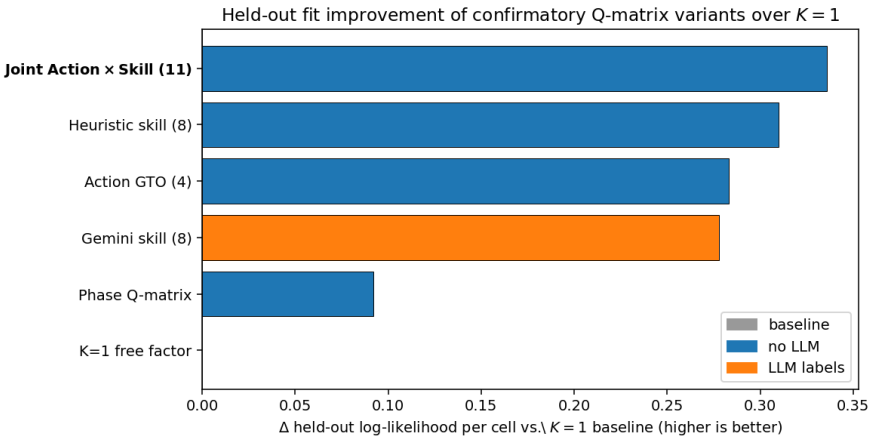


Figure 1: Held-out log-likelihood improvement per response cell vs. the $K{=}1$ free-factor baseline for all five confirmatory Q-matrices. The joint Action×Skill variant is the strongest fit; LLM-free variants (blue) span the top of the figure.

4.3 Permutation tests for label informativeness

Two complementary permutation tests probe whether the Q-matrix gains come from labels carrying real information or from the structural flexibility of the constrained model. We report the unrestricted version on the Gemini-labeled variant and the restricted version on the heuristic-labeled variant; both reach the same qualitative conclusion.

Unrestricted shuffles (Gemini labels) randomize the per-item skill assignment globally, preserving only the class size of each skill. Across $N{=}20$ shuffles, the shuffled-label fit improvement concentrates tightly at roughly half the real labels' value, reflecting the model-class flexibility that any eight-way partition enjoys. The real Gemini labels outperform every shuffled labeling we tested ($p < 1/20$).

Restricted shuffles (heuristic labels) randomize within (GTO action, phase) strata, so the shuffled fits retain the same action- and phase-aligned information as the real labels. The real heuristic labels still

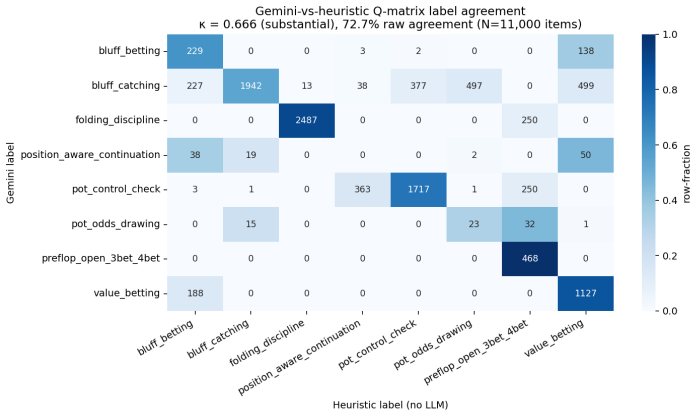


Figure 2: Heuristic vs. Gemini skill-label agreement (Cohen's $\kappa = 0.666$, 72.7% raw agreement). Rows: Gemini label. Columns: heuristic label. The two independent labelings substantially agree.

outperform every restricted shuffle we tested ($p < 1/20$), though by a smaller margin: most of the heuristic's fit advantage does come from action- and phase-conditioning, but a residual within-stratum signal remains significant.

Convergent validity from an LLM-free labeling. The heuristic Q-matrix — nine deterministic rules over PokerBench features, no LLM — substantially agrees with the Gemini labels: 72.7% raw label agreement, Cohen's $\kappa = 0.666$ across all 11,000 items (Figure 2). The two labeling methods reach a comparable held-out fit (Table 2), and the LLM-free version is actually the stronger of the two. This makes the LLM-labeling concern less central: the strongest fits come from labels that do not use an LLM.

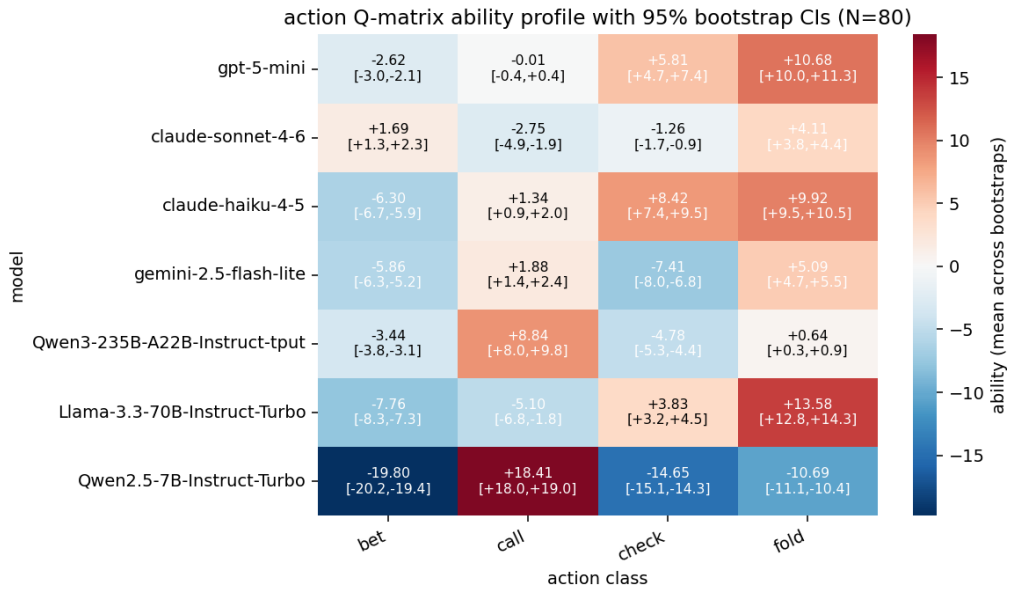
4.4 Bootstrap uncertainty and the leaderboard fallacy

Figure 3 shows the recovered ability profiles for the Action GTO and Heuristic skill Q-matrices, with 95% bootstrap percentile intervals (item-resampled). Within each column, the within-class pairwise comparisons between the seven models are CI-disjoint for 79 of 84 pairs (94%) in the Action GTO panel and 140 of 168 pairs (83%) in the Heuristic panel: the within-column rankings are statistically distinguishable. Across columns, the panel reorders substantially. The clearest illustration of this "leaderboard fallacy" is Qwen-2.5-7B, which is the highest cell anywhere in the Action GTO matrix on `call` (an indiscriminate-caller artifact) and the lowest cell on every other action class.

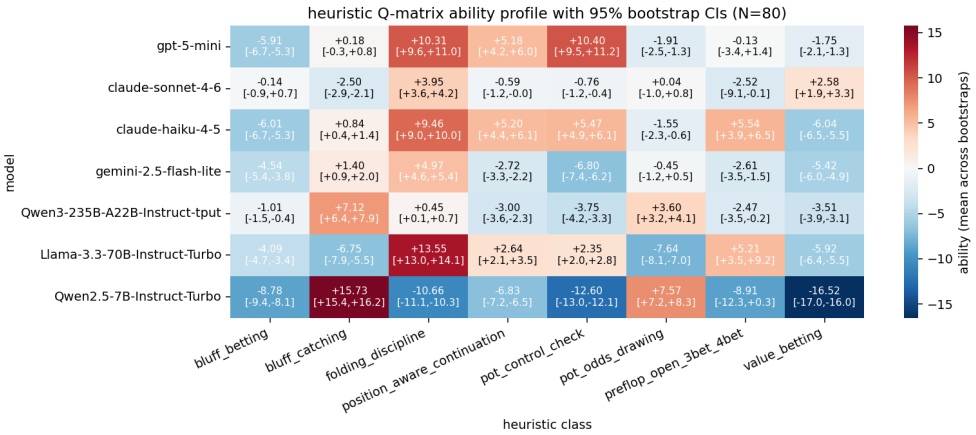
5 Discussion

What the evidence supports. Exploratory K-factor analysis does not surface stable additional dimensions in PokerBench at this panel size, but *confirmatory* Q-matrix structure does. Five distinct Q-matrix encodings — including the most theory-free choice (the GTO action class), a rule-based skill heuristic with no LLM in the labeling loop, and a joint Action×Skill refinement — explain held-out response cells substantially better than the single-ability baseline. The joint variant beats either single axis, ruling out the strongest version of the "you only found action effects" reading. Permutation tests support the position that the labels carry signal beyond what random partitions of the same shape achieve, and the LLM-free heuristic Q-matrix reaches a stronger fit than the LLM-derived one. We do not view the 8-skill taxonomy as a definitive decomposition of poker skill; it is one choice among many, and the joint variant simply combines two complementary external axes that the data prefer to either alone.

Limitations. *Panel size.* Seven LLMs is small relative to traditional IRT contexts. Individual skill-axis rankings carry meaningful per-cell uncertainty (visible in the bootstrap CIs in Figure 3); the strongest claim is the aggregate one, not any individual cell in the heatmap. Scaling the panel to dozens of LLMs is the most important next step. *Skill taxonomy is chosen, not discovered.* Our skill



(a) Action GTO Q-matrix ability profile (K=4).



(b) Heuristic skill Q-matrix ability profile (K=8).

Figure 3: Recovered ability profiles with 95% bootstrap percentile intervals (N=80 item resamples). Each cell shows the mean and the $[2.5\%, 97.5\%]$ interval; sign convention is positive = above-average ability on items of that class. The leaderboard fallacy is visible in (a): Qwen-2.5-7B is the highest cell in the entire matrix on `call` because it tends to call indiscriminately, yet is the worst on every other action class. The heuristic skill view (b) tells a similar story with a finer-grained partition.

categories are useful proxies, but they are not a definitive map of poker ability. Future work should validate them with expert poker labels or a richer ontology that includes range reasoning, board texture, and bet sizing. *Binary scoring loses EV nuance.* PokerBench publishes the GTO action class but not the per-action EV; reconstructing solver EV per item is out of scope. A graded IRT extension using EV-regret as the per-cell outcome would replace our Bernoulli likelihood with an ordered or continuous one and is a natural next step. *Feature regression pivoted.* The proposal anticipated a Lasso regression of free-factor item loadings on PokerBench features to interpret recovered latent dimensions. Because the unconstrained K-factor analysis preferred K=1, there were no extra latent dimensions to interpret in that way; we pivoted to the Q-matrix family, which imposes interpretation directly via the item-to-class map.

6 Conclusion

An unconstrained K-factor exploration over $K \in \{1, \ldots, 15\}$ does not surface stable extra dimensions in PokerBench at this panel size. Confirmatory multidimensional structure, however, is strongly supported: four single-axis Q-matrix encodings explain held-out response cells substantially better than a single ability axis, and a joint Action$\times$Skill refinement beats either single axis alone. The seven-model panel reorders along the structured axes, so a scalar accuracy score on this benchmark picks one weighting over skill axes the consumer may not endorse.

We do not view this taxonomy as the definitive decomposition of poker skill. The narrower claim is that pre-specified item structure — even structure as theory-free as the GTO action class — captures behaviorally meaningful variation that a single accuracy score averages away. Future work should scale the panel to dozens of LLMs, extend the response model to graded EV-regret rather than binary AA, validate the skill taxonomy against expert poker labels, and explore a multi-loading Q-matrix that lets each item draw on multiple skills.

References

Jimmy de la Torre. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130, 2009.

John P. Lalor and Pedro Rodriguez. py-irt: A scalable item response theory library for Python. *INFORMS Journal on Computing*, 35(1):5–13, 2023. arXiv:2203.01282.

Minhua Lin, Enyan Dai, Hui Liu, et al. How far are LLMs from professional poker players? revisiting game-theoretic reasoning with agentic tool use, 2026. arXiv:2602.00528.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. arXiv:2410.05229.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples, 2024. arXiv:2402.14992.

Provost et al. GTO wizard benchmark, 2026. arXiv:2603.23660.

Mark D. Reckase. *Multidimensional Item Response Theory*. Springer, 2009.

André A. Rupp, Jonathan Templin, and Robert A. Henson. *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press, 2010.

Stanford CS321M. AI measurement science: Lecture notes. Stanford University, 2026.

Kikumi K. Tatsuoka. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354, 1983.

Sang Truong et al. Reliable and efficient amortized model-based evaluation. Stanford CRFM, 2025.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. Oral; arXiv:2505.15055.

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. PokerBench: Training large language models to become professional poker players. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. arXiv:2501.08328.

Disclosures

Code and data. The code and data supporting this work are available at: `https://github.com/asattiraju13/poker-llm-irt`.

Plagiarism, bias, and accuracy (150 words). *How did you identify and address issues such as plagiarism (both in terms of text and in terms of non-attribution of ideas to scholars), bias, and inaccuracies?*

I developed the paper's motivation, connections to related work, and interpretation of results myself, and I utilized citations to distinguish my contributions from prior work. With respect to ideas that were not mine, I made sure to attribute them to appropriate sources in the references: Item Response Theory framework, Q matrix, PokerBench dataset, and other related work. I also conducted my own research to find ideas discussed beyond class such as the Q-matrix approach to IRT. I addressed possible bias and inaccuracies by explicitly reporting limitations. Specifically, the created skill taxonomy is heuristic and may not reflect the true structure of poker skill/ability, the LLM skill-labeling method may contain annotator bias from model training, and due to compute constraints, the model selection was limited, which also affects IRT estimates.

AI-tool use reflection (150 words). *How and why did you use AI tools, what was their impact on your learning, and how might you (or we) use them in the future?*

AI tooling supported my project, as I utilized it to help construct the IRT matrix and label items from PokerBench with skill categories. More importantly, I utilized AI IDE tools to draft code, such as module setups and library installations. This support helped reduce boilerplate work and helped me focus on the core research process: designing experiments, interpreting results, and composing a solid research paper. I also used AI chatbots to source relevant literature for understanding IRT topics beyond class and prior work conducted on poker benchmarks. LLMs also helped me revise grammar and phrasing in the paper to improve clarity and flow. In the future, I would use LLMs and AI tools within IDEs to help speed up research environment setup and boilerplate code creation so that I can focus on higher level yet detailed research tasks.

Impact statement (150 words). *What potential social, ethical, or environmental impacts could arise from this work? How might the methods, data, or findings be misused, and what steps (if any) could mitigate those risks? How have you ensured responsible attribution, data handling, and compliance with Stanford's standards of academic integrity? If the work involves human participants, sensitive data, or societal applications, what ethical considerations guided your approach?*

This project has relatively low high-stakes impact, since it is an evaluation audit instead of a deployed decision system. The main risk is misinterpretation: readers interpreting the results as advocating for LLM poker ability, which is not the key takeaway of the project. LLMs are far from good poker players, and the goal of this analysis was to analyze PokerBench as a measurement instrument and illustrate multidimensional structure in LLM poker behavior that is not captured by a single accuracy score. I have included a limitations section to address potential misinterpretations. Furthermore, all prior work is cited and data and model selections are publicly available. No human / sensitive data was used in this project.