RABBITBENCH: EVALUATING LARGE LANGUAGE MODELS ACROSS DIVERSE OCCUPATIONS THROUGH OCCUPATIONAL QUALIFICATION EXAMINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) increasingly participate in occupational work, there is a growing need to comprehensively evaluate their knowledge and capabilities across diverse occupational domains. However, existing benchmarks still lack comprehensive and robust coverage across a broad range of occupations. Occupational qualification examinations support comprehensive and robust LLM evaluation through expert-designed questions that systematically cover occupational tasks and the knowledge and capabilities they require. Following official Chinese guidelines, we collect and organize accessible and evaluable examinations from 26 types of occupational qualifications to construct RabbitBench. The benchmark spans 8 mid-level and 23 minor-level occupational groups in the Chinese occupational classification system, providing broader coverage than existing benchmarks. It includes single-choice, multiple-choice, true-or-false, and subjective questions, together with high-quality reference answers, enabling the evaluation of multiple aspects of model performance. We evaluate a diverse set of LLMs on RabbitBench. GPT-5.6-Sol and Gemini-3.1-Pro achieve the strongest performance but at substantially higher API costs, while Qwen3.5 shows a favorable cost–performance trade-off. Across occupations, general capability dominates overall performance, but similarly capable models exhibit distinct occupational strengths. Within engineering qualifications, models consistently perform better on knowledge-oriented than application-oriented subjects. Together, these results show that RabbitBench enables broad and fine-grained analysis of occupational capabilities beyond aggregate model scores. An anonymous project page is available at `https://rabbitbench.github.io/`.

1 INTRODUCTION

As large language models (LLMs) become more capable, they are increasingly participating in a growing range of occupational work, ranging from answering practitioners' questions (Brynjolfsson et al., 2025) to independently completing occupational tasks (Patwardhan et al., 2025). This growing involvement creates a substantial need to comprehensively evaluate LLMs' knowledge, technical competencies, and capabilities across a broad range of occupations. Existing studies have constructed two main types of occupational benchmarks: task-based benchmarks (Drouin et al., 2024; Boisvert et al., 2024; Xu et al., 2025; Miserendino et al., 2025; Patwardhan et al., 2025; Hu et al., 2026) and exam-based benchmarks (Hong et al., 2025a;b; Luo et al., 2025; Xun et al., 2026; Liu et al., 2023; Zhong et al., 2020). Task-based benchmarks consist of simulated occupational tasks, each typically specifying the work to be completed and providing relevant documents, tools, or other contextual resources. GDPval evaluates occupational work-product tasks across a range of occupations (Patwardhan et al., 2025). Agents' Last Exam covers long-horizon occupational tasks with verifiable outcomes (Sun et al., 2026). $OneMillion-Bench contains 400 expert-curated professional tasks spanning law, finance, healthcare, natural science, and industry (Yang et al., 2026). QualBench collects questions from multiple Chinese professional qualification examinations (Hong et al., 2025a). KMMLU-Redux and KMMLU-Pro draw on Korean national technical qualification and professional licensure examinations (Hong et al., 2025b).

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.

Table 1: Comparison of RabbitBench with existing exam-based occupational benchmarks in terms of occupational coverage, dataset size, and domain scope. Occupational coverage is measured by the numbers of mid-level and minor-level groups in the *Occupational Classification Dictionary of China*. Detailed qualification–category mappings for RabbitBench are provided in Appendix A.

Benchmark	Mid-level	Minor-level	Questions	Vertical Domain
AGIEval	5	10	8,062	Multiple
QualBench	4	8	17,316	Multiple
KMMLU-Pro/Redux	4	10	5,409	Multiple
GAOKAO-Bench	3	5	2,811	Multiple
CFinBench	1	9	99,100	Finance
CMB	1	7	280,839	Healthcare
JMed48k	1	5	48,862	Healthcare
EMPEC	1	5	157,803	Healthcare
MedBench	1	4	40,041	Healthcare
CMExam	1	4	68,119	Healthcare
MedQA	1	4	61,097	Healthcare
LawBench	1	4	10,000	Law
JEC-QA	1	4	26,365	Law
RabbitBench (ours)	8	23	26,723	Multiple

However, existing benchmarks still lack a comprehensive and robust evaluation of LLMs' occupational knowledge and technical competencies across a broad range of occupations. Current task-based benchmarks mainly evaluate whether models can independently execute specified workflows and produce digital deliverables (Drouin et al., 2024; Xu et al., 2025; Patwardhan et al., 2025; Sun et al., 2026). However, in some knowledge-intensive and regulation-bound occupational activities (e.g., interpreting occupational standards, assessing regulatory compliance), models may participate in occupational tasks by answering practitioners' questions, a form of occupational assistance that remains insufficiently represented in current task-based benchmarks. Moreover, since realistic tasks are costly to construct and grade, existing benchmarks can sample only a limited subset of occupational activities and settings, making their aggregate results sensitive to the particular cases selected (Shavelson et al., 1993; Carey, 1996; Patwardhan et al., 2025; Hua et al., 2026). In contrast, large collections of standardized examination questions, administered and scored under common procedures, can reduce this task-sampling and measurement-process variation (Shavelson et al., 1993; Carey, 1996; AERA, APA, and NCME, 2014). Although several benchmarks draw on professional examination questions, their coverage remains concentrated in a limited range of occupational fields (Hong et al., 2025a;b; Luo et al., 2025; Xun et al., 2026).

To address these limitations, we seek to bring together questions used to assess human candidates across a broad range of occupational qualifications and repurpose them for evaluating LLMs. Large question pools constructed according to defined examination syllabi can provide broader coverage of occupational knowledge and competencies, thereby supporting more robust evaluation. We therefore propose RabbitBench[1], an exam-based benchmark covering a diverse range of occupations. Following Chinese official guidance on entry-restricted qualifications[2], we collect and organize examination questions for 26 types of occupational qualifications, spanning 8 mid-level and 23 minor-level occupational categories in the Occupational Classification Dictionary of China[3]. We systematically quantify the coverage of occupational categories in this Dictionary and the number of questions, for our benchmark and related exam-based benchmarks in Table 1, which further demonstrates the breadth of occupational coverage in our benchmark. Many tasks in entry-restricted occupations involve public safety or human health; completing these tasks requires not only accomplishing the work itself but also complying with relevant national laws and regulations[2]. Our benchmark pro-

[1]The name is inspired by Miss Rabbit, a character in *Peppa Pig* who has many jobs around Peppatown; see https://peppapigworld.co.uk/pages/miss-rabbit.

[2]*Chinese National Occupational Qualification Catalogue*, https://www.mohrss.gov.cn/xxgk2020/fdzdgknr/zcfg/gfxwj/rcrs/202112/P020211202354560821450.pdf

[3]https://jg.class.com.cn/cms/resourcedetail.htm?contentUid=7ec1e51cbc4946f393f99ad6492d6b89

vides a unified platform for systematically evaluating LLM knowledge and capabilities, while also assessing whether models understand occupation-specific laws and regulations, thereby providing necessary data to support the governance of AI applications.

We evaluate a diverse set of proprietary and open-weight LLMs on RabbitBench. Overall, GPT-5.6-Sol and Gemini-3.1-Pro achieve the strongest performance, but incur substantially higher API costs than the other evaluated models. Qwen3.5 ranks behind these two models while requiring much lower API cost, demonstrating a favorable cost–performance trade-off. Across occupations, general model capability largely determines performance, with stronger models tending to perform well across a broad range of occupational domains, while occupation-specific strengths become more visible among models with similar overall performance. This broad transfer may partly reflect the general-purpose nature of the evaluated models. Within individual engineering qualifications, we further find a consistent gap between subject types: all evaluated models perform better on knowledge-oriented subjects than on application-oriented subjects, suggesting that applying occupational knowledge under complex and specialized conditions remains more challenging than understanding broadly applicable principles and regulations.

2 Related Work

We briefly review benchmarks for evaluating LLMs in occupational contexts. We introduce task-based and exam-based benchmarks in Sections 2.1 and 2.2 respectively.

2.1 Task-based occupational benchmarks

Task-based benchmarks evaluate models on concrete occupational tasks under settings that may vary in work activities, available tools and documents, assigned roles, workflow states, and deliverable requirements (Hua et al., 2026; Patwardhan et al., 2025). Existing benchmarks instantiate this approach in different ways. Interactive enterprise benchmarks such as WorkArena (Drouin et al., 2024), WorkArena++ (Boisvert et al., 2024), and TheAgentCompany (Xu et al., 2025) evaluate tool use, workflow execution, and multi-step task completion in simulated work environments. Several other benchmarks focus on economically valuable tasks and verifiable work products: SWE-Lancer (Miserendino et al., 2025) uses freelance software-engineering tasks from Upwork, while GDPval (Patwardhan et al., 2025) evaluates work-product tasks across a range of occupations. Recent efforts further broaden occupational coverage, including OccuBench (Hu et al., 2026) with simulated occupational scenarios and Agents' Last Exam (Sun et al., 2026) with long-horizon tasks and verifiable outcomes.

2.2 Occupational exam-based benchmarks

Occupational exam-based benchmarks apply questions from qualification examinations to evaluate models' occupation-related knowledge and technical competencies under standardized answer formats. Several studies construct such benchmarks by collecting examination questions from a single occupational domain. CMExam contains multiple-choice questions from the Chinese National Medical Licensing Examination, together with solution explanations and question-level annotations (Liu et al., 2023). JEC-QA constructs a legal question-answering benchmark from China's National Judicial Examination and distinguishes knowledge-driven from case-analysis questions (Zhong et al., 2020). Moving beyond single professional fields, QualBench broadens exam-based evaluation by collecting questions from multiple Chinese qualification examinations and organizing them into six dataset-specific vertical domains (Hong et al., 2025a), while KMMLU-Redux and KMMLU-Pro draw on Korean National Technical Qualification and National Professional Licensure examinations, respectively (Hong et al., 2025b). Although a gap remains between exam-based benchmarks and real occupational tasks, their standardized test items and evaluation protocols provide controlled and reproducible measurements for cross-model comparison (Hua et al., 2026; AERA, APA, and NCME, 2014). Nevertheless, the coverage of existing exam-based benchmarks is either concentrated in a single professional field or distributed across a limited set of occupational domains, restricting analyses of model strengths and weaknesses across a broader occupational landscape.

3 BENCHMARK CONSTRUCTION

This section presents the construction and composition of RabbitBench, including qualification selection, data preprocessing, and benchmark statistics. We summarize the main procedures here, while detailed descriptions of data sources, selection criteria, preprocessing, and quality control are provided in Appendix B.

3.1 DATA SOURCES

We collect questions from examinations for entry-restricted occupational qualifications listed in the National Occupational Qualification Catalogue of China. These examinations follow standardized syllabi and assessment requirements and are developed and reviewed by domain experts (Ministry of Human Resources and Social Security of the People's Republic of China, 2021).

Qualifications administered separately by level, service population, or specialization are counted as distinct qualifications, resulting in 46 candidates. We retain 26 qualifications with nationally standardized examinations, verifiable past examinations and answers, and questions suitable for text-based evaluation. For the retained qualifications, we collect available questions across examination subjects and years. Objective questions include single-choice, multiple-choice, and true-or-false questions with ground-truth answers, while the collected subjective questions are accompanied by high-quality reference answers. Detailed qualification and subject coverage is provided in Appendix A.

3.2 DATA PREPROCESSING

We transform the collected examination materials into structured question records through OCR-based extraction, structuring, and multi-stage validation. As available materials vary in completeness and quality, we consolidate multiple sources and prioritize clearer and more complete versions with corresponding answers or explanations. Source documents frequently consist of scanned PDFs with complex layouts, formulas, and tables. We use HunyuanOCR (Hunyuan Vision Team et al., 2025) to convert them into structured Markdown and subsequently parse the content into a unified JSON format. Quality control combines rule-based validation, LLM-assisted structural correction, and human review. Automated checks identify malformed records and residual document artifacts, while suspicious records undergo LLM-assisted correction and revalidation. As a final quality-control step, unresolved cases are manually reviewed against the original source PDFs and corrected where necessary, while a sample of records that pass automated validation is also inspected to verify extraction and structural accuracy. Detailed procedures are provided in Appendix B.

3.3 DATASET STATISTICS

RabbitBench contains 26,723 questions, including 25,519 objective questions and 1,204 subjective questions with reference answers. The objective questions comprise 20,498 single-choice questions, 4,646 multiple-choice questions, and 375 true-or-false questions. The subjective questions cover diverse formats, including short-answer, calculation, and case-analysis. The average lengths of single-choice, multiple-choice, true-or-false, and subjective questions are 101, 128, 42, and 611 characters, respectively. Question length is measured by the number of non-whitespace characters in the complete model-visible input, including shared materials or contextual information, the question stem, and answer options where applicable.

Figure 1 presents the occupational coverage of RabbitBench across eight mid-level occupational groups (MID) in the *Occupational Classification Dictionary of China*, together with the number of questions in each group. RabbitBench covers engineering, agriculture, healthcare, economics and finance, law, education, tourism and business services, and cultural services.

4 EXPERIMENTS

In this section, we evaluate LLMs on RabbitBench and analyze their performance at different levels. We first introduce the evaluated models and the scoring protocols for objective and subjective questions. We then report overall benchmark performance and further examine how model performance



Figure 1: Occupational coverage of RabbitBench across eight mid-level occupational groups (MID), with the number of questions in each group.

varies across occupational qualifications and their constituent subjects. For subjective questions, we additionally assess the robustness of the rubric-based evaluation using independently constructed scoring systems. We finally analyze the trade-off between model performance and API cost. A qualitative error analysis with representative question-level cases is provided in Appendix F.

4.1 Experimental Setup

Models. We evaluate eight LLMs covering both proprietary and open-weight frontier models. For proprietary models, we include GPT-5.6-Sol (OpenAI, 2026) and Gemini-3.1-Pro (Google, 2026), representing leading commercial model families from OpenAI and Google. For open-weight models, we include GLM-5.2 (Z.ai, 2026) and a series of Qwen models. In particular, we evaluate Qwen3.5 (Qwen Team, 2026) at four model scales—397B-A17B, 122B-A10B, 35B-A3B, and 9B—together with Qwen3-14B (Yang et al., 2025), allowing us to examine performance across a broad range of model sizes within a competitive open-weight model family. All models are evaluated on all 25,519 objective questions in RabbitBench. For subjective questions, which incur substantially higher generation and rubric-based evaluation costs, we evaluate a subset of 507 questions. We use a single response per question and enable the reasoning configuration supported by each model. For objective questions, we set the temperature to 0.5 and the maximum generation length to 16,384 tokens; for subjective questions, we use a temperature of 0.3 and a maximum generation length of 16,384 tokens.

Objective-question evaluation. We evaluate models in a zero-shot setting, providing only a format example to specify the required response structure and no solved demonstrations.

For objective questions, the model is prompted to first analyze the question and then provide its final answer within a designated `<FINAL_ANSWER>` field. We evaluate only the extracted final answer, allowing the evaluation to assess not only occupation-related knowledge and skills but also

Table 2: Comparison of model performance on RabbitBench across objective and subjective evaluations. Objective performance is measured by exact-match accuracy, while subjective performance is assessed using normalized rubric-based scores.

Model	Objective	Subjective
GPT-5.6-Sol	82.75	**73.81**
Gemini-3.1-Pro	**85.51**	68.26
GLM-5.2	81.84	61.38
Qwen3.5-397B	82.59	63.52
Qwen3.5-122B	79.74	62.31
Qwen3.5-35B	78.20	58.20
Qwen3-14B	71.60	47.75
Qwen3.5-9B	72.33	42.28

the model's ability to follow response instructions. For single-choice and true-or-false questions, a response is considered correct if its final answer matches the ground-truth answer. For multiple-choice questions, we use set-level exact match: a response is correct only when the predicted set of options exactly matches the ground-truth set, with no partial credit for missing or additional options. Responses that fail to produce a valid final answer within the prescribed generation budget are counted as incorrect. This evaluation therefore reflects a model's ability to solve the question and follow the required response protocol within a fixed inference budget. In subsequent analyses, objective performance is reported as accuracy over all relevant questions: across all objective questions for overall results and across all objective questions within each occupational qualification for occupation-level results.

Subjective-question evaluation. Each subjective question is associated with a high-quality reference answer. To enable fine-grained and interpretable scoring of subjective responses, we apply an LLM to derive atomic, binary criteria from each reference answer, following the criterion-level rubric design of PaperBench (Starace et al., 2025). Each criterion specifies an observable piece of content, reasoning step, judgment, or conclusion expected in the response. The rubric-generation LLM is instructed to avoid redundant criteria and assigns a point value to each criterion, with the total score for every rubric fixed at 10 points. Given a rubric for each question i, the judge LLM independently determines whether the model response satisfies each criterion. Let $z_{ik} \in 0, 1$ indicate whether criterion k is satisfied and w_{ik} denote its assigned score. The subjective score for each question i is normalized to $[0, 1]$ by dividing the earned rubric points by 10:

$$s_i^{\text{sub}} = \frac{1}{10} \sum_{k=1}^{K_i} w_{ik} z_{ik}, \qquad \sum_{k=1}^{K_i} w_{ik} = 10. \tag{1}$$

The overall subjective score is the average score across all selected subjective questions.

Unless otherwise specified, subjective performance is evaluated using GPT-5.6-Sol for both rubric generation and judging.

Multi-Judge Consistency. We evaluate subjective responses using three rubric-and-judge systems based on GPT-5.6-Sol, GLM-5.2, and Gemini-3.1-Pro, with each model generating its own rubrics and judging responses accordingly. The three systems show strong agreement at the question level, with pairwise Pearson correlations ranging from 0.852 to 0.866 and Spearman correlations from 0.845 to 0.870. They also produce highly consistent overall model rankings, with pairwise Spearman correlations between the eight-model rankings ranging from 0.976 to 1.000. Absolute scores are also broadly comparable across systems, despite modest systematic differences in scoring levels. These results indicate that subjective evaluation is stable across different rubric-and-judge models. Detailed results are provided in Appendix C.

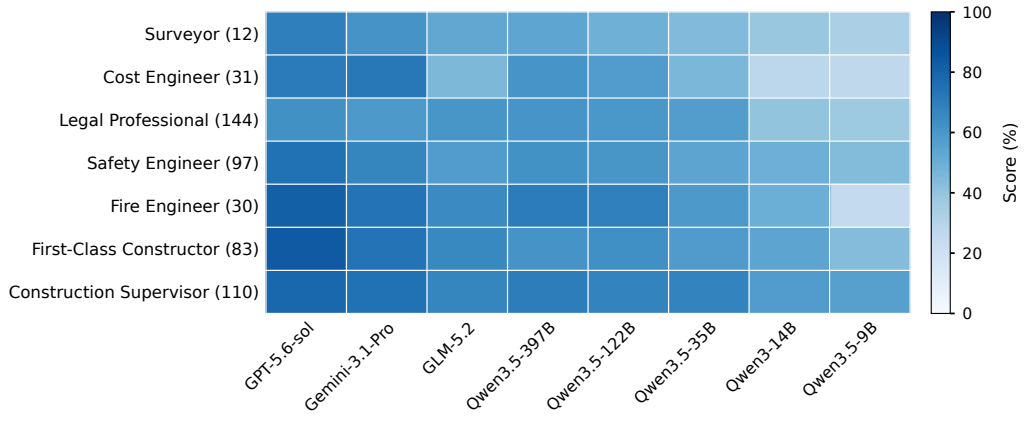


Figure 2: Heatmap of subjective model performance across seven occupational qualifications in RabbitBench. Numbers in parentheses indicate the number of subjective questions in each qualification, and color intensity represents the score level.

4.2 Overall Performance

Table 2 summarizes the overall performance of the eight evaluated models. We report objective and subjective scores separately due to different response formats and scoring mechanisms.

Several observations emerge from the overall results. First, the two proprietary frontier models achieve the strongest overall performance. Gemini-3.1-Pro obtains the highest objective accuracy at 85.51%, while GPT-5.6-Sol achieves the highest subjective score at 73.81%. Both models outperform all evaluated open-weight models on both objective and subjective evaluations. The strong objective performance of Gemini-3.1-Pro reflects its broad coverage of factual and occupational knowledge, as the objective questions span a wide range of occupation-specific knowledge and relevant laws and regulations. GPT-5.6-Sol, in contrast, performs best on subjective questions, suggesting stronger capabilities in multi-step reasoning, integrating relevant information, and constructing complete and well-structured responses that cover multiple occupational requirements.

Among open-weight models, Qwen3.5-397B achieves the highest scores on both objective and subjective questions, reaching 82.59% and 63.52%, respectively, followed by GLM-5.2 on objective questions and Qwen3.5-122B on subjective questions. Within the Qwen3.5 family, larger models consistently achieve higher scores on both forms of evaluation. Objective accuracy increases from 72.33% for Qwen3.5-9B to 78.20%, 79.74%, and 82.59% for Qwen3.5-35B, Qwen3.5-122B, and Qwen3.5-397B, respectively. A similar trend appears on subjective questions, where the corresponding scores increase from 42.28% to 58.20%, 62.31%, and 63.52%. From the perspective of question type, increasing model scale yields a substantially larger improvement on subjective questions than on objective questions.

4.3 Performance across Occupations

Overall performance can obscure distinct performance patterns across occupational domains.

In this part we examine whether occupational performance is primarily determined by a model's general capability or exhibits substantial occupation-specific specialization.

The subjective occupation-level results are shown in Figure 2, while the corresponding objective results are provided in Figure 4 in Appendix D. These results reveal a two-level pattern: general capability largely determines model performance across occupations, while domain-specific strengths and weaknesses become more apparent among models with similar overall performance.

First, improvements in general model capability tend to transfer broadly across occupational domains rather than being concentrated in a few specific fields. On objective questions, Gemini-3.1-

Table 3: Performance on knowledge- and application-oriented subjects across seven engineering qualifications, averaged across eight models. Knowledge and Application are the mean accuracies (%) for the two subject categories; Δ(pp) is Knowledge minus Application in percentage points; and #($K > A$) is the number of models scoring higher on knowledge-oriented subjects. Full qualification names and subject mappings are provided in Appendix E.

Qualification	**Knowledge (%)**	**Application (%)**	**Δ (pp)**	**#($K > A$)**
First-Class Constructor	88.04	72.56	15.48	8/8
Structural Engineer	78.40	38.67	39.73	8/8
Public Utilities Engineer	88.96	55.22	33.73	8/8
Chemical Engineer	83.11	76.90	6.21	8/8
Civil Engineer	79.69	46.44	33.25	8/8
Environmental Engineer	86.09	74.37	11.72	8/8
Electrical Engineer	77.82	42.00	35.82	8/8

Pro achieves the highest score in 20 of the 26 qualifications and outperforms GPT-5.6-Sol in 23, indicating that its overall advantage is supported by consistently strong performance across diverse bodies of occupational knowledge and relevant regulations. A similar pattern appears for subjective questions: GPT-5.6-Sol achieves the highest score in six of the seven evaluated qualifications and outperforms Gemini-3.1-Pro in the same six. This suggests that its stronger constructed-response capability generalizes across substantially different occupational domains rather than arising from strength in a particular occupation. The same pattern is also evident within the Qwen3.5 family. Qwen3.5-397B outperforms Qwen3.5-122B on all 26 qualifications for objective questions and on six of the seven qualifications for subjective questions, while Qwen3.5-122B similarly outperforms Qwen3.5-35B on 23 of 26 and six of seven qualifications, respectively. These results suggest that improvements in model capability are largely transferable across occupational domains.

Second, occupation-specific differences become more visible when models have similar overall performance. Although Qwen3.5-397B slightly outperforms GLM-5.2 in overall objective accuracy, GLM-5.2 achieves higher objective accuracy in 15 qualifications, while Qwen3.5-397B performs better in 11. Their relative advantages are concentrated in different occupational domains: GLM-5.2 shows stronger performance in several engineering and economic or financial domains, whereas Qwen3.5-397B performs better in legal and some construction-related engineering domains. An even closer split is observed between GPT-5.6-Sol and Qwen3.5-397B, whose overall objective accuracies are 82.75% and 82.59%, respectively. GPT-5.6-Sol shows relative advantages in legal, construction-management, and safety-related domains, whereas Qwen3.5-397B is stronger in engineering design, utilities, chemical engineering, and surveying. Thus, similar aggregate performance can arise from distinct occupational performance profiles, suggesting that domain-specific knowledge coverage becomes an important source of differentiation once models reach comparable overall capability.

4.4 Performance across Subjects within Individual Qualifications

We further compare exam subjects within individual qualifications, focusing on engineering examinations that distinguish knowledge-oriented subjects from application-oriented subjects. Knowledge-oriented subjects primarily assess mastery of broadly applicable principles, formulas, and regulations, whereas application-oriented subjects require these principles and standards to be applied under concrete engineering conditions, often involving domain-specific rules, multiple constraints, and multi-step calculations or decisions.

Table 3 presents the performance gap between the two subject categories. The detailed mapping of individual exam subjects to these categories is provided in Appendix E and Table 9. A consistent performance gap emerges between the two types of subjects. Across the seven engineering qualifications, all eight evaluated models achieve higher accuracy on the knowledge-oriented subjects. The differences are particularly pronounced in structural, building-services, civil, and electrical engineering, where the average gaps exceed 30 percentage points. Chemical engineering shows a much smaller gap, indicating that application-oriented subjects are not inherently more difficult in every field. Overall, these results suggest that current models are more reliable at mastering broadly applicable professional knowledge than at applying it in complex and specialized engineering contexts.

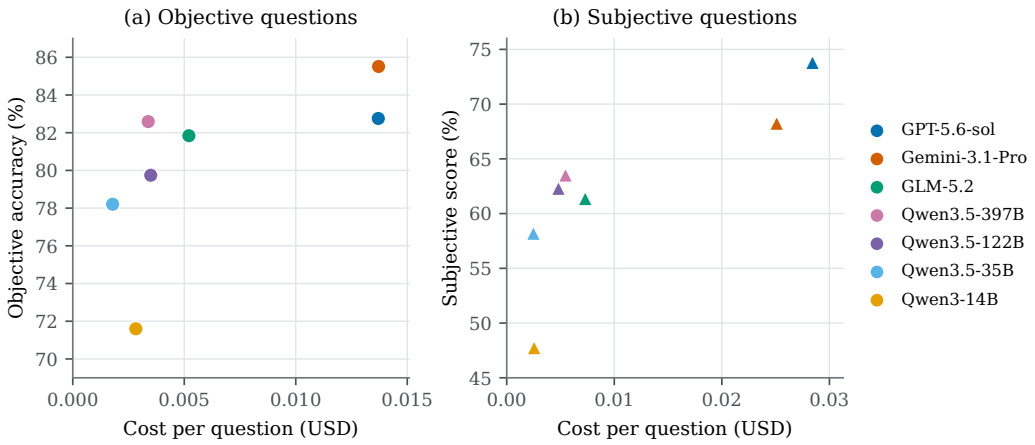


Figure 3: Cost–performance trade-offs of evaluated models on RabbitBench. Panel (a) plots objective accuracy against estimated API cost per question, and panel (b) plots subjective score against estimated API cost per question. Cost estimates are based on official list prices; for subjective questions, they include answer generation only and exclude rubric generation and judging.

4.5 Performance–Cost Trade-offs

For each question, we estimate API usage cost from its input and output token usage using official list prices, and plot cost against overall performance in Figure 3. On both objective and subjective evaluations, GPT-5.6-Sol and Gemini-3.1-Pro incur substantially higher API costs than the other models while also achieving the strongest performance. In contrast, Qwen3.5-397B achieves competitive performance at a much lower cost, ranking behind only GPT-5.6-Sol and Gemini-3.1-Pro on both evaluations. This indicates a favorable cost–performance trade-off for Qwen3.5-397B.

The cost–performance trade-off also differs between evaluation formats. On objective questions, performance differences among stronger models are relatively small despite large cost differences, suggesting a flattening efficiency frontier. In contrast, subjective performance remains more clearly separated at the high-cost end, where GPT-5.6-Sol achieves a substantial advantage. These results indicate that model selection cannot be determined by either performance or API cost alone: the efficient trade-off between the two depends on whether the target workload emphasizes constrained objective answering or constructed-response capability.

5 Conclusion

We introduce RabbitBench, a large-scale benchmark for evaluating LLMs across diverse occupations using Chinese occupational qualification examinations. RabbitBench contains 26,723 questions from 26 qualification types, spanning 8 mid-level and 23 minor-level occupational groups. By combining objective and subjective questions with broad occupational coverage, RabbitBench supports evaluation at both aggregate and occupation-specific levels.

Our evaluation of eight LLMs reveals several consistent patterns. GPT-5.6-Sol and Gemini-3.1-Pro achieve the strongest overall performance but at substantially higher API costs, while Qwen3.5-397B offers a favorable cost–performance trade-off. Across occupations, stronger models generally perform well across domains, while similarly capable models exhibit distinct occupational strengths. Within engineering qualifications, all models perform better on knowledge-oriented than application-oriented subjects, suggesting that applying occupational knowledge under specialized conditions remains more challenging than understanding broadly applicable principles and regulations. Overall, RabbitBench enables broad and fine-grained analysis of LLM occupational capabilities beyond aggregate scores.

REFERENCES

AERA, APA, and NCME. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC, 2014. URL https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. WorkArena++: Towards compositional planning and reasoning-based common knowledge work tasks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0195. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0b82662b6c32e887bb252a74d8cb2d5e-Abstract-Datasets_and_Benchmarks_Track.html.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL https://academic.oup.com/qje/article/140/2/889/7990658.

Patricia A. Carey. A review of psychometric and consequential issues related to performance assessment. Technical Report RM-96-03, Educational Testing Service, 1996. URL https://www.ets.org/research/policy_research_reports/publications/report/1996/ibzu.html.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662. PMLR, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Google. Gemini 3.1 Pro Preview, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview.

Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. QualBench: Benchmarking Chinese LLMs with localized professional qualifications for vertical domain evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5938–5953. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.303. URL https://aclanthology.org/2025.emnlp-main.303/.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. From KMMLU-Redux to KMMLU-Pro: A professional Korean benchmark suite for LLM evaluation. *arXiv preprint arXiv:2507.08924*, 2025b. URL https://arxiv.org/abs/2507.08924.

Xiaomeng Hu, Yinger Zhang, Fei Huang, Jianhong Tu, Yang Su, Lianghao Deng, Yuxuan Liu, Yantao Liu, Dayiheng Liu, and Tsung-Yi Ho. OccuBench: Evaluating AI agents on real-world professional tasks via language world models. *arXiv preprint arXiv:2604.10866*, 2026. doi: 10.48550/arXiv.2604.10866. URL https://arxiv.org/abs/2604.10866.

Yining Hua, Hongbin Na, Cyrus Ayubcha, and Levi Lian. Designing benchmarks for knowledge work. *arXiv preprint arXiv:2605.23262*, 2026. doi: 10.48550/arXiv.2605.23262. URL https://arxiv.org/abs/2605.23262.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, Qi Yang, Qiming Peng, Bin Luo, Hower Yang, Xinsong Zhang, Jinnian Zhang, Houwen Peng, Hongming Yang, Senhao Xie, Longsha Zhou, Ge Pei, Binghong Wu, Rui Yan, Kan Wu, Jieneng Yang, Bochao Wang, Kai Liu, Jianchen Zhu, Jie Jiang, Linus, Han Hu, and Chengquan Zhang. HunyuanOCR Technical Report. *arXiv preprint arXiv:2511.19575*, 2025. doi: 10.48550/arXiv.2511.19575. URL https://arxiv.org/abs/2511.19575.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. Benchmarking large language models on CMExam: A comprehensive Chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a48ad12d588c597f4725a8b84af647b5-Paper-Datasets_and_Benchmarks.pdf.

Zheheng Luo, Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. EMPEC: A comprehensive benchmark for evaluating large language models across diverse healthcare professions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9945–9958. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.518. URL https://aclanthology.org/2025.findings-acl.518/.

Ministry of Human Resources and Social Security of the People's Republic of China. Regulations on examination administration for professional and technical personnel qualifications, 2021. URL https://www.mohrss.gov.cn/xxgk2020/1_2_x/jd/202103/t20210330_412029.html. MOHRSS General Office Document No. 18 [2021].

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44412–44450. PMLR, 2025. URL https://proceedings.mlr.press/v267/miserendino25a.html.

OpenAI. GPT-5.6 Sol model, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025. URL https://arxiv.org/abs/2510.04374.

Qwen Team. Qwen3.5 model collection, 2026. URL https://huggingface.co/collections/Qwen/qwen35.

Richard J. Shavelson, Gail P. Baxter, and Xiaohong Gao. Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3):215–232, 1993. doi: 10.1111/j.1745-3984.1993.tb00424.x.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025. URL https://arxiv.org/abs/2504.01848.

Yiyou Sun et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026. doi: 10.48550/arXiv.2606.05405. URL https://arxiv.org/abs/2606.05405.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Melroy Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=LZnKNApvhG.

Yue Xun, Junyu Liu, Qian Niu, Xinyi Wang, Zheng Yuan, Zirui Li, Zequn Zhang, Bowen Zhao, Shujun Wang, Irene Li, Kan Hatakeyama-Sato, Yusuke Iwasawa, and Yutaka Matsuo. JMed48k: A multi-profession Japanese medical licensing benchmark for vision-language model evaluation. *arXiv preprint arXiv:2605.22080*, 2026. doi: 10.48550/arXiv.2605.22080. URL https://arxiv.org/abs/2605.22080.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qianyu Yang, Yang Liu, Jiaqi Li, Jun Bai, Hao Chen, Kaiyuan Chen, Tiliang Duan, Jiayun Dong, Xiaobo Hu, Zixia Jia, Yang Liu, Tao Peng, Yixin Ren, Ran Tian, Zaiyuan Wang, Yanglihong Xiao, Gang Yao, Lingyue Yin, Ge Zhang, Chun Zhang, Jianpeng Jiao, Zilong Zheng, and Yuan Gong. $OneMillion-Bench: How far are language agents from human experts? *arXiv preprint arXiv:2603.07980*, 2026. doi: 10.48550/arXiv.2603.07980. URL https://arxiv.org/abs/2603.07980.

Z.ai. GLM-5.2 model card, 2026. URL https://huggingface.co/zai-org/GLM-5.2.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9701–9708, 2020. doi: 10.1609/aaai.v34i05.6519.

A QUALIFICATION AND SUBJECT COVERAGE

This section documents RabbitBench's coverage at the occupational-qualification and examination-subject levels.

Table 4 lists the 26 qualifications and maps each one to its mid-level and minor occupational categories. Table 5 lists the examination subjects retained for each qualification.

Table 4: Occupational classification of the 26 qualification types included in RabbitBench. MID and Minor categories follow the *Occupational Classification Dictionary of China*.

Qualification Name	Qualification Name (English)	Category (MID)	Category (Minor)
注册测绘师资格考试	Registered Surveyor Qualification Examination	Engineering Technical Personnel	Surveying and Geographic Information Engineering Technical Personnel
注册化工工程师资格考试	Registered Chemical Engineer Qualification Examination	Engineering Technical Personnel	Chemical Engineering Technical Personnel
注册电气工程师资格考试	Registered Electrical Engineer Qualification Examination	Engineering Technical Personnel	Electric Power Engineering Technical Personnel
一级注册结构工程师资格考试	First-Class Registered Structural Engineer Qualification Examination	Engineering Technical Personnel	Building Engineering Technical Personnel
二级注册建筑师资格考试	Second-Class Registered Architect Qualification Examination	Engineering Technical Personnel	Building Engineering Technical Personnel
注册公用设备工程师资格考试	Registered Public Utilities Engineer Qualification Examination	Engineering Technical Personnel	Building Engineering Technical Personnel
注册土木工程师资格考试	Registered Civil Engineer Qualification Examination	Engineering Technical Personnel	Building Engineering Technical Personnel
注册核安全工程师执业资格考试	Registered Nuclear Safety Engineer Qualification Examination	Engineering Technical Personnel	Environmental Protection Engineering Technical Personnel
注册环保工程师执业资格考试	Registered Environmental Engineer Qualification Examination	Engineering Technical Personnel	Environmental Protection Engineering Technical Personnel
中级注册安全工程师职业资格考试	Intermediate Registered Safety Engineer Qualification Examination	Engineering Technical Personnel	Safety Engineering Technical Personnel
一级注册消防工程师资格考试	First-Class Registered Fire Engineer Qualification Examination	Engineering Technical Personnel	Safety Engineering Technical Personnel
二级注册计量师职业资格考试	Second-Class Registered Metrologist Qualification Examination	Engineering Technical Personnel	Standardization, Metrology, Quality, and Certification Technical Personnel
一级建造师执业资格考试	First-Class Constructor Qualification Examination	Engineering Technical Personnel	Management (Industrial) Engineering Technical Personnel
一级造价工程师职业资格考试	First-Class Cost Engineer Qualification Examination	Engineering Technical Personnel	Management (Industrial) Engineering Technical Personnel
监理工程师职业资格考试	Construction Supervisor Qualification Examination	Engineering Technical Personnel	Management (Industrial) Engineering Technical Personnel
注册城乡规划师职业资格考试	Registered Urban-Rural Planner Qualification Examination	Engineering Technical Personnel	Territorial Spatial Planning and Ecological Restoration Engineering Technical Personnel
全国执业兽医资格考试	National Veterinary Practitioner Qualification Examination	Agricultural Technical Personnel	Veterinary and Veterinary Pharmaceutical Technical Personnel
医师资格考试（中医执业医师）	Traditional Chinese Medicine Physician Qualification Examination	Health Professionals	Traditional Chinese Medicine Physicians
执业药师职业资格考试	Licensed Pharmacist Qualification Examination	Health Professionals	Pharmaceutical Technical Personnel
护士执业资格考试	Nurse Qualification Examination	Health Professionals	Nursing Personnel
注册会计师全国统一考试	National Certified Public Accountant Examination	Economic and Financial Professionals	Accountants; Auditors
专利代理师资格考试	Patent Attorney Qualification Examination	Economic and Financial Professionals	Intellectual Property Professionals

Continued on next page

Table 4 – continued from previous page

Qualification Name	Qualification Name (English)	Category (MID)	Category (Minor)
国家统一法律职业资格考试	National Unified Legal Professional Qualification Examination	Legal, Social, and Religious Professionals	Judges; Prosecutors; Lawyers; Notaries
小学教师资格考试	Primary School Teacher Qualification Examination	Teaching Professionals	Primary and Secondary School Teachers
全国导游资格考试	National Tour Guide Qualification Examination	Leasing and Business Service Personnel	Tourism and Public Recreation Service Personnel
演出经纪人员资格认定考试	Performing Arts Agent Qualification Examination	Cultural and Educational Service Personnel	Social and Cultural Activity Service Personnel

Table 5: Subjects included for each qualification type in RabbitBench. Objective and subjective records from the same examination subject are listed once.

Qualification Name	Qualification Name (English)	Short Name	Included Subjects
一级建造师执业资格考试	First-Class Constructor Qualification Examination	First-Class Constructor	Construction Engineering Economics; Construction Engineering Regulations and Related Knowledge; Construction Project Management; Highway, Municipal Public, Building, Mechanical and Electrical, and Water Resources and Hydropower Engineering Management and Practice
一级注册消防工程师资格考试	First-Class Registered Fire Engineer Qualification Examination	Fire Engineer	Fire Safety Technical Practice; Fire Safety Technical Comprehensive Ability; Fire Safety Case Analysis
一级注册结构工程师资格考试	First-Class Registered Structural Engineer Qualification Examination	Structural Engineer	Shared Public Foundation*; Professional Foundation; Professional Examination
一级造价工程师职业资格考试	First-Class Cost Engineer Qualification Examination	Cost Engineer	Construction Engineering Cost Management; Construction Engineering Pricing; Construction Engineering Technology and Measurement (Civil Engineering and Installation Engineering); Construction Engineering Cost Case Analysis
专利代理师资格考试	Patent Attorney Qualification Examination	Patent Attorney	Patent Law; Related Laws
中级注册安全工程师职业资格考试	Intermediate Registered Safety Engineer Qualification Examination	Safety Engineer	Other Safety Professional Practice; Chemical Safety Professional Practice; Construction Safety Professional Practice
二级注册建筑师资格考试	Second-Class Registered Architect Qualification Examination	Second-Class Architect	Architectural Design, Materials, and Construction; Building Structures, Building Physics, and Equipment; Construction Economics, Construction, and Design Practice Management
二级注册计量师职业资格考试	Second-Class Registered Metrologist Qualification Examination	Metrologist	Metrology Laws, Regulations, and Comprehensive Knowledge; Metrology Professional Practice and Case Analysis
全国导游资格考试	National Tour Guide Qualification Examination	Tour Guide	Policies and Laws and Regulations; Tour Guide Business; National Tour Guide Basic Knowledge; Local Tour Guide Basic Knowledge
全国执业兽医资格考试	National Veterinary Practitioner Qualification Examination	Veterinary Practitioner	Veterinary General Examination
医师资格考试（中医执业医师）	Traditional Chinese Medicine Physician Qualification Examination	TCM Physician	Medical Comprehensive Examination
国家统一法律职业资格考试	National Unified Legal Professional Qualification Examination	Legal Professional	Comprehensive Subject; Legal Professional Qualification Subjective Questions; Criminal Law Subjective Questions
小学教师资格考试	Primary School Teacher Qualification Examination	Primary School Teacher	Comprehensive Quality (Primary School); Educational Teaching Knowledge and Ability (Primary School)
执业药师职业资格考试	Licensed Pharmacist Qualification Examination	Pharmacist	Pharmacy Professional Knowledge I and II; Comprehensive Pharmacy Knowledge and Skills; Traditional Chinese Pharmacy Professional Knowledge I and II; Comprehensive Traditional Chinese Pharmacy Knowledge and Skills; Pharmacy Administration and Regulations

Continued on next page

Table 5 – continued from previous page

Qualification Name	Qualification Name (English)	Short Name	Included Subjects
护士执业资格考试	Nurse Qualification Examination	Nurse	Professional Practice; Practical Ability
注册会计师全国统一考试	National Certified Public Accountant Examination	Certified Public Accountant	Corporate Strategy and Risk Management; Auditing; Tax Law; Economic Law; Financial Cost Management
注册公用设备工程师资格考试	Registered Public Utilities Engineer Qualification Examination	Public Utilities Engineer	Shared Public Foundation*; Professional Foundation; Professional Knowledge; Professional Case Analysis
注册化工工程师资格考试	Registered Chemical Engineer Qualification Examination	Chemical Engineer	Shared Public Foundation*; Professional Foundation; Professional Knowledge; Professional Case Analysis
注册土木工程师资格考试	Registered Civil Engineer Qualification Examination	Civil Engineer	Shared Public Foundation*; Geotechnical Professional Foundation, Professional Knowledge, and Professional Case Analysis; Road Engineering Professional Foundation, Professional Knowledge, and Professional Case Analysis
注册城乡规划师职业资格考试	Registered Urban-Rural Planner Qualification Examination	Urban-Rural Planner	Urban-Rural Planning Principles; Related Urban-Rural Planning Knowledge; Urban-Rural Planning Management and Regulations
注册核安全工程师执业资格考试	Registered Nuclear Safety Engineer Qualification Examination	Nuclear Safety Engineer	Nuclear Safety Professional Practice; Nuclear Safety Laws and Regulations; Comprehensive Nuclear Safety Knowledge
注册测绘师资格考试	Registered Surveyor Qualification Examination	Surveyor	Surveying Comprehensive Ability; Surveying Management and Laws and Regulations; Surveying Case Analysis
注册环保工程师执业资格考试	Registered Environmental Engineer Qualification Examination	Environmental Engineer	Shared Public Foundation*; Professional Foundation; Professional Knowledge; Professional Case Analysis
注册电气工程师资格考试	Registered Electrical Engineer Qualification Examination	Electrical Engineer	Shared Public Foundation*; Professional Foundation; Professional Case Analysis
演出经纪人员资格认定考试	Performing Arts Agent Qualification Examination	Performing Arts Agent	Ideological and Political Theory and Legal Fundamentals; Performing Arts Market Policies and Brokerage Practice
监理工程师职业资格考试	Construction Supervisor Qualification Examination	Construction Supervisor	Construction Project Contract Management; Basic Theory of Construction Supervision and Related Regulations; Construction Project Target Control; Construction Supervision Case Analysis (Civil Engineering and Water Resources)

Note. *The same pool of 931 Shared Public Foundation questions is associated with the First-Class Registered Structural Engineer, Registered Public Utilities Engineer, Registered Chemical Engineer, Registered Civil Engineer, Registered Environmental Engineer, and Registered Electrical Engineer qualifications. These questions are counted only once in benchmark-level totals.

B Data Sources and Preprocessing

B.1 Data Sources

We use questions from examinations for entry-restricted occupational qualifications listed in the *National Occupational Qualification Catalogue of China* as the primary data source for RabbitBench. These qualifications span a broad range of knowledge-intensive occupations and are established on a legal or regulatory basis, with standardized requirements for professional practice. Their examinations are systematically designed around occupational requirements: relevant authorities translate occupational tasks into examination syllabi that specify the knowledge and skills to be assessed, after which domain experts develop and review examination questions against these specifications (Ministry of Human Resources and Social Security of the People's Republic of China, 2021). Many examinations also assess knowledge of domain-specific laws and regulations, providing an additional dimension for evaluating whether models can operate under explicit professional rules and constraints. Together, these properties make entry-restricted qualification examinations a suitable basis for evaluating occupational knowledge and professional problem solving across diverse domains.

We treat qualifications that are separately administered by level, service population, or specialization as distinct qualification types, resulting in 46 candidates from the official catalogue. We retain 26 qualification types according to four criteria:

1. **National standardization**: standardized examination syllabi, question setting, and scoring procedures;
2. **Verifiable past examinations**: past examination papers and corresponding answers are publicly available through traceable sources;
3. **Multi-year coverage**: sufficient past examinations are available for the major subjects of the qualification; and
4. **Suitability for text-based evaluation**: the question content and expected answers can be faithfully represented and evaluated in text.

Accordingly, we exclude examinations that are locally administered without nationally standardized papers, rely on confidential question banks without verifiable past examinations, or primarily involve practical tests, interviews, or other components that cannot be reliably represented in text.

For the retained qualifications, we collect available questions satisfying the text-based evaluation criterion across subjects and years. The resulting benchmark covers both objective and subjective questions. Objective questions include single-choice, multiple-choice, and true-or-false questions, each paired with a ground-truth answer. The collected subjective questions are accompanied by high-quality reference answers, enabling RabbitBench to evaluate constructed responses in addition to objective questions.

B.2 Data Preprocessing

We transform the collected examination materials into structured question records through three stages: **examination collection**, **question extraction and structuring**, and **inspection and validation**. Throughout the pipeline, we rely primarily on rule-based processing, while using LLMs to assist with structurally anomalous cases.

Examination collection. Past examinations for entry-restricted occupational qualifications are not distributed through a unified official repository, and publicly available materials vary substantially in completeness and quality. We therefore collect examination materials from three main sources: published compilations of past examination papers, electronic preparation materials distributed through e-commerce platforms, and downloadable resources from examination-material and question-bank websites. To avoid relying on a single source, we consolidate materials across multiple sources. When multiple versions of the same examination are available, we prioritize versions that are clearer and more complete and that contain answers or explanations, while using complementary sources to fill missing subjects or years.

Question extraction and structuring. Most collected examination papers are distributed as scanned PDFs and frequently contain complex layouts, mathematical formulas, tables, and other structures that cannot be reliably recovered through direct text extraction. We therefore use HunyuanOCR to convert each PDF page into structured Markdown, preserving textual content and structural elements such as mathematical formulas and tables.

Since the Markdown structure varies across examination types, we use a parsing script to convert the different formats into a unified JSON representation.

Inspection and validation. We apply a multi-stage quality-control procedure combining rule-based validation, LLM-assisted correction, and human review. First, rule-based validation addresses formatting-related issues by checking for structural anomalies, e.g., missing answers or incomplete options. It also removes irrelevant content introduced by the source documents, such as residual headers, footers, watermarks, and promotional text.

Records or examinations that fail these checks, or exhibit suspicious structures, are then passed to an LLM for structural correction. The LLM is used to reorganize source-derived content, for example, separating merged options, realigning answers and explanations. All corrected records are subsequently passed through rule-based validation again.

Finally, three members of our team manually review unresolved cases against the original source PDFs and sample-check records that pass the automated validation procedure. When discrepancies are identified, minor issues are corrected manually, whereas substantial issues trigger reprocessing of the corresponding examination or replacement of the source PDF followed by re-extraction.

C MULTI-JUDGE CONSISTENCY

We compare three scoring systems, each using GPT-5.6-Sol, GLM-5.2, or Gemini-3.1-Pro for both rubric generation and judging. Table 6 compares the mean subjective scores of each answer model under the three scoring systems. The absolute scores are broadly consistent across systems: although Gemini tends to assign slightly higher scores and GLM slightly lower scores, the range for any answer model is at most 8.4 percentage points. Table 7 reports pairwise Pearson and Spearman correlations of question-level scores, together with the Spearman correlation of model-level mean scores. The consistently high correlations indicate strong agreement between systems at both the individual-question and overall-model levels. Table 8 reports pairwise Spearman correlations of model rankings within each qualification. The generally high correlations show that the relative ordering of models is also largely preserved across occupational qualifications. Together, these results indicate that subjective evaluation is stable across different rubric-and-judge models in both absolute scores and relative model rankings.

Table 6: Mean subjective scores (%) under three rubric-and-judge systems. Each column uses the named model for both rubric generation and judging.

Answer model	**GPT**	**GLM**	**Gemini**
GPT-5.6-Sol	73.81	71.40	77.11
Gemini-3.1-Pro	68.26	65.16	72.22
GLM-5.2	61.38	59.66	66.26
Qwen3.5-397B	63.52	60.14	68.03
Qwen3.5-122B	62.31	58.46	66.35
Qwen3.5-35B	58.20	54.35	61.81
Qwen3-14B	47.75	42.81	51.21
Qwen3.5-9B	42.28	40.80	46.65

Table 7: Pairwise consistency between scoring systems. r_q and ρ_q are Pearson and Spearman correlations; ρ_m is the Spearman correlation of the eight model-level mean scores.

Systems	r_q	ρ_q	ρ_m
GPT–GLM	0.857	0.856	0.976
GPT–Gemini	0.852	0.845	1.000
GLM–Gemini	0.866	0.870	0.976

Table 8: Spearman correlations of the eight model rankings within each qualification under the three scoring systems. N is the number of subjective questions in that qualification.

Qualification	N	**GPT–GLM**	**GPT–Gemini**	**GLM–Gemini**
First-Class Constructor	83	1.000	1.000	1.000
Fire Engineer	30	1.000	1.000	1.000
Cost Engineer	31	0.929	0.857	0.929
Safety Engineer	97	1.000	1.000	1.000
Legal Professional	144	0.905	0.905	0.952
Surveyor	12	0.905	0.833	0.810
Construction Supervisor	110	0.833	0.810	0.976

D OBJECTIVE PERFORMANCE ACROSS OCCUPATIONAL QUALIFICATIONS

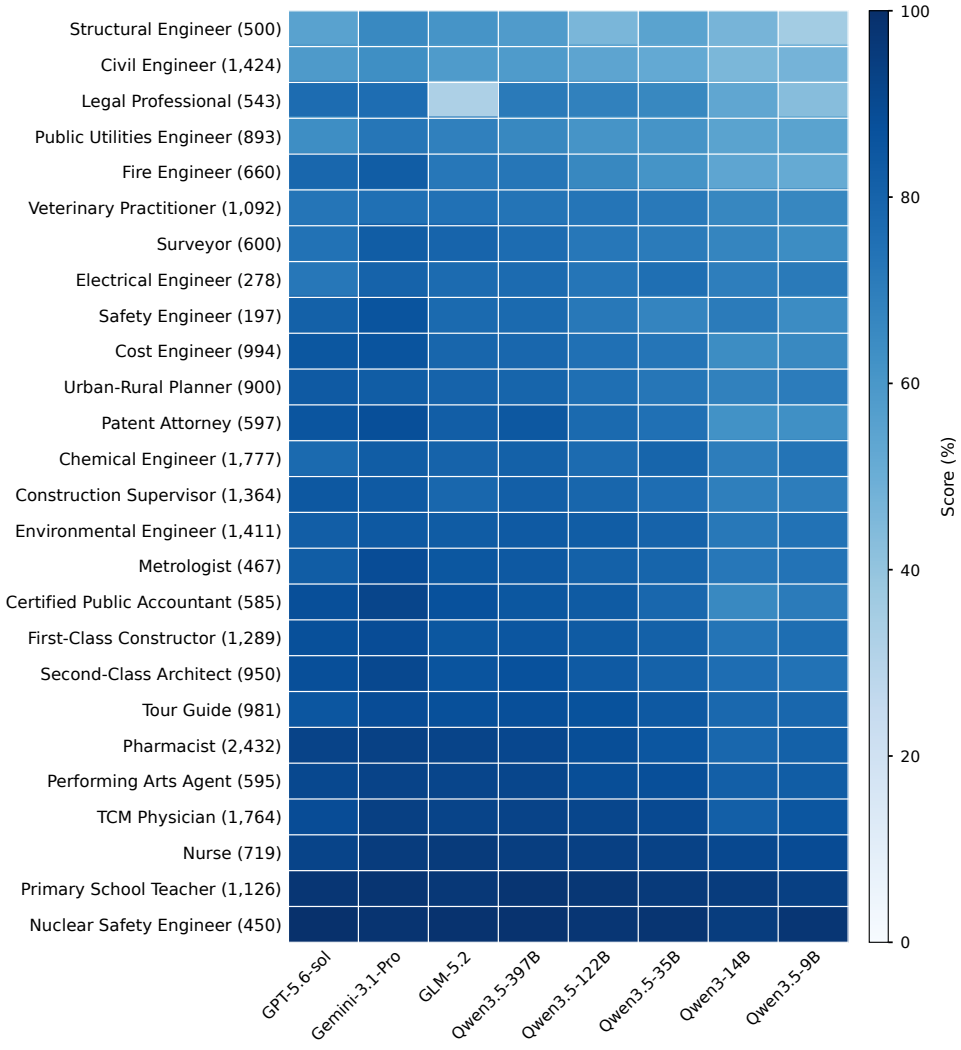


Figure 4: Heatmap of objective model performance across 26 occupational qualifications in RabbitBench. Numbers in parentheses indicate the number of objective questions in each qualification, and color intensity represents accuracy.

E SUBJECT CATEGORIZATION WITHIN ENGINEERING QUALIFICATIONS

Table 9: Mapping of exam subjects to the knowledge-oriented and application-oriented categories used in Table 3. Knowledge-oriented subjects primarily assess understanding of broadly applicable occupational principles, formulas, and regulations, whereas application-oriented subjects require these principles and standards to be applied under concrete engineering conditions.

Qualification	**Knowledge-oriented subjects**	**Application-oriented subjects**
First-Class Constructor	Construction Engineering Economics; Construction Engineering Laws and Regulations; Construction Project Management	Highway Engineering Management and Practice; Municipal Engineering Management and Practice; Building Engineering Management and Practice; Mechanical and Electrical Engineering Management and Practice; Water Resources and Hydropower Engineering Management and Practice
Structural Engineer	Occupational Fundamentals	Occupational Examination
Public Utilities Engineer	Occupational Fundamentals	Occupational Knowledge; Case Analysis
Chemical Engineer	Occupational Fundamentals	Occupational Knowledge; Case Analysis
Civil Engineer	Geotechnical Engineering Fundamentals; Road Engineering Fundamentals	Geotechnical Engineering Knowledge; Geotechnical Case Analysis; Road Engineering Knowledge; Road Engineering Case Analysis
Environmental Engineer	Occupational Fundamentals	Occupational Knowledge; Case Analysis
Electrical Engineer	Occupational Fundamentals	Case Analysis

F Error Analysis

Table 10 examines five objective questions and one subjective question drawn from 6 examination subjects across five qualifications. Each case includes the Chinese question, an English translation, and results from three models. We inspect the recorded responses to locate erroneous steps or omitted content, using the reference answers and, for the subjective case, the stored rubric and judgments. These selected cases illustrate failure modes without estimating their prevalence.

Table 10: Error categories and examples with multiple model results. Questions are condensed where needed. Quoted reasoning and answer excerpts are translated from the models' visible responses; ellipses mark omissions, and bold phrases highlight erroneous claims. Subjective scores use the same GPT-5.6-Sol rubric and judge.

Error category	Question, LLM results, and error analysis
Legal and regulatory scope. Confusing the circumstances of an act with the legal conditions for assigning liability.	**Question.** 某县公安局干警张某在协助县林业局执行公务时，遇到往日有嫌隙的王某，张某借故将王某殴打致死。对王某的死亡承担赔偿义务的是? **English.** While assisting the county forestry bureau with official duties, police officer Zhang encounters Wang, with whom he has a personal grudge, and uses a pretext to beat him to death. Who is liable to pay compensation? A. 该县公安局/ The county public security bureau; B. 该县人民政府/ The county government; C. 张某/ Zhang; D. 该县林业局/ The county forestry bureau. **Reference answer: C.** **GPT-5.6-Sol: A.** "Even if he acted out of personal animosity and with intent, **this does not directly turn the act into a purely personal act**." **Gemini-3.1-Pro: D.** "His conduct was, in substance, **the exercise of administrative authority under the forestry bureau's delegation/direction**." **Qwen3.5-397B: C.** "The act . . . has no intrinsic connection with performing official duties and is a personal act unrelated to the exercise of authority." **Error analysis.** The two incorrect responses treat the on-duty setting as establishing official action. Gemini additionally infers a delegation relationship. The reference interpretation assigns liability to Zhang's personal act; Qwen identifies this distinction.
Context-specific professional terminology. Applying a general definition to a term with a specific occupational meaning.	**Question.** 导游领队的"一岗双责"是指导游领队应兼具哪两项职责? **English.** Which two responsibilities does "one post, two responsibilities" refer to for tour guides and tour leaders? A. 保护旅游者安全与预防新冠肺炎疫情/ Protecting tourist safety and preventing COVID-19; B. 为旅游者提供服务与保护旅游者安全/ Providing tourism services and protecting tourist safety; C. 预防新冠肺炎疫情与引导文明旅游/ Preventing COVID-19 and promoting civilized tourism; D. 为旅游者提供服务与引导文明旅游/ Providing tourism services and promoting civilized tourism. **Reference answer: D.** **GPT-5.6-Sol: B.** "'One post, two responsibilities' means that, while performing one's primary duties, one must also **undertake the corresponding safety responsibility**." **Qwen3.5-122B: B.** "In industry regulation, this term places greater emphasis on the dual baseline of **'business + safety'**." **GLM-5.2: D.** "Although protecting tourist safety in B is one of a guide's duties, it does not match the standard wording in the policy documents on 'one post, two responsibilities'." **Error analysis.** Both incorrect explanations substitute a familiar safety-management interpretation for the tourism-specific definition. GLM distinguishes a generally valid duty from the particular duties named by the term.

Continued on next page

Table 10 – continued from previous page

Error category	Question, LLM results, and error analysis
Selecting the correct quantitative target. Computing an intermediate quantity correctly but failing to obtain the quantity requested.	**Question.** 某医用输液泵在20 mL/h 设置点重复测量三次，数据为20.8、21.4、20.5 mL/h。以三次测量的算术平均值作为测得值，则测得值及其实验标准偏差为多少？极差系数$C_2 = 1.13$、$C_3 = 1.69$、$C_4 = 2.06$。 **English.** Three flow measurements of a medical infusion pump at a 20 mL/h setting are 20.8, 21.4, and 20.5 mL/h. The reported measured value is their arithmetic mean. What are this value and its experimental standard deviation? The range coefficients are $C_2 = 1.13$, $C_3 = 1.69$, and $C_4 = 2.06$. **Options** (measured value, standard deviation; both in mL/h): A. $(20.9, 0.31)$; B. $(20.9, 0.53)$; C. $(20.8, 0.31)$; D. $(20.8, 0.53)$. **Reference answer: A.** **GPT-5.6-Sol: B.** "$s = R/C_3 = 0.9/1.69 \approx 0.53$ mL/h. Therefore, the measured value is 20.9 mL/h and **the experimental standard deviation is 0.53 mL/h**." **GLM-5.2: B.** "Experimental standard deviation $s = R/C_3 = 0.9/1.69 \approx 0.5325 \approx 0.53$ mL/h. … **This corresponds to option B**." **Gemini-3.1-Pro: A.** "Here, 'its' refers to the measured value (the arithmetic mean), so the experimental standard deviation of the arithmetic mean is required. … $s_{\bar{x}} = s/\sqrt{n} \approx 0.31$ mL/h." **Error analysis.** GPT and GLM stop at the single-measurement standard deviation. The missing step is division by $\sqrt{3}$ to obtain the standard deviation of the mean. The error concerns the statistical target, not the preceding arithmetic.
Technical operating sequences. Reversing the required relationship between interacting mechanical components.	**Question.** 关于压力机离合器和制动器的说法，正确的是? **English.** Which statement about the clutch and brake of a mechanical press is correct? A. 大中型压力机离合器接合前，制动器必须处于制动状态/ On large and medium presses, the brake must be engaged before clutch engagement; B. 停机时，离合器脱开，制动器不制动/ When idle, the clutch is disengaged and the brake is released; C. 组合结构旨在提高二者同时结合的可能性/ A combined clutch–brake structure increases the likelihood of simultaneous engagement; D. 瞬时停机时，离合器立即脱开，制动器立即接合/ On an instantaneous stop command, the clutch immediately disengages and the brake immediately engages. **Reference answer: D.** **GPT-5.6-Sol: A.** "Before clutch engagement on large and medium presses, **the brake should be engaged** to ensure that the slide cannot move unexpectedly." **Qwen3.5-9B: A.** "**Only after the clutch is fully engaged and the system is stable is the brake released**, allowing the slide to move." **Gemini-3.1-Pro: D.** "Before clutch engagement, the brake must first be released; otherwise, the two interfere mechanically and may damage the equipment." **Error analysis.** GPT and Qwen justify keeping the brake engaged as a precaution against motion, reversing the release-before-engagement sequence in the reference explanation. Gemini identifies the mechanical interference and selects the specified stop action.

Continued on next page

Table 10 – continued from previous page

Error category	Question, LLM results, and error analysis
Exact thresholds and answer justification. Using an incorrect numerical rule, or choosing an option that contradicts the model's own stated rule.	**Question.** 关于在电力杆塔周围取土的说法，哪一项正确? **English.** Which statement correctly specifies where soil may be excavated around a power-line tower? A. 110 kV 杆塔的4 m 外/ Beyond 4 m from a 110 kV tower; B. 220 kV 杆塔的5 m 外/ Beyond 5 m from a 220 kV tower; C. 330 kV 杆塔的6 m 外/ Beyond 6 m from a 330 kV tower; D. 500 kV 杆塔的7 m 外/ Beyond 7 m from a 500 kV tower. **Reference answer: B.** **GPT-5.6-Sol: D.** "When excavating around a 500 kV tower, a safety distance of **at least 7 m** should normally be maintained." **Gemini-3.1-Pro: D.** "For 110 kV, 220 kV, 330 kV, and 500 kV towers, excavation is prohibited within 10 m. … If forced to match a safety-distance value … **some question banks give D as the reference answer**." **Qwen3.5-397B: B.** "110–220 kV: within 5 m; 330–500 kV: within 8 m. … Beyond 5 m [for 220 kV] is outside the prohibited area." **Error analysis.** The reference gives 5 m for 110–220 kV and 8 m for 330–500 kV. GPT supplies an incorrect 7 m threshold. Gemini rejects every option under its recalled 10 m rule, then selects D by speculating about the answer key.
Incomplete constructed responses. Listing relevant work without specifying all required construction steps.	**Context and question.** 主坝为混凝土面板堆石坝。写出混凝土面板的施工内容。 **English.** The main dam is a concrete-faced rockfill dam. List the construction activities for its concrete face slab. **Reference answer.** 面板分块、垂直缝砂浆条铺设、钢筋架立、面板混凝土浇筑、面板养护。 **English.** Panel subdivision; placing mortar strips at vertical joints; erecting reinforcement; placing face-slab concrete; and curing. Each of these five rubric criteria is worth 2 points. **GPT-5.6-Sol: 6/10.** Answer excerpt: "Steel fabrication and installation; … construction of perimeter and vertical joints and waterstops; … concrete mixing, transport, placement, spreading, and vibration; … curing and defect treatment." **Gemini-3.1-Pro: 8/10.** Answer excerpt: "Panel subdivision; … steel fabrication and tying; … concrete placement; … concrete curing; … joint and waterstop construction." **GLM-5.2: 10/10.** Answer excerpt: "Subdivision of the concrete panels (or alternate-panel casting); placing mortar strips at vertical joints; erecting reinforcement; … concrete placement; … concrete curing." **Error analysis.** Inspection of the full answers confirms that GPT omits panel subdivision and mortar-strip placement. Gemini includes subdivision but leaves mortar-strip placement unspecified. GLM states both explicitly. General references to joint or waterstop construction do not identify the missing mortar-strip operation.