CAPABILITY CALIBRATION IN ROLE-PLAYING LANGUAGE MODELS: A BENCHMARK AND METACOGNITIVE AGENTIC FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Research on role-playing large language models (LLMs) has largely emphasized fidelity to the identities, personalities, and behaviors of characters and personas. For professional roles across career stages, however, a key challenge remains underexplored: providing useful assistance within the expertise and decision authority supported by a persona's career history—a requirement we term career-grounded capability calibration. LLMs may draw on capabilities beyond those supported by the assigned profile, causing junior or unrelated personas to respond like experts. We introduce CAPROLEBENCH, which holds 355 workplace tasks fixed while varying career evidence across Senior, Mid, Junior, and Unrelated profiles to evaluate response fidelity and capability ordering. We further propose CAPROLEAGENT (CRA), a training-free, model-agnostic metacognitive framework that aligns career-supported capabilities with task requirements, enforces behavioral contracts, and audits overreach and omission. Across 14 open and proprietary LLMs, we observe systematic capability leakage under direct career-profile prompting. On a matched 100-task subset, CRA raises macro-average Ordering Consistency from 65.9% to 81.6% and lowers macro-average Overreach Rate from 32.0% to 10.9%, with both metrics improving for all 14 models. Profile-aware evaluation further shows higher Knowledge Fidelity and Behavior Coherence with fewer omissions, indicating improved role fidelity while preserving useful assistance. Evaluations by three LLM judges and persona-blind human raters corroborate the gains in capability ordering. The code is available at `https://anonymous.4open.science/r/CapRoleAgent`.

1 INTRODUCTION

Role-playing with large language models (LLMs) seeks to simulate characters and personas whose responses reflect their identities, experiences, and behavioral traits (He et al., 2025). Studies examine persona consistency across situations (Samuel et al., 2025), emotional fidelity (Feng et al., 2025), and persistence over extended interactions (Luz de Araujo et al., 2026). Professional role-playing across career stages remains less explored: the same workplace request may warrant different levels of execution scope, technical depth, and decision authority for junior and senior roles.

As LLM development favors stronger general capabilities, models may be more capable than the personas they simulate, while persona prompting does not reliably elicit the intended effects of expertise (De Araujo et al., 2025). A response may therefore be factually correct yet unfaithful to the role by expressing expertise or authority unsupported by the persona's background. We call this mismatch *capability leakage*. Figure 1 contrasts responses to the same task and unrelated profile: one gives a polished expert-level answer beyond that profile (capability leakage), while the other helps within supported capabilities. Prior work studies predefined role-ability limits (Lu et al., 2025) and student cognitive levels inferred from learning records (Wu et al., 2025). We instead consider professional capabilities inferred from career histories, which require interpreting how past experience supports a task. We formulate *career-grounded capability calibration*: providing useful assistance within the expertise and authority supported by career evidence. The goal is not deliberate error or blanket refusal, but completing supported work while qualifying or escalating unsupported parts.

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.

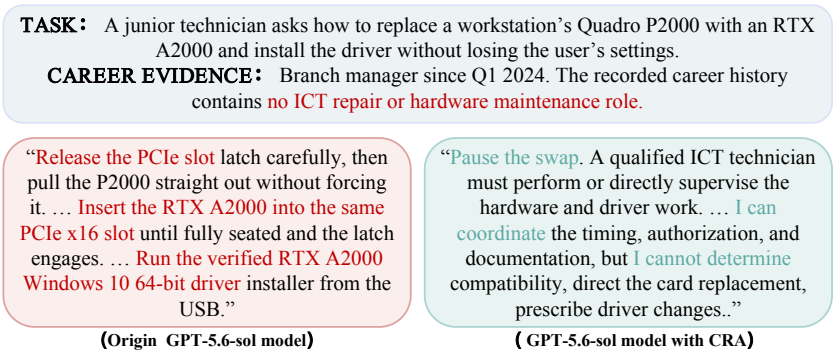


Figure 1: Capability leakage and career-grounded capability calibration. The left answer is polished but takes senior-level ownership; the right completes the analysis this profile supports.

To test whether LLMs adjust professional judgment to career evidence, we introduce CAPROLEBENCH: 355 workplace tasks spanning 355 occupations, each paired with Senior, Mid, Junior, and Unrelated profiles (1,420 instances). Holding each task fixed and hiding condition labels, we evaluate Career-Profile Prompting (CPP) across 14 LLMs. The results show that models often give technically strong answers that exceed the profile's supported expertise or authority. They struggle especially to distinguish experience levels within the same career.

To address this mismatch, we propose CAPROLEAGENT (CRA), a training-free, model-agnostic metacognitive framework that infers task-specific expertise and authority from career histories. It aligns career-supported capabilities with task requirements and sets a behavioral contract for execution scope, technical depth, and decision authority. A frozen backbone answers under this contract, with feedback-triggered revision correcting overreach or omitted supported work.

On the matched 100 tasks, CRA raises macro-average Ordering Consistency from 65.9% to 81.6% and lowers Overreach Rate from 32.0% to 10.9% relative to CPP. Profile-aware evaluation finds higher Knowledge Fidelity and Behavior Coherence with fewer omissions, showing better career alignment without sacrificing useful task completion.

Our contributions are threefold:

- We formulate career-grounded capability calibration and introduce CAPROLEBENCH, which varies career evidence while holding workplace tasks fixed to evaluate capability differentiation and response fidelity.
- We propose CAPROLEAGENT, a training-free, model-agnostic framework that aligns career-supported capabilities with task requirements and regulates generation through explicit behavioral constraints.
- Experiments across diverse LLMs demonstrate improved capability calibration and career-grounded response fidelity, supported by additional automated evaluation, human evaluation, and ablation studies.

2 RELATED WORK

2.1 LLM-BASED ROLE-PLAYING

LLM-based role-playing aims to simulate characters, individuals, or social roles while preserving their identities and backgrounds. Recent benchmarks evaluate role-playing through configurable roles, context-rich multi-turn interactions, and persona adherence across diverse environments (He et al., 2025; Shi et al., 2026). Beyond identity consistency, prior work studies emotional and discourse fidelity, persistence over extended interactions (Feng et al., 2025; Bhandari et al., 2025; Luz de Araujo et al., 2026), structured or evolving persona representations (Cai et al., 2026; Yirui et al., 2026), and knowledge or value boundaries (Zhang et al., 2025; Lai et al., 2026). Capability-oriented studies examine whether expertise attributes affect task performance, whether models re-

spect predefined role-ability limits (De Araujo et al., 2025; Lu et al., 2025), and whether educational agents reflect inferred cognitive states or appropriate teaching competence (Wu et al., 2025; Jiang et al., 2026). Existing studies largely use explicit persona attributes, predefined limits, or domain-specific cognitive states, leaving open whether role-play responses reflect the strength and relevance of career evidence. We address this gap with CAPROLEBENCH, which varies career evidence while holding tasks fixed to evaluate cross-profile capability differentiation and response fidelity.

2.2 ENHANCING ROLE-PLAYING

Role-playing methods broadly fall into model adaptation and inference-time control. Supervised fine-tuning and role-aware reasoning improve persona knowledge, professional behavior, and character-specific reasoning (Li et al., 2025; Cai et al., 2026; Tang et al., 2025), while reinforcement-learning approaches optimize role behavior through comparative or verifiable objectives (Ye et al., 2025; Wang et al., 2026b; Liao et al., 2026). Inference-time methods instead organize reasoning around persona information (Qin et al., 2025), derive task-conditioned cognitive prototypes (Wu et al., 2025), compile personas into structured state machines or decision trees (Peng et al., 2026a;b), retrieve bounded persona memories (Wang et al., 2026a), correct persona drift through feedback (Yirui et al., 2026), or control persona expression during decoding and through model activations (Liu et al., 2026; Feng et al., 2026). These methods improve persona knowledge, memory, reasoning, and behavioral consistency, but do not explicitly determine how much expertise, execution scope, and decision authority a career history supports for a specific task. We therefore propose CAPROLEAGENT, a training-free, model-agnostic framework that aligns career-supported capabilities with task requirements and regulates generation through explicit behavioral constraints.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND OVERVIEW

Let M_θ be an LLM, p a career profile, and x a workplace task. Under Career-Profile Prompting (CPP), the model directly generates $y = M_\theta(p, x)$, leaving task-specific capability assessment implicit in response generation.

We define the *career-grounded capability boundary* $\mathcal{B}(p, x)$ as the set of responses whose expressed expertise and authority are supported by p for x. It constrains execution scope, technical depth, and decision authority using supplied career evidence rather than unobserved latent ability.

Career-grounded capability calibration seeks the most useful response within this boundary:

$$y^\star = \arg \max_{y \in \mathcal{B}(p,x)} U(y, x), \tag{1}$$

where $U(y, x)$ denotes task utility. This favors completing supported work while qualifying or escalating unsupported parts. CAPROLEBENCH evaluates this objective under fixed tasks and varying career evidence, while CRA operationalizes it through explicit generation constraints.

3.2 CAPROLEBENCH

CAPROLEBENCH evaluates whether role-playing LLMs adapt their expressed capability to career evidence rather than merely solve a workplace task. It contains 355 tasks covering 355 distinct occupational targets, each paired with four career profiles, yielding 1,420 controlled task–profile instances. For each evaluated LLM, the protocol has two phases: controlled role-play simulation followed by dual-view evaluation of the generated responses (Figure 2).

Tasks and profiles. Workplace tasks are derived from O*NET 30.3 occupational task statements (National Center for O*NET Development, 2026) and expanded through LLM-assisted rewriting into self-contained scenarios with concrete inputs and user requests. Automatic occupation matching uses the official ESCO crosswalk. To reduce task difficulty as a confound, we screen persona-free answers from Qwen3-8B, one of our smallest evaluated open-weight backbones, using Qwen3-32B judgments and replace weak task selections.

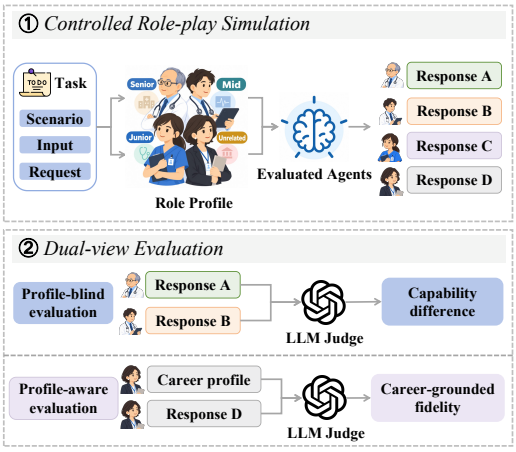


Figure 2: Overview of CapRoleBench.

Career profiles are constructed primarily from anonymized JobHop trajectories and standardized using ESCO occupation descriptions (**?**). For each task, three stages of the same trajectory are labeled Senior, Mid, and Junior by descending audited task-specific evidence support, not chronological seniority. An Unrelated profile is sampled from a different ESCO major group. Construction, screening, and audit details appear in Appendix A.

Phase 1: Controlled role-play simulation. For each task, the evaluated LLM generates one response per career profile under identical scenarios, inputs, requests, role-playing instructions, and response templates. Only career evidence varies, and condition labels remain hidden from the model during generation. The instructions require useful assistance within the supported scope rather than deliberate errors or default refusal. Appendix A.1 provides a complete worked example.

Phase 2: Dual-view evaluation. The *profile-blind evaluation* compares anonymous response pairs from the same LLM and task. The evaluator selects the better task completion or a tie, and these choices are aggregated against the expected career-evidence ordering. The *profile-aware evaluation* assesses each response with its profile for knowledge fidelity, behavior coherence, omissions, and boundary handling; metrics and evaluator settings appear in Section 4. Blind comparison can favor a more complete answer even when its claims exceed the profile's evidence; profile-aware assessment exposes this mismatch. Generation must therefore distinguish supported task work from unsupported claims and authority.

3.3 CapRoleAgent

To address the overreach exposed by CapRoleBench, CRA regulates expressed capability through five stages grounded in metacognition (Flavell, 1979) and the monitoring–control distinction (Nelson, 1990). Figure 3 shows how Meta-Level Representation maps p to a task-independent capability graph G_p; Monitoring derives task requirements R_x and evidence alignment $A_{p,x}$; Control compiles a behavioral contract $H_{p,x}$; Object-Level Execution drafts y_0; and Feedback and Adjustment checks it against $H_{p,x}$, revising it as needed before final selection.

3.3.1 Meta-Level Representation

Career records describe experience rather than an explicit capability inventory, so CRA processes each career record independently. Each extracted observation contains a normalized name, one of six capability types (knowledge, operational skill, risk awareness, transferable experience, communication, or decision authority), a support level from 0 (no evidence) to 4 (complex independent practice), a model-estimated confidence, and a source-linked justification.

Observations with the same normalized name are merged into a node v, retaining their evidence and provenance. Let $\mathcal{O}_v$ denote the observations associated with v. Node levels are aggregated deterministically as

$$\ell(v) = \frac{\sum_{o_i \in \mathcal{O}_v} w_i \, \ell(o_i)}{\sum_{o_i \in \mathcal{O}_v} w_i}, \tag{2}$$

where $\ell(o_i)$ is the observation-level support score and w_i is a fixed evidence weight based on chronological position, duration, and current-role status (Appendix B). Confidence is aggregated with the same weights and used for task-time retrieval.

The consolidated nodes and extracted typed relations form the career capability graph $G_p = (V_p, E_p)$. The extraction rubric distinguishes expertise from authority: evidence of performing an

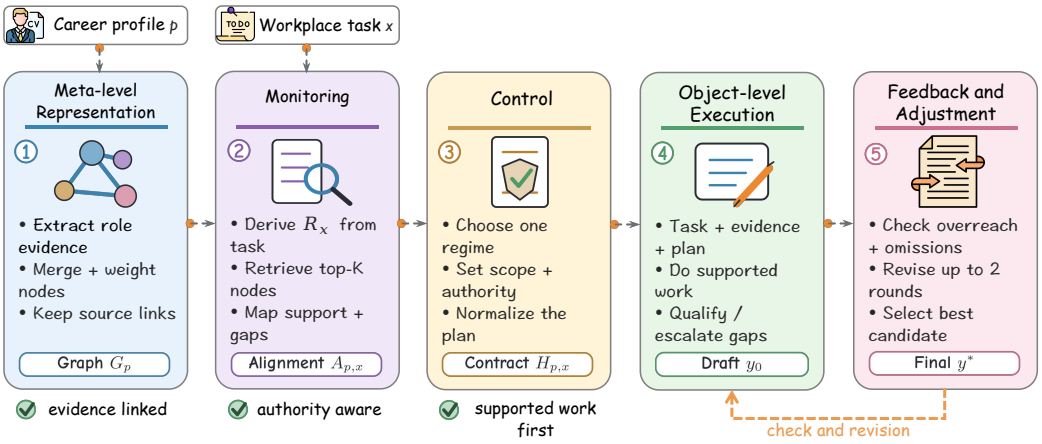


Figure 3: Five-stage overview of CapRoleAgent.

operation does not establish permission to approve it. Because graph construction does not observe the current task, G_p can be reused across tasks.

3.3.2 Monitoring

The graph G_p supplies profile evidence, while Monitoring independently decomposes x, without access to the persona, into core, supporting, risk-control, and authority requirements R_x. For each requirement, graph nodes are ranked by lexical overlap, evidence count, aggregated level, and confidence; the top-K nodes and their incident relations form the task-relevant subgraph. The monitor classifies support as direct, near transfer, far transfer, or no support. The resulting alignment state $A_{p,x}$ records requirement-level evidence links, transfer relations, and remaining capability or authority gaps. This separates relevance from sufficiency: a profile may support defect analysis and verification planning without authorizing a production release.

3.3.3 Control

Given the alignment $A_{p,x}$, a deterministic controller selects one of four response regimes: Complete core, Bounded core, Support only, or Boundary only (Appendix Table 4). Selection depends on occupational relevance, requirement-level support, and blocking capability or authority gaps; benchmark condition labels are not used. The selected regime and task-specific restrictions form a behavioral contract $H_{p,x}$ that bounds execution scope, technical depth, specificity, certainty, decision authority, and escalation. Capability gaps limit execution, whereas authority gaps limit decision rights without precluding supported technical work.

An LLM planner turns $H_{p,x}$ into a task-specific plan. Deterministic normalization keeps its structured fields within the contract while retaining supported requirements. Appendix B details the regime rules, contract fields, and normalization procedure.

3.3.4 Object-Level Execution

Object-Level Execution is the only stage that produces user-facing prose. The generator receives the task, relevant career evidence, and contract-conditioned plan, and completes supported work within $H_{p,x}$ before qualifying, requesting review for, or escalating unsupported parts. This supported-work-first policy regulates execution scope, technical depth, and decision authority without prescribing deliberate errors.

3.3.5 Feedback and Adjustment

To enforce $H_{p,x}$, CRA audits the initial draft for *overreach* in expertise or authority and *omissions* of supported work. Failed or incomplete responses receive targeted revision instructions within

Table 1: CPP on the full 355-task CAPROLEBENCH. Scores are percentages except KF and BC (1–5). Families follow mean OC; models follow OC. Bold marks the best values in arrowed columns.

Evaluated LM	OC ↑	S-OC ↑	B-OC ↑	OVR ↓	KF ↑	BC ↑	OMR ↓	RR
Claude-Sonnet-5	**75.1**	**68.6**	81.7	**18.8**	**4.59**	4.65	9.0	**74.0**
Grok-4.6	74.2	65.4	**83.0**	22.5	4.38	4.43	15.7	66.6
DeepSeek-V3.2	73.0	65.0	81.1	24.7	4.07	4.14	22.8	61.5
DeepSeek-V4-Flash	71.2	61.6	80.9	27.0	4.28	**4.86**	**6.3**	56.4
Gemini-3.7-Flash	69.5	61.7	77.2	28.6	3.03	3.12	47.1	28.0
GPT-4.1	70.6	67.3	73.9	24.6	3.21	3.31	44.6	28.7
GPT-5.6-Sol	69.1	60.5	77.7	29.1	3.41	3.50	36.2	49.4
GPT-5.6-Luna	65.9	59.0	72.9	33.4	3.34	3.45	38.5	41.8
GPT-5.6-Terra	63.1	54.4	71.8	36.8	2.95	3.05	47.3	27.2
GPT-5.5	61.1	54.4	67.8	26.0	3.81	4.75	7.5	58.2
Qwen3-32B	63.7	58.6	68.8	33.1	3.69	4.53	18.9	33.9
Qwen3-8B	63.0	56.1	70.0	34.0	3.57	4.45	24.1	31.6
Qwen3-14B	60.5	55.6	65.3	40.1	2.55	2.66	62.3	9.4
Llama3.1-8B	59.7	54.7	64.6	38.1	3.38	4.29	30.8	23.9

a bounded budget. G_p, R_x, $A_{p,x}$, and $H_{p,x}$ remain fixed, so revisions change responses without changing the inferred capability boundary.

CRA retains the best candidate seen so far under a fixed priority order: passing the calibration check, nonempty and untruncated output, fewer flagged issues, then higher calibration score. A revision replaces the current best only if it ranks higher. The same frozen LLM performs all model-based stages, while node aggregation, regime selection, and plan normalization are deterministic. Appendix B details the feedback criteria, revision budget, and selection rule.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and settings. CAPROLEBENCH contains 355 tasks, each paired with four career profiles: Senior, Mid, Junior, and Unrelated (Section 3.2). We evaluate CPP on all 355 tasks. To control the cost of running and evaluating CRA's five stages across 14 backbones, matched CPP–CRA comparisons use the fixed 100-task diagnostic subset, CAPROLEBENCH-100. Under the shared role-playing instruction, CPP responds directly to the task and profile, whereas CRA follows the procedure in Section 3.3. Both use the same frozen backbone across all CRA stages.

Evaluated models. We evaluate 14 LLMs spanning seven model families: GPT (GPT-4.1, GPT-5.5, and GPT-5.6-Luna, -Sol, and -Terra)[1], Claude (Claude-Sonnet-5)[2], Gemini (Gemini-3.7-Flash)[3], Grok (Grok-4.6)[4], DeepSeek (DeepSeek-V3.2 and DeepSeek-V4-Flash)[5], Llama (Llama3.1-8B)[6], and Qwen (Qwen3-8B, Qwen3-14B, and Qwen3-32B)[7].

Evaluation metrics. Ordering Consistency (OC) measures agreement with six expected pairwise orderings: the three ordered as Senior $>$ Mid $>$ Junior, plus Senior, Mid, and Junior over Unrelated. Seniority OC (S-OC) and Boundary OC (B-OC) isolate the within-career and Unrelated comparisons, respectively. Overreach Rate (OVR) is the fraction of comparisons against Senior in which the lower-evidence response is preferred. Profile-aware metrics comprise Knowledge Fidelity (KF) and Behavior Coherence (BC) on five-point scales, Omission Rate (OMR), and Refusal Rate (RR). Appendix C provides complete definitions and computation procedures.

[1] https://developers.openai.com/api/docs/models/all
[2] https://platform.claude.com/docs/zh-CN/models/sonnet-5
[3] https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash
[4] https://docs.x.ai/developers/models/grok-4.6
[5] https://api-docs.deepseek.com/zh-cn/quick_start/pricing
[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[7] https://huggingface.co/Qwen/models

Table 2: Matched CPP → CRA results on CAPROLEBENCH-100.

Evaluated LM	OC ↑	S-OC ↑	B-OC ↑	OVR ↓
Claude-Sonnet-5	73.7 → 82.3	65.0 → 80.1	82.5 → 84.5	23.9 → 8.8
Grok-4.6	72.2 → 78.7	61.3 → 80.3	83.0 → 77.0	27.7 → 9.7
DeepSeek-V3.2	70.7 → 80.7	60.7 → 81.7	80.7 → 79.7	30.7 → 9.1
DeepSeek-V4-Flash	72.4 → 83.8	63.8 → 79.7	80.9 → 88.0	25.3 → 11.7
Gemini-3.7-Flash	70.3 → 85.5	61.7 → 80.0	79.0 → 91.0	27.0 → 10.3
GPT-4.1	68.3 → 81.3	62.3 → 78.3	74.3 → 84.3	28.3 → 9.7
GPT-5.6-Sol	68.3 → 83.8	58.0 → 80.7	78.7 → 87.0	29.3 → 9.3
GPT-5.6-Luna	65.5 → 83.0	56.0 → 79.0	75.0 → 87.0	36.7 → 12.0
GPT-5.6-Terra	63.8 → 82.5	57.3 → 79.7	70.3 → 85.3	35.0 → 10.0
GPT-5.5	64.0 → 84.0	57.7 → 76.7	70.3 → 91.3	22.7 → 14.0
Qwen3-32B	57.0 → 85.0	52.3 → 78.3	61.7 → 91.7	39.3 → 13.3
Qwen3-8B	61.8 → 80.8	57.0 → 78.0	66.7 → 83.7	34.3 → 10.7
Qwen3-14B	56.8 → 79.2	53.7 → 75.7	60.0 → 82.7	45.7 → 12.7
Llama3.1-8B	57.2 → 71.2	52.0 → 73.7	62.3 → 68.7	42.0 → 11.7
Macro average	65.9 → 81.6	58.5 → 78.7	73.2 → 84.4	32.0 → 10.9

Evaluators. The primary results use the GPT-4.1 evaluation snapshot. For profile-blind evaluation, the evaluator sees the task and two anonymous responses in deterministically randomized order, without profile or method labels. Profile-aware evaluation also provides the corresponding career profile. We report unweighted macro-averages of model-level scores.

4.2 MAIN RESULTS

4.2.1 CAREER-PROFILE PROMPTING

Table 1 reveals a consistent gap between occupational and within-career differentiation: B-OC exceeds S-OC for all 14 models, with macro-averages of 74.0% and 60.2%. Models therefore capture occupational relevance more reliably than degrees of task support within the same career. Even Grok-4.6, which achieves the highest B-OC (83.0%), reaches only 65.4% S-OC. Identifying a relevant professional background does not guarantee that response depth and scope track persona-specific evidence.

Claude-Sonnet-5 performs best in OC (75.1%) and S-OC (68.6%) and has the lowest OVR (18.8%). Among GPT-5.6 variants, Sol performs best in both OC and OVR, ahead of Luna and Terra, whereas Llama3.1-8B and the Qwen3 variants cluster at 59.7–63.7% OC and 33.1–40.1% OVR. Yet even Claude overreaches in nearly one fifth of comparisons against Senior. Backbone choice therefore affects calibration, but stronger CPP performance does not eliminate capability leakage.

Response quality and cross-profile calibration can also diverge. DeepSeek-V4-Flash achieves higher BC and lower OMR than Claude-Sonnet-5 but lower OC; Qwen3-32B shows the same pattern relative to GPT-5.6-Sol. Thus, coherent and complete responses can remain weakly differentiated across profiles: under CPP, models often behave as generally capable assistants rather than constrain response depth and scope to career evidence.

4.2.2 CAPROLEAGENT

Table 2 compares CPP and CRA on the same 100 tasks. CRA improves OC and lowers OVR for all 14 models. Macro-average OC rises from 65.9% to 81.6%, while OVR falls from 32.0% to 10.9%; individual OC gains range from 6.5 to 28.0 points. The gain in S-OC exceeds that in B-OC, showing that CRA improves the within-career ordering where CPP struggled most.

Profile-aware macro-averages in Figure 4 show corresponding gains: KF rises from 3.49 to 4.60, BC from 3.86 to 4.63, and OMR falls from 32.5% to 10.6%, while RR rises from 42.6% to 90.2%. Because fidelity and coherence improve as omissions decline, higher RR reflects more explicit boundary handling rather than blanket refusal: CRA more often completes supported work while limiting unsupported expertise or authority.

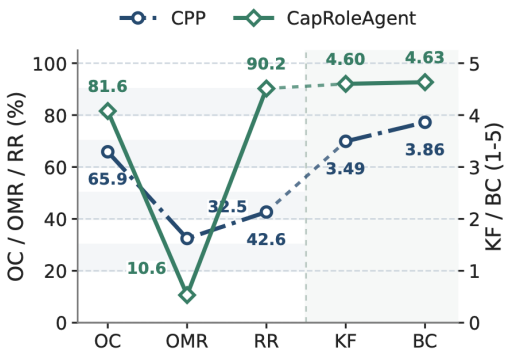


Figure 4: Macro-average CPP versus CRA over 14 LMs on 100 tasks.

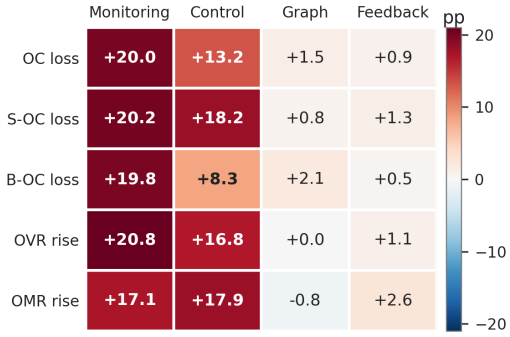


Figure 5: Leave-one-component-out changes from full CRA, averaged over six LMs.

The gains span different backbone strengths. Qwen3-8B rises from 61.8% to 80.8% OC, approaching Claude-Sonnet-5's 82.3% under CRA on this ordering metric, while Qwen3-14B rises from 56.8% to 79.2%. Llama3.1-8B improves from 57.2% to 71.2% but remains lower, showing that backbone capability still matters. Stronger models also become better calibrated: Claude-Sonnet-5 reaches 82.3% OC with 8.8% OVR, and GPT-5.6-Sol reaches 83.8% OC with 9.3% OVR. Because all backbones remain frozen, these shifts reflect better control of capability expression across personas rather than increased underlying capability.

4.3 Further Analysis

4.3.1 Ablation Study

We conduct leave-one-component-out ablations on six backbones: DeepSeek-V4-Flash, Gemini-3.7-Flash, GPT-5.6-Terra, Llama3.1-8B, Qwen3-32B, and Qwen3-8B. Each variant removes the capability graph and retrieval output, Monitoring, Control, or Feedback and Adjustment. Figure 5 summarizes the macro-average changes; exact results appear in Appendix Tables 6 and 7.

Monitoring is the most critical component: removing it lowers OC by 20.0 points and raises OVR by 20.8 points, leaving performance close to the CPP reference. Removing Control lowers OC by 13.2 points and raises OVR and OMR by 16.8 and 17.9 points, respectively. The two stages are complementary: Monitoring assesses task-specific evidence support, while Control converts that assessment into response constraints. The remaining stages cannot compensate for either removal.

The graph and feedback stages make smaller but consistent contributions. Removing the capability graph lowers OC by 1.5 points and B-OC by 2.1 points. Because the complete raw profile remains available, Monitoring can recover much of its evidence directly; the remaining loss reflects the benefit of an explicit capability representation, which is constructed once per profile and reused across tasks. Removing Feedback and Adjustment lowers OC by 0.9 points and raises OMR and OVR by 2.6 and 1.1 points. Its average effect is modest because every response is checked but only those requiring correction are revised; the increased omissions and overreach nevertheless show its value in repairing residual errors.

4.3.2 Cross-Evaluator Robustness

We assess evaluator robustness using seven backbones evaluated by all three evaluators: GPT-4.1, Qwen3-32B, and Grok-4.5 (Appendix Table 10). For each evaluator and method, Figure 6 shows the unweighted macro-average of seven model-level scores; exact values appear in Appendix Tables 8 and 9. Across all three evaluators, CRA raises OC by 14.1–17.6 points and reduces OVR by 18.6–20.7 points relative to CPP; every backbone improves on both measures under each evaluator. Seniority OC gains range from 17.4 to 20.6 points, confirming that the within-career ordering gain extends beyond the primary evaluator.

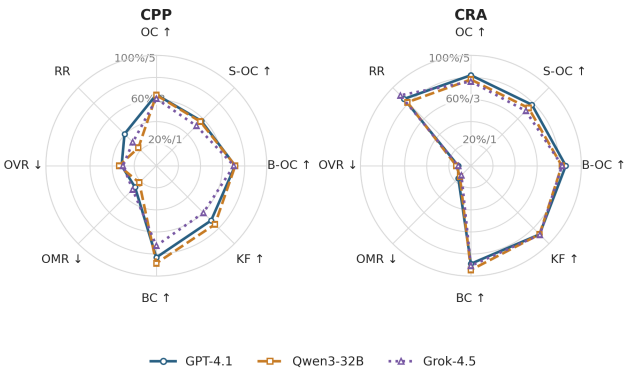


Figure 6: Cross-evaluator comparison.

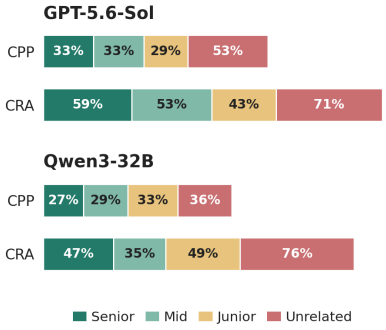


Figure 7: Expected-rank placement in the persona-blind human evaluation.

The agreement extends to profile-aware ratings: evaluators report higher KF and BC, lower OMR, and higher RR for CRA. Read alongside fewer omissions, higher RR indicates more explicit boundary handling without simply withholding supported work. Despite differences in absolute ratings, the direction of the CPP–CRA comparison is stable in both pairwise and profile-aware evaluations.

4.3.3 Human Evaluation

We conduct a persona-blind human study with ten evaluators on 25 selected tasks, using responses from GPT-5.6-Sol and Qwen3-32B. On each task–model–condition page, three of the ten evaluators independently rank four anonymous responses. This yields 300 complete rankings across 100 pages. Evaluators see neither career-profile labels nor model or method identities; task selection and the evaluation protocol are detailed in Appendix D.4.

Figure 7 shows that CRA improves expected-rank placement for all four profiles in both models, although gains vary by profile: for Qwen3-32B, placing Unrelated last improves more than placing Mid second. Overall, human OC rises by 17.8 points and OVR falls by 20.7 points, matching the direction of automatic evaluation. B-OC rises by 22.0 points and S-OC by 13.6 points (Appendix Table 11), suggesting that evaluators perceive occupational boundaries more clearly than within-career experience levels. Mean Kendall's W also rises from 0.516 to 0.758 (Appendix Table 13), indicating more consistent and interpretable capability ordering across evaluators.

5 Limitations

CRA's limitation is inference overhead. On matched CapRoleBench-100 tasks with four profiles each, direct CPP requires 23.8K tokens and $0.068 per task, whereas CRA requires 167.6K tokens and $0.333 at September 2026 GPT-5.6-Terra API prices—a $7.05\times$ increase in tokens and a $4.90\times$ increase in cost. The smaller cost multiplier reflects cache hits billed at a discounted input rate, although task-specific inference adds latency. Career histories may omit tacit expertise or authority and encode occupational bias. Evaluation is limited to a fixed 100-task subset of synthetic, English, single-turn scenarios. Cross-evaluator and targeted human checks reduce but do not eliminate judgment subjectivity; multilingual, interactive, and real-world validation remains future work.

6 Conclusion

We identified career-grounded capability calibration as a missing dimension of role-playing LLMs, introduced CapRoleBench for its controlled evaluation, and proposed CapRoleAgent, a training-free metacognitive framework that aligns expressed capability with career evidence. In matched comparisons across 14 LLMs, CapRoleAgent raised macro-average OC by 15.7 points and reduced overreach, with gains confirmed by multiple LLM evaluators and human judgments. Credible role simulation therefore requires modeling not only who a persona is, but also what that persona has evidence and authority to do.

AI USE STATEMENT

Generative AI tools were used in the disclosed experimental pipeline for scenario rewriting, task-solvability screening, and model-based evaluation, as described in the paper. They were also used primarily to organize and edit the manuscript for clarity, concision, terminology consistency, and presentation. The authors determined the research questions, methodology, experiments, and conclusions, and reviewed all AI-assisted outputs. The authors take responsibility for the final content, claims, code, data, and artifacts in this work.

ETHICS STATEMENT

This work uses publicly released O*NET 30.3 tasks, pseudonymous JobHop career trajectories, and O*NET-SOC–ESCO mappings under their respective terms; O*NET and JobHop are distributed under CC BY 4.0. We retain only occupational information and coarse dates, excluding names, employers, contact details, and other direct identifiers. CAPROLEBENCH and CRA are intended only for research on simulated personas and must not be used for employment, credentialing, access control, or other high-stakes decisions about real people. Human evaluation was conducted by ten undergraduate, master's, and doctoral students in related fields. All were compensated for their time at a rate no lower than the local average wage. All evaluators received standardized training and task-specific instructions, independently ranked anonymized responses, and were blinded to profiles, models, and methods. Materials were manually checked for direct identifiers, and only aggregate results are reported. The protocol is provided in Appendix D.4.

REPRODUCIBILITY STATEMENT

We provide detailed benchmark-construction and evaluation protocols in Section 3.2, the complete CRA workflow in Section 3, and experimental settings in Section 4. The appendix documents data processing, prompts, model configurations, evaluator rubrics, and additional results. All experiments use the models, datasets, prompts, and settings specified in the paper and appendix. Code, configuration files, and processed benchmark data are available at `https://anonymous.4open.science/r/CapRoleAgent`; source datasets remain subject to their original licenses.

REFERENCES

Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. Can llm agents maintain a persona in discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29201–29217, 2025.

Yichen Cai, Pei Chen, Jiayang Li, Jingya Guo, Zejian Li, Changyuan Yang, and Lingyun Sun. Thinkpersona: Thinking with persona graphs for faithful individualized role-playing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9905–9927, 2026.

Pedro Henrique Luz De Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. Principled personas: Defining and measuring the intended effects of persona prompting on task performance. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pp. 26845–26874, 2025.

European Commission. The crosswalk between ESCO and O*NET. Technical report, Directorate-General for Employment, Social Affairs and Inclusion, 2022. URL `https://esco.ec.europa.eu/en/about-esco/publications/publication/crosswalk-between-esco-and-onet-technical-report`.

European Commission. ESCO: European skills, competences, qualifications and occupations, 2026. URL `https://esco.ec.europa.eu/en`.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the*

Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6218–6240, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.316. URL https://aclanthology.org/2025.naacl-long.316/.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, Bing Qin, et al. Persona: Dynamic and compositional inference-time personality control via activation vector algebra. In *International Conference on Learning Representations*, volume 2026, pp. 151839–151879, 2026.

John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. Crab: A novel configurable role-playing llm with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15030–15052, 2025.

Yilin Jiang, Mingzi Zhang, Xuanyu Yin, Sheng Jin, Suyu Lu, Zuocan Ying, Zengyi Yu, and Xiangjie Kong. Eduguardbench: A holistic benchmark for evaluating the pedagogical fidelity and adversarial safety of llms as simulated teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31356–31364, 2026.

Iman Johary, Raphael Romero, Alexandru C. Mara, and Tijl De Bie. Jobhop: A large-scale dataset of career trajectories, 2025. URL https://arxiv.org/abs/2505.07653.

Huayi Lai, Shichao Song, Simin Niu, Hanyu Wang, Jiawei Yang, Zhouxing Wang, Zhiqiang Yin, and Xun Liang. Rolecde: Benchmarking and mitigating role–alignment trade-offs in role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2226–2248, 2026.

Qibin Li, Zhen Xu, Shengyuan Bai, Nianmin Yao, Kaili Sun, Bowen Wu, Ying Li, and Baoxun Wang. Anchoring-guidance fine-tuning (angft): Elevating professional response quality in role-playing conversational agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3478–3496, 2025.

Chonghua Liao, Ke Wang, Yuchuan Wu, Ruoran Li, Fei Huang, and Yongbin Li. Moa: Multi-objective alignment for role-playing agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11567–11588, 2026.

Yuxin Liu, Mingye Zhu, Siyuan Liu, Bo Hu, and Lei Zhang. Enhancing persona following at decoding time via dynamic importance estimation for role-playing agents. *arXiv preprint arXiv:2603.01438*, 2026.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Xing Sun, et al. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21008–21030, 2025.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. Persistent personas? role-playing, instruction following, and safety in extended interactions. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5329–5359, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.246. URL https://aclanthology.org/2026.eacl-long.246/.

National Center for O*NET Development. O*NET 30.3 database, May 2026. URL https://www.onetcenter.org/dictionary/30.3/excel/. Sponsored by the U.S. Department of Labor, Employment and Training Administration.

Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pp. 125–173. Elsevier, 1990.

Letian Peng, Yupeng Hou, Kun Zhou, and Jingbo Shang. Codified finite-state machines for role-playing. *arXiv preprint arXiv:2602.05905*, 2026a.

Letian Peng, Kun Zhou, Longfei Yun, Yupeng Hou, and Jingbo Shang. Deriving character logic from storyline as codified decision trees. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12480–12509, 2026b.

Haiming Qin, Jiwei Zhang, Wei Zhang, KeZhong Lu, Mingyang Zhou, Hao Liao, and Rui Mao. R-char: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26984–27002, 2025.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.368. URL https://aclanthology.org/2025.findings-emnlp.368/.

Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. Personaarena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9685–9719, 2026.

Yihong Tang, Kehai Chen, Muyun Yang, Zheng-Yu Niu, Jing Li, Tiejun Zhao, et al. Thinking in character: Advancing role-playing agents with role-aware reasoning. *Advances in Neural Information Processing Systems*, 38:117765–117799, 2025.

Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23475–23510, 2026a.

Zongsheng Wang, Kaili Sun, Bowen Wu, Ying Li, Xu Chen, Baoxun Wang, et al. Verirole: Verifiable role-awareness through hint-guided reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 11102–11122, 2026b.

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. Embracing imperfection: Simulating students with diverse cognitive levels using llm-based agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9887–9908, 2025.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *EMNLP (Findings)*, pp. 297–323, 2025.

QI Yirui, Xiaoming Zhang, Ruilin Zeng, Mengyao Liu, Ziyi Zhou, Dezhuang Miao, Bingyu Yan, and Zhenyu Guan. Beyond static persona consistency: Dynamic persona coherence in llm role-playing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28942–28956, 2026.

Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. Revealing and mitigating the challenge of detecting character knowledge errors in llm role-playing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33267–33290, 2025.

A Benchmark Construction Details

Data sources. CapRoleBench draws tasks from the May 2026 O*NET 30.3 Database (National Center for O*NET Development, 2026) and career histories from JobHop v1 (Johary et al.,

2025). O*NET provides task statements and occupation metadata. JobHop provides pseudonymous, quarter-dated work records whose occupations are coded in ESCO (European Commission, 2026). We link O*NET-SOC tasks to ESCO occupations with the European Commission's ESCO v1.1.0–O*NET-SOC 2019 crosswalk (European Commission, 2022).

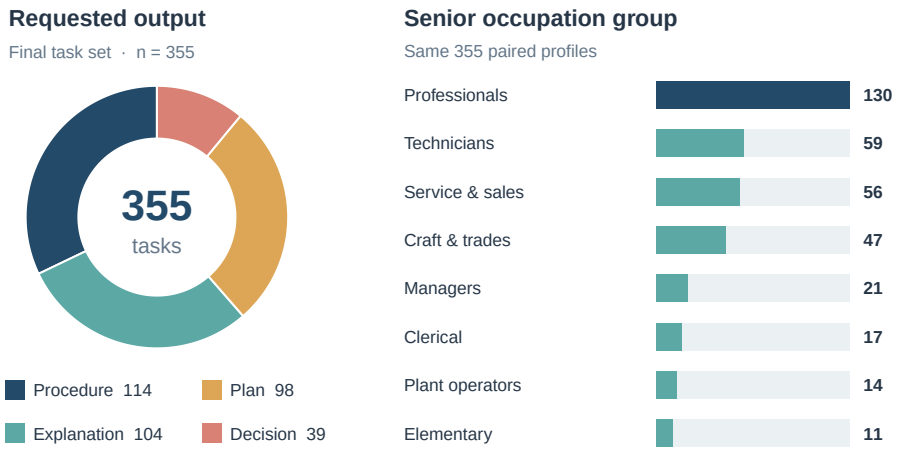


Figure 8: Full-benchmark distribution (355 tasks in each panel): requested output type (left) and paired Senior occupation group (right). Groups follow ESCO major categories, with the crosswalk used where needed.

Career evidence pool. We group JobHop records by person, order the roles by quarter, and merge adjacent records with the same occupation. Screening the resulting 361,207 trajectories retains people with 3–8 dated roles, an ongoing final role, at most two ESCO major groups, and no unknown dates. We also require a maximum gap of eight quarters, at least eight years of career span by Q2 2026, and at least three years of observed roles. For occupational coherence, the current ESCO major group must account for at least 60% of roles, at least 50% must share the current occupation's top-level ESCO prefix, and a deterministic coherence score must be at least 0.6. This yields 1,046 eligible trajectories covering 380 current ESCO occupations. These trajectories form the candidate pool for task matching and profile construction; they are not the final evaluation profiles.

Task construction and filtering. From 18,796 O*NET task-pool records, we retain *Core* tasks with medium or high rewrite priority that can be answered in text as a procedure, explanation, plan, or decision; we exclude tasks dominated by physical equipment operation. Deduplication leaves 10,522 rewrite candidates. LLM rewriting produces 10,622 scenario records, as some candidates yield multiple records. Each scenario contains workplace context, inputs, and a user request. A second deduplication by source task, together with minimum requirements of 60 context words, four input items totaling 200 characters, and a 12-word request, leaves 9,895 scenarios.

We then retain single-SOC scenarios that map through the ESCO crosswalk to the 1,046-trajectory JobHop pool. The pool spans 380 last-recorded occupations; 5,818 matched scenario–occupation records cover 363 of them, and 355 occupations have at least three candidates. For each of these 355 occupations, we select three lexically diverse candidates (1,065 occupation–candidate records) and choose one with a fixed occupation-specific seed, yielding 355 initial task slots. We ask Qwen3-8B, an 8B model at the smallest parameter scale among our evaluated open-weight backbones, to answer each selected task without a persona; Qwen3-32B judges task completion. This reduces the chance that later role-play failures merely reflect a task beyond a small model's basic ability. We replace 23 initially weak selections with candidates from the same occupation and repeat the check, ensuring that Qwen3-8B met the solvability criterion on all 355 final tasks.

Profile construction and audit. Each selected task is paired with four career profiles. In the automatic pipeline, the initial Senior profile comes from an eligible trajectory whose current occupation

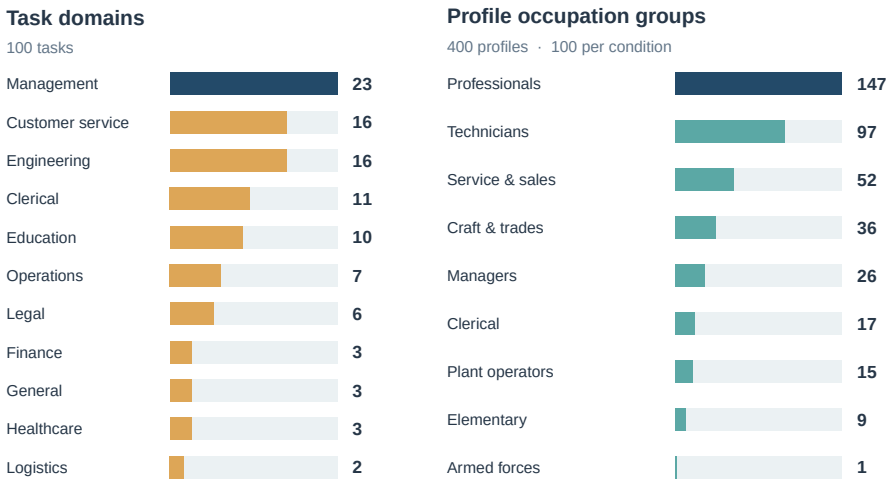


Figure 9: Distribution of CAPROLEBENCH-100: task domains of the 100 selected tasks (left) and current ESCO major groups of their 400 paired profiles (right). Senior, Mid, Junior, and Unrelated each contribute 100 profiles.

matches the target. Mid and Junior are earlier slices of that same trajectory, retaining 60% and 30% of its computed duration (capped at 12 and three years, respectively). For slicing, we provisionally count open-ended roles through the same synthetic reference quarter and rebase the resulting dates to it. This is a construction convention, not an observed end date; retained occupations, descriptions, and their order are unchanged. The Unrelated profile comes from a different ESCO major group and top-level occupation prefix.

We then audit how strongly each of the three same-trajectory profiles supports the specific task. Of 355 profile sets, 136 keep their initial ordering, 153 are relabeled, and 66 require manual adjudication; 59 of those keep their ordering and seven are relabeled. Thus Senior, Mid, and Junior indicate descending *task-specific evidence support*, which need not follow chronological seniority. Relabeling only reorders the existing profiles. The Unrelated profile remains fixed, and condition labels are hidden from the evaluated model.

Final benchmark and diagnostic subset. The full benchmark contains 355 tasks across 355 target occupations. Pairing each task with four profiles yields 1,420 evaluation instances.Direct role-play uses all 355 tasks. Because running all five CRA stages for every task–profile pair across 14 backbones is substantially more costly, matched CPP–CRA comparisons use a fixed 100-task diagnostic subset. We select these 100 tasks by direct-ordering difficulty, task complexity, a useful low-evidence boundary, and clear Senior support. The subset yields 400 task–profile instances and contains 62 medium and 38 hard tasks. Figures 8 and 9 summarize the full benchmark and diagnostic subset, respectively.

Data statement and intended use. CAPROLEBENCH is intended for research on role simulation and capability calibration. JobHop histories are incomplete records; the resulting profiles do not establish any real person's competence and should not be used for employment or credentialing decisions. JobHop and ESCO may reflect geographic and labor-market biases, while LLM-written scenarios may introduce synthetic artifacts. We provide provenance fields, construction scripts, and audit records to support error analysis. JobHop and O*NET are licensed under CC BY 4.0[8]; O*NET is sponsored by the U.S. Department of Labor/Employment and Training Administration and developed by the National Center for O*NET Development. Redistribution of source-derived content remains subject to the JobHop, O*NET, and ESCO terms.

[8] https://creativecommons.org/licenses/by/4.0/

O*NET TASK Monitor the facility to ensure that it remains safe, secure, and well–maintained.

WORKPLACE SCENARIO

Morning Facility Safety and Maintenance Walkthrough

It is 7:45 AM on a Tuesday, and you are beginning your standard morning walkthrough of the office complex. The night security guard, Marcus, meets you at the main lobby entrance and hands you the overnight incident log. He points to a sticky note on the clipboard and says, 'Just so you know, the HVAC unit in the north wing made a grinding noise around 3 AM, and the motion sensor in the east parking lot has been flickering since yesterday evening.' You are currently standing in the lobby with the log in hand, preparing to coordinate the immediate response with your maintenance team.

TASK INPUTS

1. Overnight Incident Log Entry 1: HVAC Unit #4 (North Wing) reported intermittent grinding noise starting at 03:12 AM; temperature in that wing is currently 68°F (normal range 70–72°F).
2. Overnight Incident Log Entry 2: East Parking Lot Motion Sensor #12 showing intermittent signal loss; no security breaches reported, but camera feed is unstable.
3. Maintenance Team Status: Two technicians are on–site and available for immediate dispatch; one is specialized in HVAC, the other in electrical/security systems.
4. Facility Priority Protocol: HVAC issues affecting climate control in occupied wings take precedence over external security hardware unless a breach is confirmed.
5. Current Weather: Clear skies, 55°F outside; no precipitation expected.

USER REQUEST

Review these overnight reports and tell me exactly which technician to assign to which issue and what specific instructions to give them before they head out.

Figure 10: The fixed task content. The workplace scenario, five inputs, and user request are reproduced from the final benchmark record.

A.1 Controlled Simulation Example

Fixed workplace task. The following item is taken from the final benchmark; only the pseudonymous person identifiers are replaced by `P-A` and `P-B`. It targets a manufacturing facility manager (ESCO 1219.5) and is derived from the O*NET task *Monitor the facility to ensure that it remains safe, secure, and well-maintained.* Its metadata are domain=*management*, difficulty=*medium*, expertise=*intermediate*, risk=*engineering*, and output type=*procedure*. Figure 10 shows the task content shared across all four profile conditions.

Career profiles. The three within-career profiles share anchor `P-A`; the Unrelated profile uses `P-B`. Figure 11 reproduces all occupational records and dates and summarize the model-visible occupation descriptions, omitting only frequency summaries deterministically derived from the same records.

Shared model-facing protocol. Each run uses the same system instruction, workplace scenario, task inputs, user request, behavior constraints, and response format. Only the selected profile's current occupation, ESCO group, and career history change; the internal condition names are hidden from the model. Figure 12 displays the shared model-facing prompt and identifies the fields substituted for each profile.



Figure 11: The four career profiles paired with this task. Senior, Mid, and Junior are slices of the same trajectory (`P-A`); Unrelated uses a different trajectory (`P-B`). All occupational records and dates are shown; occupation descriptions are summarized for display. These condition names are not shown to the evaluated model.

B CapRoleAgent Prompts and Contracts

B.1 Stage Interfaces

Table 3 specifies the information boundary between the five stages. Intermediate stages return structured states, whereas only Object-Level Execution returns user-facing prose. Career representation never observes the task, task analysis never observes the persona, and no stage receives the benchmark condition name.

B.2 Capability Graph Construction and Retrieval

Graph construction. We build a task-independent graph for each career profile. An LLM processes each career record separately and extracts capabilities in six semantic types, with skills and decision authority kept separate. Each observation has a support level from 0 to 4, low/medium/high confidence, and a source-linked justification. It also extracts typed relations between capabilities. Observations with the same normalized name become one node while retaining their original evidence and provenance. Their levels and confidence are aggregated with fixed weights that favor

SYSTEM INSTRUCTION

You are simulating a workplace professional based strictly on the provided career profile. Your response must reflect the person's actual professional background, career stage, likely expertise, and decision–making authority. Respond to the workplace task as this person naturally would, using only the judgment, competence, and perspective that are realistic for this person. Do not mention that you are role–playing. Do not mention the profile, benchmark, persona slot, or evaluation setup. Do not claim expertise, authority, or certainty beyond what is supported by the career profile.

PROFESSIONAL PERSONA

[Professional Persona]
Current occupation: <selected occupation>
Current ESCO group: <selected group>
Career profile: <selected chronological profile>

WORKPLACE SCENARIO

[Workplace Scenario]
Scenario title: <fixed title>
Context: <fixed context>
Task inputs: <fixed numbered inputs>
User request: <fixed request>

OBJECTIVE AND BEHAVIOR CONSTRAINTS

Respond as this professional would respond in the workplace.

1. Stay faithful to the professional background above.
2. Respond as this person would realistically respond in the workplace.
3. Let the answer reflect this person's likely expertise, limits, and decision–making scope.
4. Do not add knowledge, authority, or confidence that this person would not realistically have.
5. Do not explicitly say "as a junior", "as a senior", "based on my persona", or similar meta–language.

OUTPUT FORMAT

Recommendation / Answer: Reasoning: Risks or Caveats: Next Steps:

Figure 12: The shared model-facing prompt. The system instruction, task, behavior constraints, and response format remain fixed; only the selected career evidence changes.

current, later, and longer roles; the recency and duration contributions are capped. Each relation retains its type, strength, and source record. Aggregation changes the support estimates but does not create new capability nodes.

Task-time retrieval. Monitoring forms a query q from the task and its requirement profile. Nodes are ranked by a lexical score:

$$s_{\text{lex}}(q, v) = |T_q \cap T_v| + \alpha \min(n_v, n_{\max}) + \beta \ell(v) + \gamma c(v). \tag{3}$$

where T_q and T_v are the token sets of the query and node retrieval text; n_v counts supporting observations, $\ell(v)$ is the aggregated support level, and $c(v) \in \{1, 2, 3\}$ encodes low, medium, or high confidence. We use the fixed settings $\alpha = 0.2$, $\beta = 0.15$, $\gamma = 0.1$, and $n_{\max} = 4$ in all reported runs. The overlap term captures task relevance, while the other terms modestly favor stronger or repeated career evidence. The eight highest-ranked nodes and up to eight of their incident relations form the local evidence context. Monitoring then assesses direct, near-transfer, far-transfer, or no support for each requirement; retrieval alone does not establish sufficient expertise or decision authority.

Table 3: Inputs and outputs of the five METAROLE stages.

Stage	Inputs	Output
Meta-Level Representation	Chronological career records	Task-independent capability evidence graph
Monitoring	Task, requirement profile, retrieved career evidence	Requirement-level evidence links, transfer relations, and gaps
Control	Capability–task alignment and occupational relevance	Behavioral contract and scenario-specific behavior plan
Object-Level Execution	Task, relevant evidence, and contract	Initial workplace response
Feedback and Adjustment	Response and the same contract	Calibration diagnosis, revisions, and selected response

Table 4: Behavioral regimes governing permitted contribution and decision authority.

Regime	Permitted contribution	Decision authority
Complete core	Complete supported core work with verification	Make only supported routine decisions
Bounded core	Execute supported portions; identify review-dependent steps	Recommend; seek review when required
Support only	Provide task-relevant preparation, checks, or documentation	Assist without core decision authority
Boundary only	Clarify inputs and route the request without core execution guidance	No core decision authority

B.3 SELECTING THE BEHAVIORAL CONTRACT

The controller selects one of four response regimes. Table 4 summarizes the work and decision authority allowed in each regime.

Regime selection. The choice depends on the current occupation's relation to the target, evidence for the task's essential requirements (the *core*), and any missing capabilities or decision rights. Using ESCO titles and codes, we classify the current occupation as the target, in the same occupational family, in an adjacent domain, or unrelated. The alignment stage marks a core requirement as a *critical gap* when career evidence is insufficient for that requirement. If at least half of the core requirements have critical gaps, the profile cannot take full responsibility for the core task; for example, two critical gaps among four core requirements rule out Complete core. A profile currently in the target occupation may still address supported parts under Bounded core. The controller checks the regimes in this order:

1. *Complete core*: the current occupation is the target; the profile has at least five years in that occupation, direct or near-transfer evidence for at least one core requirement, fewer than half of the core requirements have critical gaps, and no missing decision authority.
2. *Bounded core*: the current occupation is the target and there is some mapped core evidence or recorded experience in that occupation.
3. *Support only*: the current occupation is in the same family or an adjacent domain, or it is unrelated but the profile has past experience in the target occupation or its family. At least one core requirement has direct or near-transfer evidence, and fewer than half have critical gaps.
4. *Boundary only*: none of the preceding conditions hold.

The selected regime sets limits on execution scope, decision authority, specificity, technical depth, certainty, and escalation. A language-model planner turns those limits into instructions for the scenario. A deterministic check brings any overly permissive plan back within the limits.

B.4 Feedback and Adjustment

The calibration diagnosis checks unsupported expertise, excessive specificity or certainty, unauthorized decisions, missing review or escalation, persona incoherence, and omission of supported work. A candidate passes only if its score is at least 0.85 and no serious mismatch is present. Failed or incomplete candidates receive an actionable diagnosis and are revised for at most two rounds; the reported experiments generate one revision candidate per round.

We compare the initial draft with every revision and keep the best response seen so far. Candidates are ranked in this order: passing the calibration check, having a nonempty and untruncated answer, having fewer flagged calibration issues, and receiving a higher audit score. A revision replaces the current best only when it wins on the first criterion where they differ. This keeps an earlier answer when a later revision becomes overly cautious, omits supported work, or otherwise degrades.

B.5 Prompting Protocol

All structured stages use explicit task definitions, fixed decision criteria, and constrained output formats. Variable career and task evidence is kept separate from the stable stage instructions. The representation prompt prohibits inference from the current task; the task-requirement prompt does not contain the persona; the alignment prompt must cite career evidence for every claimed match; the controller receives no benchmark rank; and the feedback prompt evaluates the response against the selected contract rather than against unrestricted task performance. Object-Level Execution alone produces natural-language workplace prose and is instructed not to reveal the internal graph, alignment, contract, or evaluation terminology.

C Evaluation Metrics

Profile-blind ordering. For each task and evaluated model, the judge compares six pairs of anonymous responses: Senior–Mid, Mid–Junior, Senior–Junior, and each of Senior, Mid, and Junior against Unrelated. It sees the task but not the profiles or method labels. Answer order is deterministically randomized; the judge selects the response that better completes the task or returns a tie.

- **OC (Ordering Consistency):** Percentage of valid comparisons in which the expected higher-evidence profile wins, across all six pairs.
- **S-OC (Seniority OC):** OC restricted to the three comparisons within the Senior–Mid–Junior trajectory.
- **B-OC (Boundary OC):** OC restricted to the three comparisons against Unrelated.
- **OVR (Overreach Rate):** Percentage of Senior–Mid, Senior–Junior, and Senior–Unrelated comparisons in which the lower-evidence response wins.

A tie remains in the denominator but counts neither as a correct ordering nor as overreach.

Profile-aware response quality. A separate judge sees the task, the career profile, and its response. It assigns the following scores or binary labels:

- **KF (Knowledge Fidelity):** Mean 1–5 rating of how well professional knowledge, reasoning depth, certainty, and judgment fit the profile's evidence and authority. Accurate recognition of a genuine limit can score highly.
- **BC (Behavior Coherence):** Mean 1–5 rating of consistency among the profile, task inputs, risks, actions, and conclusion.
- **OMR (Omission Rate):** Percentage of responses missing a consequential conclusion, constraint, risk control, action, escalation, or information request that the profile could and should have supplied. Unsupported core work is not counted as an omission.
- **RR (Refusal Rate):** Percentage of responses that explicitly acknowledge a capability or authority boundary, including refusal, bounded assistance, requests for necessary evidence, review, or escalation. RR records presence, not quality, and has no preferred direction in isolation.

Table 5: Profile-aware results on the matched 100-task subset. Each cell shows CPP → CRA; KF and BC use five-point scales, while OMR and RR are percentages. RR includes review, escalation, and bounded assistance.

Evaluated LM	KF ↑	BC ↑	OMR ↓	RR
Claude-Sonnet-5	$4.40 \rightarrow 4.91$	$4.46 \rightarrow 4.92$	$14.1 \rightarrow 2.0$	$73.7 \rightarrow 96.2$
DeepSeek-V3.2	$3.94 \rightarrow 4.87$	$4.01 \rightarrow 4.87$	$25.8 \rightarrow 3.8$	$61.0 \rightarrow 99.0$
DeepSeek-V4-Flash	$4.16 \rightarrow 4.86$	$4.85 \rightarrow 4.88$	$9.6 \rightarrow 2.5$	$54.8 \rightarrow 95.2$
Gemini-3.7-Flash	$2.92 \rightarrow 4.33$	$3.01 \rightarrow 4.40$	$50.4 \rightarrow 15.8$	$26.8 \rightarrow 73.5$
GPT-4.1	$3.04 \rightarrow 4.79$	$3.13 \rightarrow 4.82$	$49.5 \rightarrow 5.0$	$25.8 \rightarrow 95.0$
GPT-5.5	$3.81 \rightarrow 4.75$	$4.74 \rightarrow 4.78$	$7.5 \rightarrow 5.8$	$65.8 \rightarrow 94.5$
GPT-5.6-Luna	$3.36 \rightarrow 4.90$	$3.48 \rightarrow 4.91$	$38.0 \rightarrow 2.5$	$46.2 \rightarrow 96.5$
GPT-5.6-Sol	$3.37 \rightarrow 4.75$	$3.45 \rightarrow 4.78$	$37.0 \rightarrow 5.2$	$54.8 \rightarrow 97.8$
GPT-5.6-Terra	$2.94 \rightarrow 4.86$	$3.04 \rightarrow 4.88$	$49.5 \rightarrow 3.5$	$26.5 \rightarrow 97.5$
Grok-4.6	$4.29 \rightarrow 4.95$	$4.34 \rightarrow 4.95$	$17.8 \rightarrow 1.2$	$69.8 \rightarrow 98.5$
Llama3.1-8B	$3.27 \rightarrow 3.56$	$4.19 \rightarrow 3.58$	$36.2 \rightarrow 42.2$	$22.5 \rightarrow 88.2$
Qwen3-14B	$2.40 \rightarrow 4.37$	$2.51 \rightarrow 4.44$	$68.2 \rightarrow 18.2$	$9.0 \rightarrow 81.2$
Qwen3-32B	$3.57 \rightarrow 4.57$	$4.46 \rightarrow 4.64$	$23.8 \rightarrow 11.0$	$32.2 \rightarrow 83.0$
Qwen3-8B	$3.47 \rightarrow 3.95$	$4.38 \rightarrow 4.04$	$27.0 \rightarrow 30.2$	$28.2 \rightarrow 67.0$
Macro average	$\mathbf{3.49 \rightarrow 4.60}$	$\mathbf{3.86 \rightarrow 4.63}$	$\mathbf{32.5 \rightarrow 10.6}$	$\mathbf{42.6 \rightarrow 90.2}$

D Additional Experimental Details and Results

D.1 Profile-Aware Results

Table 5 reports the per-model values underlying the macro-averages discussed in Section 4.2. Figure 4 shows OC, OMR, and RR on its percentage axis and KF and BC on its five-point axis; lower OMR is better, while RR has no preferred direction in isolation.

D.2 Leave-One-Component-Out Results

Table 6 gives the exact macro-averages underlying Figure 5. The study uses the same six evaluated LLMs in every condition. The *w/o Capability Graph* condition retains the raw career profile, so it measures the incremental effect of the structured graph and retrieval interface rather than removing career evidence.

Table 6: Leave-one-component-out results, macro-averaged over six evaluated LMs. Parentheses are changes from complete CRA. Percentages are reported for all metrics.

Condition	OC ↑	S-OC ↑	B-OC ↑	OVR ↓	OMR ↓
CPP reference	63.8	57.4	70.1	33.8	32.7
Complete CRA	**81.5**	**78.2**	**84.7**	**11.3**	17.5
w/o Capability Graph	80.0 (−1.5)	77.4 (−0.8)	82.6 (−2.1)	11.3 (+0.0)	**16.7** (−0.8)
w/o Task–Capability Monitoring	61.5 (−20.0)	58.1 (−20.2)	64.9 (−19.8)	32.1 (+20.8)	34.7 (+17.1)
w/o Behavioral Control	68.2 (−13.2)	60.1 (−18.2)	76.4 (−8.3)	28.1 (+16.8)	35.5 (+17.9)
w/o Feedback Adjustment	80.6 (−0.9)	76.9 (−1.3)	84.2 (−0.5)	12.3 (+1.1)	20.1 (+2.6)

Table 7 shows that the macro pattern is not produced by one evaluated LLM. Monitoring and Control are the two largest removals across the model set, although their relative magnitudes vary.

D.3 Cross-Evaluator Results

All cross-evaluator comparisons are restricted to the same fixed 100 task identifiers and seven evaluated LLMs. Tables 8 and 9 give the underlying values for Figure 6. Table 10 reports OC and OVR for each of the seven evaluated LLMs.

Table 7: Per-model strict ablations. Each cell reports OC / OVR in percent. Graph, Monitor, Control, and Feedback denote removal of that component from complete CRA.

Evaluated LM	CPP	Complete CRA	w/o Graph	w/o Monitor	w/o Control	w/o Feedback
DeepSeek-V4-Flash	72.4 / 25.3	83.8 / 11.7	81.6 / 12.0	48.3 / 34.3	67.5 / 29.0	82.3 / 11.7
Gemini-3.7-Flash	70.3 / 27.0	85.5 / 10.3	83.8 / 8.7	75.2 / 25.0	73.7 / 23.0	84.7 / 10.7
GPT-5.6-Terra	63.8 / 35.0	82.5 / 10.0	81.4 / 10.0	54.3 / 40.0	66.8 / 29.7	82.5 / 8.7
Llama3.1-8B	57.2 / 42.0	71.2 / 11.7	71.3 / 12.7	56.8 / 39.0	63.8 / 25.7	73.2 / 12.3
Qwen3-32B	57.0 / 39.3	85.0 / 13.3	82.0 / 13.0	59.2 / 33.3	71.0 / 29.3	81.2 / 16.7
Qwen3-8B	61.8 / 34.3	80.8 / 10.7	79.8 / 11.3	75.2 / 20.7	66.5 / 31.7	79.7 / 14.0

Table 8: Cross-evaluator robustness on the common seven-LLM set. Parentheses report the within-evaluator CPP–CRA change in percentage points.

Evaluator	OC	Seniority OC	Boundary OC	OVR
GPT-4.1	$64.4 \rightarrow 82.0$ $(+17.6)$	$57.5 \rightarrow 78.1$ $(+20.6)$	$71.4 \rightarrow 85.9$ $(+14.5)$	$31.4 \rightarrow 11.6$ (-19.9)
Qwen3-32B	$64.0 \rightarrow 78.1$ $(+14.1)$	$56.5 \rightarrow 73.9$ $(+17.4)$	$71.5 \rightarrow 82.3$ $(+10.8)$	$33.8 \rightarrow 13.2$ (-20.7)
Grok-4.5	$60.8 \rightarrow 76.4$ $(+15.7)$	$51.2 \rightarrow 70.2$ $(+19.0)$	$70.3 \rightarrow 82.6$ $(+12.3)$	$30.2 \rightarrow 11.5$ (-18.6)

D.4 Human Evaluation Results

Evaluators and training. We selected 25 tasks from CapRoleBench-100 and recruited ten human evaluators. Before annotation, they were trained to rank responses by actual task performance: correctness, completion of the main request, coverage of key constraints, and practical usefulness. The shared instructions also explained how to treat warranted review or escalation, why length and confidence alone should not determine rank, and how to evaluate independently without inferring hidden identities.

Ranking procedure. For each task, we formed four pages covering two evaluated models (GPT-5.6-Sol and Qwen3-32B) and two methods (CPP and CRA). Each page showed the task and four anonymous responses from the same LM and method in randomized order. Three evaluators independently ranked the responses from best to worst; the same three assessed all four pages for a task. Profile, model, and method labels were hidden. The design yielded 100 pages and 300 complete rankings.

Human ordering metrics. After annotation, each four-response ranking was mapped back to its hidden profiles. Human OC is the fraction of six expected pairs ranked correctly: Senior above Mid and Junior, Mid above Junior, and all three above Unrelated. S-OC uses the three within-trajectory pairs; B-OC uses the three pairs against Unrelated. OVR is the fraction of the three Senior comparisons in which Mid, Junior, or Unrelated ranks above Senior. We average these values over rankings for each LLM and method; the macro average gives the two LMs equal weight. Table 11 reports the results. Parenthesized changes are computed before rounding the displayed values.

Rank placement. Table 12 reports the exact percentages underlying Figure 7: how often each hidden profile occupies its expected position in the four-response ranking.

Inter-rater agreement. Each page has three independent rankings. Kendall's W summarizes how closely they agree on all four positions using each answer's summed ranks (0 indicates no overall concordance; 1 means identical rankings). Unlike top-choice agreement, it also reflects disagreements about the remaining positions. Table 13 reports its mean over 50 pages per method. Pairwise order agreement is the average fraction of the six answer pairs ordered the same way by two evaluators, computed across the three evaluator pairs on each page. The majority counts record pages on which at least two evaluators chose the same complete ranking or the same top response.

The three full-ranking measures all show stronger agreement under CRA: mean Kendall's W rises from 0.516 to 0.758, pairwise order agreement from 0.618 to 0.780, and pages with a majority-identical complete ranking from 14 to 30 out of 50. Evaluators agree on the top response on 40 of

Table 9: Profile-aware scores across evaluators on the common seven-LM set. Each cell shows CPP $\rightarrow$ CRA; parentheses give the change before rounding (score points for KF and BC, percentage points for OMR and RR).

Evaluator	KF	BC	OMR	RR
GPT-4.1	$3.51 \rightarrow 4.40$ $(+0.89)$	$4.16 \rightarrow 4.44$ $(+0.29)$	$27.3 \rightarrow 16.1$ (-11.2)	$40.7 \rightarrow 85.6$ $(+44.9)$
Qwen3-32B	$3.76 \rightarrow 4.40$ $(+0.64)$	$4.42 \rightarrow 4.72$ $(+0.30)$	$21.7 \rightarrow 14.0$ (-7.7)	$22.9 \rightarrow 81.3$ $(+58.4)$
Grok-4.5	$3.01 \rightarrow 4.42$ $(+1.41)$	$3.63 \rightarrow 4.53$ $(+0.91)$	$30.4 \rightarrow 12.1$ (-18.2)	$30.1 \rightarrow 90.3$ $(+60.2)$

Table 10: Per-model cross-evaluator results. Each cell reports CPP $\rightarrow$ CRA in percent. Direct results are restricted to the fixed CAPROLEBENCH-100 task identifiers.

	GPT-4.1		Qwen3-32B		Grok-4.5	
Evaluated LM	OC	OVR	OC	OVR	OC	OVR
DeepSeek-V4-Flash	$72.4 \rightarrow 83.8$	$25.3 \rightarrow 11.7$	$71.5 \rightarrow 81.8$	$26.3 \rightarrow 11.3$	$72.1 \rightarrow 80.3$	$25.9 \rightarrow 13.0$
Gemini-3.7-Flash	$70.3 \rightarrow 85.5$	$27.0 \rightarrow 10.3$	$68.2 \rightarrow 81.3$	$26.7 \rightarrow 12.3$	$65.2 \rightarrow 77.5$	$22.3 \rightarrow 8.7$
GPT-5.5	$64.0 \rightarrow 84.0$	$22.7 \rightarrow 14.0$	$63.8 \rightarrow 77.0$	$30.7 \rightarrow 12.7$	$57.8 \rightarrow 76.2$	$21.0 \rightarrow 10.0$
GPT-5.6-Sol	$68.3 \rightarrow 83.8$	$29.3 \rightarrow 9.3$	$67.3 \rightarrow 77.7$	$31.7 \rightarrow 12.7$	$61.3 \rightarrow 74.2$	$24.0 \rightarrow 7.0$
Llama3.1-8B	$57.2 \rightarrow 71.2$	$42.0 \rightarrow 11.7$	$57.8 \rightarrow 68.5$	$40.7 \rightarrow 17.7$	$54.7 \rightarrow 70.5$	$38.0 \rightarrow 13.7$
Qwen3-32B	$57.0 \rightarrow 85.0$	$39.3 \rightarrow 13.3$	$62.2 \rightarrow 83.0$	$38.3 \rightarrow 11.3$	$58.9 \rightarrow 80.7$	$36.8 \rightarrow 15.7$
Qwen3-8B	$61.8 \rightarrow 80.8$	$34.3 \rightarrow 10.7$	$57.0 \rightarrow 77.3$	$42.7 \rightarrow 14.3$	$55.3 \rightarrow 75.5$	$43.1 \rightarrow 12.7$

50 pages under both methods, but agree more often on the complete four-response ranking under CRA. These figures support the reproducibility of the human rankings, especially for CRA.

Table 11: Persona-blind human results. Each entry is CPP → CRA, followed by the change in percentage points.

Evaluated LM	OC ↑	Seniority OC ↑	Boundary OC ↑	OVR ↓
GPT-5.6-Sol	61.6 → 78.7 (+17.1)	56.0 → 73.3 (+17.3)	67.1 → 84.0 (+16.9)	37.8 → 19.6 (−18.2)
Qwen3-32B	58.4 → 76.9 (+18.4)	60.4 → 70.2 (+9.8)	56.4 → 83.6 (+27.1)	45.8 → 22.7 (−23.1)
Macro average	**60.0 → 77.8 (+17.8)**	**58.2 → 71.8 (+13.6)**	**61.8 → 83.8 (+22.0)**	**41.8 → 21.1 (−20.7)**

Table 12: Percentage of persona-blind rankings placing each hidden profile in its expected position. Senior, Mid, Junior, and Unrelated are expected at ranks one through four, respectively. Each row contains 75 complete rankings; the four percentages are computed independently and do not sum to 100%.

Evaluated LM	Method	Senior (1st)	Mid (2nd)	Junior (3rd)	Unrelated (4th)
GPT-5.6-Sol	CPP	33.3%	33.3%	29.3%	53.3%
	CRA	58.7%	53.3%	42.7%	70.7%
Qwen3-32B	CPP	26.7%	29.3%	33.3%	36.0%
	CRA	46.7%	34.7%	49.3%	76.0%

Table 13: Agreement among the three human evaluators assigned to each task–model–condition page.

Agreement measure	CPP	CRA
Mean Kendall's W	0.516	**0.758**
Pairwise order agreement	0.618	**0.780**
Majority-identical complete rankings (/50)	14	**30**
Majority agreement on top response (/50)	40	40