Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia

Fajri Koto
Department of Natural Language Processing
MBZUAI, Abu Dhabi, UAE
fajri.koto@mbzuai.ac.ae

Abstract

While knowledge evaluation in large language models has predominantly focused on academic subjects like math and physics, these assessments often fail to capture the practical demands of real-world professions. In this paper, we introduce IndoCareer, a dataset comprising 8,834 multiple-choice questions designed to evaluate performance in vocational and professional certification exams across various fields. With a focus on Indonesia, IndoCareer provides rich local contexts, spanning six key sectors: (1) healthcare, (2) insurance and finance, (3) creative and design, (4) tourism and hospitality, (5) education and training, and (6) law. Our comprehensive evaluation of 27 large language models shows that these models struggle particularly in fields with strong local contexts, such as insurance and finance. Additionally, while using the entire dataset, shuffling answer options generally maintains consistent evaluation results across models, but it introduces instability specifically in the insurance and finance sectors.[1]

1 Introduction

The evaluation of large language models (LLMs) has shifted from traditional natural language processing (NLP) tasks (Mikheev et al., 1999; Straka and Straková, 2017) to more complex, knowledge-intensive, and reasoning-based challenges. One of the key datasets used to assess these abilities is the massive multitask language understanding (MMLU) (Hendrycks et al., 2021). Initially introduced in English, MMLU datasets have also been developed in other languages, including Indonesian (Koto et al., 2023), Chinese (Li et al., 2024), and Arabic (Koto et al., 2024a). These datasets consist of school exam questions across various subjects and education levels, tailored to local curricula.[2] However, they primarily focus on academic subjects, often overlooking vocational and professional expertise, which are more relevant to real-world applications.

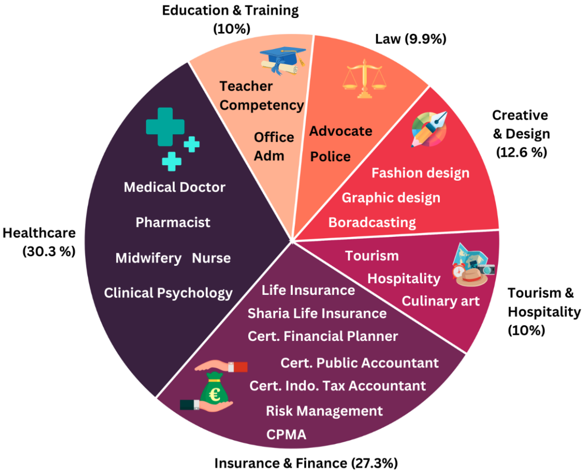


Figure 1: Distribution of professions in IndoCareer.

Due to the recent widespread adoption of LLMs across various domains, including health (Zhang et al., 2024), education (Weijers et al., 2024; Srivatsa and Kochmar, 2024), and finance (Lee and Soon, 2024), evaluating a model's knowledge across professional fields has become crucial. For instance, in healthcare, the model must adhere to ethical standards (Gundersen and Bærøe, 2022) and possess expertise in prevalent regional diseases. We should not trust AI-based health recommendations from models that have not passed a competency exam. Similarly, in education, the model needs to understand and align with local government teaching guidelines. Despite the importance of certification exams in professional fields, such exams have been largely excluded from prior work (Koto et al., 2023).

[1] Data can be accessed at https://huggingface.co/datasets/indolem/IndoCareer.

[2] The English MMLU is based on the U.S. curriculum, while the Indonesian MMLU follows the Indonesian curriculum.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 938–948
April 30, 2025 ©2025 Association for Computational Linguistics

In this paper, we introduce IndoCareer, a dataset comprising 8,834 multiple-choice questions collected from various Indonesian competency exams, certification exams, and vocational school exams. Our focus on Indonesian addresses the limitations of prior work (Koto et al., 2023) and aims to enrich language diversity and local context nuances in NLP datasets, which are predominantly English-centric (Liu et al., 2024). Figure 1 shows the distribution of IndoCareer, which covers 22 different professions across 6 categories: (1) healthcare, (2) insurance and finance, (3) creative and design, (4) tourism and hospitality, (5) education and training, and (6) law. Additionally, we demonstrate that IndoCareer is generally robust to option shuffling (Zhou et al., 2024) when using the entire dataset, but it specifically introduces instability in insurance and finance professions.

2 Related Work

Indonesian Language Models IndoBERT (Koto et al., 2020; Wilie et al., 2020), IndoBERTweet (Koto et al., 2021), IndoGPT (Cahyawijaya et al., 2021), and IndoBART (Cahyawijaya et al., 2021) are among the earliest transformer-based language models developed from scratch for Indonesian. These models have been widely adopted by industry and academia across various applications. For models exceeding 1 billion parameters, no foundational models have been pre-trained exclusively on Indonesian text. Instead, research has focused on adapting multilingual models through fine-tuning techniques. Notable examples include Bactrian-X (Li et al., 2023), which employs LoRA (Hu et al., 2022) for fine-tuning LLama-1 (Touvron et al., 2023a), and Merak (Ichsan, 2023), Cendol (Cahyawijaya et al., 2024), and Komodo (Owen et al., 2024), which are fine-tuned adaptations of LLama-2 (Touvron et al., 2023b). Despite growing interest in deploying Indonesian LLMs across various domains and job sectors, there remains a lack of suitable benchmarks tailored to evaluate their performance. To address this gap, we introduce IndoCareer.

Benchmarks for Evaluating Language Models NusaCrowd (Cahyawijaya et al., 2023) represents a significant effort to consolidate scattered datasets for Indonesian NLP. While most high-quality datasets focus on classical NLP tasks such as sentiment analysis, summarization, and text classification, benchmarks for knowledge-intensive and reasoning tasks have been notably limited until very recently. The introduction of IndoMMLU (Koto et al., 2023), COPAL-ID (Wibowo et al., 2024), and IndoCulture (Koto et al., 2024b) marks a step forward in this direction. COPAL-ID and IndoCulture focus on cultural commonsense reasoning, while IndoMMLU evaluates exam questions across different education levels in Indonesia, from primary to high school.

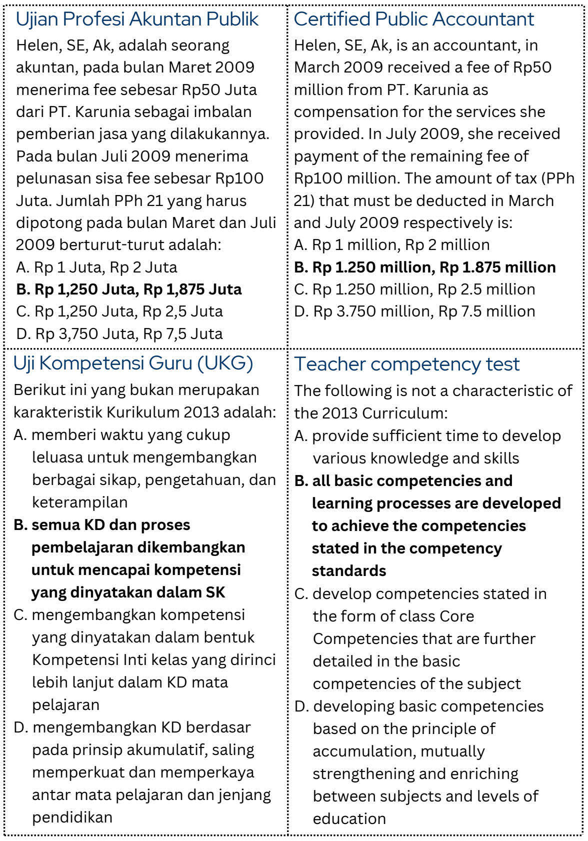
Despite recent advancements, a significant gap remains in evaluating LLMs on professional tasks in the Indonesian context, as IndoMMLU does not include questions from professional exams. This limitation is not unique to Indonesia; professional exam coverage is also limited in similar benchmarks for other languages. For example, English MMLU (Hendrycks et al., 2021) and Chinese MMLU (Li et al., 2024) include professional exam questions in only 20% of their datasets, while Arabic MMLU (Koto et al., 2024a) has an even lower coverage of just 4%.

As LLMs are increasingly applied across various domains (Zhang et al., 2024; Lee and Soon, 2024), there is a pressing need for a benchmark that evaluates their readiness for professional job sectors. IndoCareer addresses this gap, offering a comprehensive benchmark of professional exams spanning 22 professions, making it the first of its kind in Indonesia.

3 IndoCareer

IndoCareer comprises 8,834 multiple-choice questions compiled from Indonesian competency exams, certification exams, and vocational school exams across 22 professions. In Indonesia, competency exams are commonly required in healthcare professions by the government. Certification exams, on the other hand, focus on specific skills within a profession, such as tax accounting in finance. At the high school level, vocational schools offer specialized training in areas like tourism, culinary arts, and fashion design. In Figure 1, Table 1, and Table 2, we present detailed statistics for the 22 professions covered in IndoCareer. In this dataset, we exclude engineering-related professions, as their certification exams are generally conducted in English.

Data Construction We manually collected exam questions from publicly available sources across 22 professions. A majority (78%) of the questions



Ujian Profesi Akuntan Publik

Helen, SE, Ak, adalah seorang akuntan, pada bulan Maret 2009 menerima fee sebesar Rp50 Juta dari PT. Karunia sebagai imbalan pemberian jasa yang dilakukannya. Pada bulan Juli 2009 menerima pelunasan sisa fee sebesar Rp100 Juta. Jumlah PPh 21 yang harus dipotong pada bulan Maret dan Juli 2009 berturut-turut adalah:

A. Rp 1 Juta, Rp 2 Juta
B. Rp 1,250 Juta, Rp 1,875 Juta
C. Rp 1,250 Juta, Rp 2,5 Juta
D. Rp 3,750 Juta, Rp 7,5 Juta

Certified Public Accountant

Helen, SE, Ak, is an accountant, in March 2009 received a fee of Rp50 million from PT. Karunia as compensation for the services she provided. In July 2009, she received payment of the remaining fee of Rp100 million. The amount of tax (PPh 21) that must be deducted in March and July 2009 respectively is:

A. Rp 1 million, Rp 2 million
B. Rp 1.250 million, Rp 1.875 million
C. Rp 1.250 million, Rp 2.5 million
D. Rp 3.750 million, Rp 7.5 million

Uji Kompetensi Guru (UKG)

Berikut ini yang bukan merupakan karakteristik Kurikulum 2013 adalah:

A. memberi waktu yang cukup leluasa untuk mengembangkan berbagai sikap, pengetahuan, dan keterampilan
B. semua KD dan proses pembelajaran dikembangkan untuk mencapai kompetensi yang dinyatakan dalam SK
C. mengembangkan kompetensi yang dinyatakan dalam bentuk Kompetensi Inti kelas yang dirinci lebih lanjut dalam KD mata pelajaran
D. mengembangkan KD berdasar pada prinsip akumulatif, saling memperkuat dan memperkaya antar mata pelajaran dan jenjang pendidikan

Teacher competency test

The following is not a characteristic of the 2013 Curriculum:

A. provide sufficient time to develop various knowledge and skills
B. all basic competencies and learning processes are developed to achieve the competencies stated in the competency standards
C. develop competencies stated in the form of class Core Competencies that are further detailed in the basic competencies of the subject
D. developing basic competencies based on the principle of accumulation, mutually strengthening and enriching between subjects and levels of education

Figure 2: Example of questions in IndoCareer. The English translation is only for illustrative purposes.

were sourced from Scribd,[3] a document-sharing platform, while the remaining were obtained from local government websites[4] and shared Google Drive folders. We ensured that all collected questions were relevant to their respective professions and suitable for distribution for research purposes. Importantly, 99% of the exam questions were retrieved from file formats, such as PDFs and Word documents, rather than directly from web pages, minimizing the risk of overlap with training data used by LLMs.

To extract the questions and answers, we hired three professional teachers with Bachelor's degrees in Education for a one-month period. Their task focused exclusively on text-based questions, excluding any questions containing images (see Figure 2 for examples). Each worker was responsible for extracting approximately 3,000 questions. To ensure ethical practices, they were compensated above the minimum wage in Indonesia, with the total workload equivalent to five full-time workdays.

Field	Professions	Exam Type	#Q
Healthcare	Medical Doctor	Competency Exam	805
	Pharmacist	Competency Exam	598
	Midwifery	Competency Exam	680
	Nurse	Competency Exam	497
	Clinical Psychology	Other	95
Insurance & Finance	Life Insurance	Certification Exam	476
	Sharia Life Insurance	Certification Exam	558
	CFP	Certification Exam	96
	CPA	Certification Exam	663
	CPMA	Certification Exam	169
	CITA	Certification Exam	253
	Risk Management	Certification Exam	194
Tourism & Hospitality	Tourism	Vocational School	222
	Hospitality	Vocational School	367
	Culinary Art	Vocational School	294
Creative & Design	Graphic Design	Vocational School	423
	Fashion Design	Vocational School	267
	Broadcasting	Vocational School	422
Law	Advocate	Certification Exam	591
	Police	Other	280
Education & Training	Teacher Competency Test	Certification Exam	538
	Office Administration	Vocational School	346

Table 1: Number of questions in IndoCareer across different professions. CFP stands for Certified Financial Planner, CPA stands for Certified Public Accountant, CPMA stands for Certified Professional Management Accountant, and CITA stands for Certified Indonesian Tax Accountant.

Quality Control We ensure the high quality of our dataset through a rigorous and multi-step quality control process. Although we employ "expert" workers who are native Indonesian speakers with at least a Bachelor's degree, additional measures are implemented to maintain and verify quality. First, all data sources are manually checked and validated by the author before being distributed to the workers. Workers also participate in a 1-hour workshop prior to data collection, ensuring they fully understand the guidelines and the expected data standards.

After the workers complete their tasks, we apply automated filtering to eliminate repetitive questions and entries without answer keys. To further validate the dataset, we conducted a manual review of 300 randomly selected samples (3.3% of the dataset), performed by the authors of this paper. During this review, we verified the accuracy of the questions, answer options, and answer keys. The manual review achieved an accuracy rate of 99%, demonstrating the dataset's reliability and representing the highest meaningfully achievable score for IndoCareer.

Data Statistics Table 1 summarizes the distribution of questions in IndoCareer across 22 pro-

[3] https://www.scribd.com/

[4] For example: https://badanbahasa.kemdikbud.go.id

Field	# Questions	# Chars	
		Question	Answer
Healthcare	2675	277.3	95.9
Insurance and Finance	2409	156.3	165.3
Tourism and Hospitality	883	99.8	96.2
Creative and Design	1112	101.0	100.5
Law	871	130.7	141.2
Education and Training	884	159.5	165.9

Table 2: Average question and answer length (in characters) for each profession fields.

fessions, organized into six main fields: Healthcare, Insurance & Finance, Tourism & Hospitality, Creative & Design, Law, and Education & Training. Each profession corresponds to specific exam types, including competency exams, certification exams, vocational school exams, and others. Healthcare encompasses five professions, such as Medical Doctor and Pharmacist, contributing a total of 2,675 questions. Insurance & Finance, the largest category with seven professions, includes fields like Life Insurance, Certified Public Accountant (CPA), and Risk Management, with 2,409 questions. Tourism & Hospitality covers three professions—Tourism, Hospitality, and Culinary Art—comprising 883 questions, while Creative & Design features 1,112 questions. The Law field includes Advocate and Police exams, with a total of 871 questions, while Education & Training, with Teacher Competency Tests and Office Administration, adds another 884 questions.

According to Table 2, healthcare questions are the longest, averaging 2 to 3 times the length of those in tourism and hospitality, and creative and design. The number of multiple-choice options is generally consistent across professional fields, averaging 4 options. However, the total character count of the options varies, with insurance and finance, and education and training having the longest options, exceeding 160 characters.

Additionally, we manually examined 300 random samples to assess whether answering the questions required local context.[5] Our analysis revealed that 34% of the questions incorporated Indonesian local context, with a notable concentration in the fields of insurance and finance, tourism and hospitality, and law.

4 Experiments

Pezeshkpour and Hruschka (2024); Zhou et al. (2024) demonstrated that LLMs are highly sensitive to the order of options in multiple-choice questions. To ensure a more robust evaluation, we report the average performance across three evaluations for each model: one using the original order of options and two with the options shuffled.[6] We evaluated one closed-source model (GPT-4o) and 26 open-weight LLMs, comprising 18 multilingual models (BLOOMZ (Muennighoff et al., 2022), mT0 (Muennighoff et al., 2022), Gemma-2 (Team et al., 2024), Aya-23 (Üstün et al., 2024), LLaMA3.1[7]) and 8 Indonesian-centric models (IndoGPT (Cahyawijaya et al., 2021), Bactrian-ID (Li et al., 2023), Merak (Ichsan, 2023), Komodo (Owen et al., 2024), SeaLLM (Nguyen et al., 2023), SEA-LION (Singapore, 2023), and Cendol (Cahyawijaya et al., 2024)). Details for each model can be found in the Appendix.

Our focus is on zero-shot experiments using the Indonesian prompt: *Ini adalah soal [subject] untuk [exam type]. Pilihlah salah satu jawaban yang dianggap benar!*.[8] For evaluation, we use the `LM-Harness` package (Gao et al., 2024), selecting the answer based on the highest probability of the first token (i.e., A, B, C, D) in the generated output. Specifically, for GPT-4o, we used the `gpt-4o` model from OpenAI,[9] selecting the answer based on the first letter generated in the output.[10]

4.1 Results

Table 3 summarizes the zero-shot performance of various large language models (LLMs) across professional fields in `IndoCareer`, highlighting significant differences in their ability to handle Indonesian professional exams. GPT-4o and LLaMA-3.1 (70B) emerge as the top-performing models, with GPT-4o achieving the highest overall accuracy at 72.3%, followed closely by LLaMA-3.1 (70B) with 68.5%. This 4-point gap demonstrates GPT-4o's superior capability in handling complex tasks across diverse professions. In contrast, other multilingual models show significantly lower accuracy, ranging

[5] The 300 random samples are the same as those used for the manual review. Given the 99% accuracy rate from the initial review, we included an additional 1% of randomly selected correct samples for the local context assessment.

[6] For reproducibility, we also release two versions of `IndoCareer` with shuffled options, available at `https://huggingface.co/datasets/indolem/IndoCareer`.

[7] `https://github.com/meta-llama/llama3`

[8] The English translation is "This is a [subject] question for [exam type]. Please choose the correct answer!"

[9] `https://openai.com/`

[10] For GPT-4o, we slightly adjusted the prompt, instructing the model to output only one of the options as the answer.

Model (#parameters)	**Healthcare**	**Insurance & Finance**	**Tourism & Hospitality**	**Law**	**Creative & Design**	**Education & Training**	**Average**
Random	20.6	25.8	20.0	24.1	20.1	22.8	22.5
BLOOMZ (560M)	17.9	23.9	19.3	27.5	17.6	24.9	21.3
BLOOMZ (1.7B)	28.2	34.7	40.2	32.6	39.0	35.9	33.7
BLOOMZ (3B)	29.8	39.2	42.2	37.3	44.2	40.8	37.3
BLOOMZ (7B)	32.9	41.7	47.1	40.3	48.9	45.1	40.7
mT0$_{\text{small}}$ (300M)	22.3	26.2	21.7	23.5	22.2	19.5	23.1
mT0$_{\text{base}}$ (580M)	23.3	26.5	24.8	24.3	23.0	24.0	24.4
mT0$_{\text{large}}$ (1.2B)	25.0	26.8	25.3	24.2	24.3	23.3	25.2
mT0$_{\text{xl}}$ (3.7B)	27.7	38.9	43.8	36.0	42.4	43.3	36.6
mT0$_{\text{xxl}}$ (13B)	29.4	41.1	44.3	40.0	46.1	44.1	38.7
Gemma-2 (2B)	35.7	51.0	55.5	44.4	55.0	52.1	46.8
Gemma-2 (9B)	54.3	62.2	68.0	56.9	68.1	60.8	60.5
Gemma-2 (27B)	58.3	64.2	71.7	60.2	71.7	62.6	63.5
Aya-23 (8B)	37.0	46.1	51.7	44.3	51.7	47.5	44.6
Aya-23 (35B)	43.9	52.9	59.0	50.4	61.8	53.3	51.7
LLaMA-3.1 (8B)	35.9	46.7	51.9	41.2	53.0	45.3	44.1
LLaMA-3.1$_{\text{Instruct}}$ (8B)	44.8	53.6	61.1	47.7	63.3	54.9	52.4
LLaMA-3.1 (70B)	61.4	65.0	69.4	64.0	72.3	61.4	64.8
LLaMA-3.1$_{\text{Instruct}}$ (70B)	**64.4**	**69.3**	**74.2**	**68.1**	**75.1**	**65.3**	**68.5**
Bactrian-ID (7B)	20.5	29.0	22.7	26.6	25.5	25.1	24.7
IndoGPT (117M)	21.5	26.6	24.5	23.2	18.1	23.6	23.2
Merak (7B)	37.2	45.6	49.7	43.8	50.8	46.9	44.1
SeaLLM (7B)	**41.1**	**54.7**	**56.0**	**44.7**	**61.3**	**50.8**	**50.1**
SEA-LION (7B)	19.2	28.9	20.0	27.6	20.9	27.3	23.8
Komodo (7B)	25.5	29.7	27.4	30.5	29.8	31.8	28.5
Cendol$_{\text{mT5-xxl}}$ (13B)	20.8	24.8	22.9	22.9	21.8	21.4	22.5
Cendol$_{\text{LLaMA2}}$ (13B)	23.3	28.6	22.7	24.7	24.0	25.2	25.1
GPT-4o	**68.3**	**73.5**	**75.7**	**75.4**	**78.3**	**67.4**	**72.3**

Table 3: Zero-shot LLM performance (% accuracy), combined across professional fields. "Average" means the average across all questions in IndoCareer.

between 38.0% and 60.0%, indicating their struggles with Indonesian-specific professional exams.

Indonesian-centric models, including SEA-LION, Komodo, and Cendol, underperform dramatically, with results close to random guessing in some fields. These findings suggest that existing Indonesian-centric models are not yet optimized for professional exam tasks, limiting their utility in practical applications. Notably, the SEA-LION (7B) and Komodo (7B) models achieve only 23.8% and 28.5% average accuracy, respectively, underscoring the gap between local adaptations and the more capable multilingual models.

Healthcare stands out as the most challenging professional field, with an average performance across all models at only 37.2%.[11] This poor performance underscores the limitations of current off-the-shelf LLMs as reliable health advisors in the Indonesian context. These findings highlight the critical need for robust model adaptations and fine-tuning specifically tailored to Indonesian professional tasks to enhance performance and to ensure applicability in high-stakes domains such as healthcare.

4.2 Analysis

Shuffling the multiple-choice options leads to unstable results in insurance and finance. Table 4 lists the top 10 professions with the highest standard deviation (σ) in performance across three evaluation runs. While the standard deviations are relatively low, ranging from 1.5 to 3.0, they indicate minor instabilities in model predictions when the multiple-choice options are shuffled. For certain professions, such as Certified Financial Planner, Certified Indonesian Tax Accountant, and Certified Professional Management Accountant, the average rank correlation (τ) drops below 0.9, indicating reduced consistency in model performance across evaluation runs. Although their deviations are not severe, they highlight areas where models are less robust to option shuffling, particularly in domains requiring nuanced reasoning. Across the entire

[11] This figure is calculated by averaging all values in the Healthcare column of Table 3.

Profession	$\sigma\downarrow$	$\tau\uparrow$
Clinical Psychology	3.00	0.93
Cert. Financial Planner	2.91	0.68
Cert. Professional Management Accountant	2.00	0.90
Fashion Design	1.98	0.93
Advocate	1.96	0.91
Police	1.86	0.95
Cert. Indo. Tax Accountant	1.81	0.85
Sharia Life Insurance	1.81	0.97
Risk Management	1.76	0.97
Tourism	1.63	0.96
All	1.57	0.98

Table 4: Top 10 professions with the highest standard deviation (σ). τ represents the average rank correlation across three runs. The red cells are the three worse score. The scores are based on evaluations across 27 models.

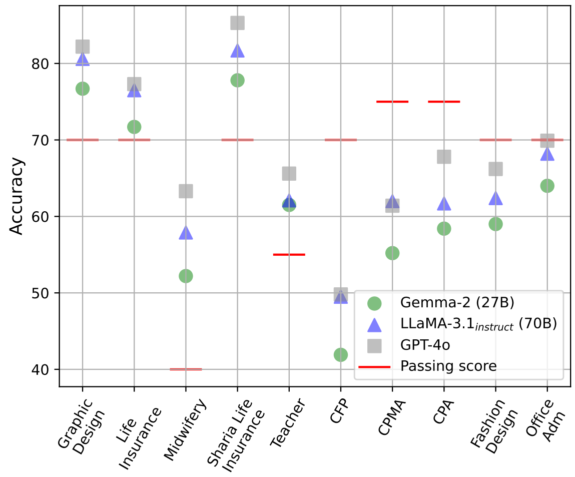


Figure 3: Top 5 and bottom 5 professions based on the model's accuracy disparity relative to the passing score.

.

dataset, however, the rank correlation remains high, with an average of 0.98. This indicates that while minor instabilities exist at the profession level, the overall dataset maintains stable performance.

LLMs perform well in life insurance certification but struggle with finance-related certifications. Figure 3 illustrates the performance of LLMs across the top 5 and bottom 5 professions in terms of accuracy relative to the passing scores. The passing scores for each exam, represented by red horizontal lines, were sourced from publicly available information. The figure highlights that while GPT-4o, LLaMA-3.1 (70B), and Gemma-2 (27B) achieve passing scores for professions such as life insurance, sharia life insurance, graphic design, midwifery, and teacher competency, they fall significantly short for finance-related certifications. None of the models evaluated pass the exams for Certified Financial Planner (CFP), Certified Professional Management Accountant (CPMA), Certified Public Accountant (CPA), fashion design, or office administration. Notably, GPT-4o, the best-performing model overall, falls over 20 points below the passing score for CFP, emphasizing the difficulty of finance-related tasks. The results suggest that finance-related certifications, which often require domain-specific reasoning and detailed calculations, remain a challenge for current LLMs. On the other hand, professions with more straightforward knowledge requirements, such as life insurance or midwifery, align better with the strengths of existing LLMs. These findings highlight the need for targeted fine-tuning and adaptation to improve performance in specialized and calculation-heavy fields like finance.

Questions with local context and numerical analysis pose greater challenges. We conducted an error analysis on the best-performing open-weight model, LLaMA-3.1 (70B), by examining 100 incorrectly predicted samples and 100 correctly predicted samples for comparison. These samples were drawn from the original questions, without applying option shuffling. The analysis showed that questions with Indonesian local context were more common among the incorrectly predicted samples, with 50% of the incorrect predictions containing local context, compared to only 22% among the correct predictions. Considering that `IndoCareer` contains 34% local context overall, as discussed in Section 3, this suggests that questions incorporating local context are particularly challenging for language models. This finding aligns with prior research (Koto et al., 2024b), indicating that questions grounded in local context often introduce cultural or situational nuances not well-captured in the models' pretraining data.

In addition to local context, questions involving numerical analysis also posed significant challenges for LLaMA-3.1 (70B). Among the incorrectly predicted samples, 43 required numerical reasoning, compared to only 29 among the correctly predicted ones. Numerical questions often involve calculations or logical reasoning steps, which many LLMs are not explicitly optimized to handle. These results reveal two key areas where model performance could be improved: understanding and addressing culturally specific content and enhancing their capabilities for numerical reasoning.

5 Conclusion

We introduce IndoCareer as the most comprehensive dataset of professional exams across various job sectors in Indonesia. The dataset encompasses 22 professions, categorized into healthcare, insurance and finance, creative and design, tourism and hospitality, education and training, and law. Evaluations across different LLMs show that most off-the-shelf models demonstrate vocational and professional expertise below the passing scores. We believe IndoCareer will be valuable in supporting LLM adaptation for various job sectors in Indonesia.

Limitations

There are three main limitations to our work: (1) IndoCareer excludes multimodal data such as tables, audio, images, and videos. Including these would make the benchmark more comprehensive and reflective of real-world scenarios. However, since our focus is on LLM evaluation, we only include text-based questions; (2) Engineering-related professions are excluded from IndoCareer because the language used in these exams is primarily English, while our focus is on the Indonesian language; (3) The evaluation is limited to multiple-choice questions and does not include text generation tasks. We follow prior work in using the multiple-choice format as an initial step to address the lack of professional and vocational exam benchmarks in Indonesian.

Ethical Considerations

IndoCareer is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[12] and is intended solely for academic research. The questions included in IndoCareer are sourced from publicly available materials. We collected these questions in compliance with Indonesian Copyright Law No. 28 of 2014, specifically Article 44. This article states that the use, reproduction, and/or modification of works or related rights, in whole or in part, is not considered copyright infringement, provided the source is properly cited and the purpose is for education or research.[13]

[12] https://creativecommons.org/licenses/by-nc-sa/4.0/

[13] https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/id/id064en.pdf

References

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Torbjørn Gundersen and Kristine Bærøe. 2022. The future ethics of artificial intelligence in medicine: making sense of collaborative models. *Science and engineering ethics*, 28(2):17.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Muhammad Ichsan. 2023. Merak-7b: The llm for bahasa indonesia. *Hugging Face Repository*.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *arXiv preprint arXiv:2404.01854*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Meisin Lee and Lay-Ki Soon. 2024. 'finance wizard' at the FinLLM challenge task: Financial text summarization. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 153–158, Jeju, South Korea. -.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into Indonesia's regional languages. *arXiv preprint arXiv:2403.09362*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

AI Singapore. 2023. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Kv Aditya Srivatsa and Ekaterina Kochmar. 2024. What makes math word problems challenging for LLMs? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1138–1148, Mexico City, Mexico. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Ruben Weijers, Gabrielle Fidelis de Castilho, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Quantifying learning-style adaptation in effectiveness of LLM teaching. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 112–118, St. Julians, Malta. Association for Computational Linguistics.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. LLM-based medical assistant personalization with short- and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398, Mexico City, Mexico. Association for Computational Linguistics.

Wenjie Zhou, Qiang Wang, Mingzhou Xu, Ming Chen, and Xiangyu Duan. 2024. Revisiting the self-consistency challenges in multi-choice question formats for large language model evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14103–14110, Torino, Italia. ELRA and ICCL.

A Models

Models (#parameters)	Source
BLOOMZ (560M)	`bigscience/bloomz-560m`
BLOOMZ (1.1B)	`bigscience/bloomz-1b1`
BLOOMZ (1.7B)	`bigscience/bloomz-1b7`
BLOOMZ (3B)	`bigscience/bloomz-3b`
BLOOMZ (7.1B)	`bigscience/bloomz-7b1`
mT0$_{\text{small}}$ (300M)	`bigscience/mt0-small`
mT0$_{\text{base}}$ (580M)	`bigscience/mt0-base`
mT0$_{\text{large}}$ (1.2B)	`bigscience/mt0-large`
mT0$_{\text{xl}}$ (3.7B)	`bigscience/mt0-xl`
mT0$_{\text{xxl}}$ (13B)	`bigscience/mt0-xxl`
Gemma-2 (2B)	`google/gemma-2-2b-it`
Gemma-2 (9B)	`google/gemma-2-9b-it`
Gemma-2 (27B)	`google/gemma-2-27b-it`
Aya-23 (8B)	`CohereForAI/aya-23-8B`
Aya-23 (35B)	`CohereForAI/aya-23-35B`
LLaMA3.1 (8B)	`meta-llama/Meta-Llama-3.1-8B`
LLaMA3.1-Instruct (8B)	`meta-llama/Meta-Llama-3.1-8B-Instruct`
LLaMA3.1 (70B)	`meta-llama/Meta-Llama-3.1-70B`
LLaMA3.1-chat (70B)	`meta-llama/Meta-Llama-3.1-70B-Instruct`
Bactrian-ID (7B)	`haonan-li/bactrian-id-llama-7b-lora`
IndoBART (132M)	`indobenchmark/indobart-v2`
IndoGPT (117M)	`indobenchmark/indogpt`
Merak (7B)	`Ichsan2895/Merak-7B-v5-PROTOTYPE1`
SeaLLM (7B)	`SeaLLMs/SeaLLMs-v3-7B-Chat`
SEA-LION (7B)	`aisingapore/sea-lion-7b`
Komodo (7B)	`Yellow-AI-NLP/komodo-7b-base`
Cendol$_{\text{mT5-xxl}}$ (13B)	`indonlp/cendol-mt5-xxl-merged-inst`
Cendol$_{\text{LLaMA2}}$ (13B)	`indonlp/cendol-llama2-13b-merged-chat`

Table 5: With the exception of GPT-4o, all the models used in this study were sourced from Huggingface (Wolf et al., 2020).

B Full Results

Table 6 presents the accuracy of each model across various professions. The passing scores for each exam were sourced from publicly available information. We found that GPT-4o passes most of the exams, with the exceptions being Certified Financial Planner, Certified Public Accountant, Certified Professional Management Accountant (CPMA), and Office Administration. LLaMA-3.1 and Gemma-2 also pass some Indonesian exams, but no Indonesian-centric model has yet passed the professional and vocational exams in Indonesia.

Profession	**P.Score**	**BLOOMZ**	**mT0**	**Aya-23**	**Gemma-2**	**LLaMA3.1**	**Merak**	**SeaLLM**	**SEA-LION**	**Komodo**	**Cendol**	**GPT-4o**
Healthcare												
Medical Doctor	66.0	33.4	27.3	45.1	61.6	70.9	39.6	42.9	19.9	24.1	23.1	74.8
Pharmacist	57.0	30.0	24.4	41.2	55.8	62.9	36.2	40.3	19.4	23.0	23.0	64.5
Midwifery	40.0	29.4	30.9	42.2	52.2	57.9	33.1	37.4	20.5	22.8	22.2	63.3
Nurse	60.0	35.3	33.9	45.5	58.4	60.8	36.2	44.7	23.8	25.6	25.7	66.4
Clinical Psycology	70.0	53.3	50.7	62.7	73.6	71.4	53.3	62.7	26.4	33.7	34.1	74.3
Insurance & Finance												
Life Insurance	70.0	48.7	48.1	61.3	71.7	76.5	49.0	59.3	27.3	35.0	30.2	77.3
Sharia Life Insurance	70.0	45.9	47.1	61.6	77.8	81.7	51.7	64.1	32.2	30.0	31.8	85.3
Cert. Financial Planner	70.0	29.4	27.6	36.9	41.9	49.5	25.4	38.0	24.0	22.6	22.2	49.8
Cert. Public Accountant	75.0	37.9	36.7	44.7	58.4	61.7	42.8	48.2	25.9	27.0	26.6	67.8
Cert. Indo. Tax Accountant	60.0	37.5	39.1	41.7	47.3	50.4	34.3	45.5	32.3	33.2	31.5	60.7
CPMA	75.0	32.7	28.7	40.6	55.2	62.0	38.6	40.8	26.9	25.9	24.1	61.4
Risk Management	70.0	41.7	37.9	51.1	64.0	67.2	45.0	53.2	26.9	29.3	32.1	70.9
Tourism & Hospitality												
Tourism	70.0	51.3	53.6	58.1	72.6	74.3	45.8	58.8	21.8	30.1	24.2	76.6
Hospitality	70.0	43.0	43.4	54.8	67.2	69.0	47.6	55.4	23.4	25.1	26.4	71.7
Culinary Art	70.0	47.0	45.2	62.0	73.5	76.7	50.1	60.4	22.8	28.3	22.2	79.5
Creative & Design												
Fashion Design	70.0	34.8	35.2	47.1	59.0	62.4	36.4	49.1	20.6	25.3	21.7	66.2
Graphic Design	70.0	52.9	53.8	65.3	76.7	80.6	54.8	65.0	23.6	29.3	28.0	82.2
Broadcasting	70.0	51.6	49.2	63.9	75.3	77.3	54.7	63.9	23.4	30.7	25.0	79.8
Law												
Advocate	70.0	34.6	39.9	47.1	59.9	68.7	36.7	41.9	26.4	27.3	26.4	72.9
Police	60.0	44.2	37.4	47.7	56.2	64.0	45.4	47.5	21.7	27.0	26.5	67.6
Education & Training												
Teacher Competency	55.0	46.6	44.1	53.4	61.5	62.1	45.5	48.8	27.5	29.3	27.7	65.6
Office Administration	70.0	38.9	42.1	52.1	64.0	68.2	42.0	50.0	24.2	27.7	23.0	69.9

Table 6: Zero-shot LLM performance (% accuracy) across professions for each model. "P.Score" indicates the passing score for each exam. The models used in this table include BLOOMZ (7B), mT0_{xxl}, Aya-23 (35B), Gemma-2 (27B), $\text{LLaMA-3.1}_{\text{Instruct}}$, Merak (7B), SeaLLM (7B), SEA-LION (7B), Komodo (7B), $\text{Cendol}_{\text{LLaMA2}}$ (13B) and GPT-4o. Green cells indicate that the model meets or exceeds the passing score.