TMCL: Measuring Taiwanese Criminal Law Understanding

Anonymous ACL submission

Abstract

We present TMCL (**T**aiwan **M**ultitask **C**riminal **L**aw), a new benchmark designed for Taiwanese criminal legal questions in Traditional Chinese. TMCL provides 11 tasks with a total of 1,914 multi-choice questions, including Taiwan's official examinations and human-annotated datasets for basic and conceptual features. We evaluate 10 recent LLMs supporting Simplified Chinese or Traditional Chinese, and the results indicate that understanding Taiwan's criminal legal knowledge remains a challenging task. Our dataset is publicly available on: https://anonymous.4open.science/r/TMCL_ACL-8421/

1 Introduction

The judgment corpus contains rich information and has become a popular research focus for social science and law researchers over the past few decades. With the rapid growth of large language models (LLMs), evaluating their understanding of legal documents is an essential issue across different regions and legal systems. Existing legal benchmarks for LLMs are predominantly designed for Anglophone legal systems, such as the United States (Guha et al., 2023). While efforts have been made to develop benchmarks for Chinese legal texts–e.g., (Fei et al., 2023) for the People's Republic of China and (Tam et al., 2024; Chen et al., 2024) for Taiwan–these benchmarks primarily evaluate broad general knowledge rather than in-depth, domain-specific legal analytic skills.

Given the specialized nature of jurisprudence, we argue that a more granular and domain-focused benchmark is essential for meaningful legal evaluation. Specifically, an effective benchmark should differentiate LLMs that exhibit lawyer-like reasoning from those that merely possess general legal knowledge. To address this gap, we concentrate on Taiwanese criminal law and curate a diverse set of questions tailored to this domain. This approach enables a more precise assessment of each model's ability to comprehend and reason about legal documents in a domain-specific context.

Question:
甲持偽造之信用卡至某影城的自動售票機，盜刷購買取得電影票及餐飲券，價值共計約5 萬元。下列敘述，何者正確？
(Person A, carrying a forged credit card, went to the self-service ticket machine at a certain movie theater and fraudulently swiped it to purchase movie tickets and dining vouchers, with a total value of approximately NT$50,000. Which of the following descriptions is correct?)
Choices:
A: 成立刑法第339條之1不正利用收費設備取財罪
(It constitutes the offense of improperly using fee-collecting equipment to obtain property under Article 339-1 of the Criminal Code.)
B: 成立刑法第339條詐欺罪
(It constitutes the offense of fraud under Article 339 of the Criminal Code.)
C: 成立刑法第320條竊盜罪
It constitutes the offense of theft under Article 320 of the Criminal Code.
D: 成立刑法第335條侵占罪
(It constitutes the offense of embezzlement under Article 335 of the Criminal Code.)
Answer: A

Figure 1: An example from the *Examinations for Judges and Prosecutors* task.

We formulate two challenges corresponding to the two task groups. The **Conceptual Features from Judgments** (CFFJ) challenge requires LLMs to correctly identify key sentencing reference features used by judges in practice based on the established facts of the crime. This task assesses the model's ability to extract legally-relevant facts, an essential analytical skills for legal professionals. The **Examinations for Judicial Personnel** (EJP) challenge evaluates whether an LLM can perform at a level comparable to passing professional legal qualification exams for judges, lawyers, clerks, etc. This task primarily tests the model's reasoning capabilities as well as legal knowledge.

In this paper, our contributions are as follows:

1. We present TMCL, a novel benchmark that comprises multi-level exams and feature-

extracted questions organized into 11 tasks, totaling 1,914 multiple-choice questions. We constructed two groups of tasks: (1) 1,086 human-annotated questions on conceptual features, and (2) 828 questions from three types of national examinations for legal practitioners, including judges, prosecutors, lawyers, and judicial clerks.

2. We test our dataset on ten recent LLMs, including API-based models (e.g., GPT-4o (OpenAI, 2024), Claude-3.5 (Anthropic, 2024)) and locally hosted models (e.g., Llama-3 variants (Llama Team, 2024), Qwen2.5 (Team, 2024)). The results indicate that current models face challenges in featured-based questions and exam-level questions related to Taiwanese legal knowledge.

3. We discussed the performance of the current Traditional Chinese models, raised potential reasons for their shortcomings, and compared them with other models.

2 Related Works

2.1 Benchmarks for Traditional Chinese and Law Tasks

Existing benchmarks in Traditional Chinese emphasizes general usage but not in legal systems. TC-Eval (Hsu et al., 2023) is the first benchmark for measuring LLMs understanding 55 subjects in Traditional Chinese, including one subject for basic law knowledge. TMLU (Chen et al., 2024) collects 37 subjects of Taiwanese official tests from 9th grades students to professionals, including one subject for lawyer qualification test. TMMLU+ (Tam et al., 2024) contains seven subjects focusing on different law domains out of 66 subjects. These datasets mainly follows the structure of MMLU (Hendrycks et al., 2021). Table 1 lists the evaluation benchmarks with the total count of questions.

In other legal systems, there are specific benchmarks for law tasks. LegalBench (Guha et al., 2023) collects 162 tasks covering six types of legal reasoning, including tasks containing judgments from real world, such as *Canada_tax_court_outcomes*. LawBench (Fei et al., 2023) collects 20 tasks covering three cognitive levels (memorization, understanding and applying), where the legal knowledge applying level as five tasks are based on the fact of judgments from two legal AI contests, CAIL2018 (Xiao et al., 2018) and LAIC2021[1].

Benchmark	test	dev
TC-Eval	25	5
TMLU	272	5
TMMLU+	1,763	35
TMCL (Ours)	1,859	55

Table 1: Counts of law-related questions in Traditional Chinese benchmarks. TMCL only contains questions based on Criminal Law and Criminal Procedure Law while questions in other benchmarks are general. TMMLU+ have other 197 questions for validation.

2.2 LLMs for Traditional Chinese and Law

To the best of our knowledge, three foundational models specialize in Taiwan domain knowledge: Taiwan-LLama (Lin and Chen, 2023), Breeze (Hsu et al., 2024), and TAIDE (TAIDE, 2024). Taiwan-LLama underwent continued pretraining on 35B tokens of Taiwan-specific data, followed by two supervised fine-tuning (SFT) stages. Breeze was pretrained on 650 GB of data[2], followed by one SFT stage. TAIDE, as a trustworthy model, underwent continued pretraining on 43B tokens (140 GB) of public data and an instruction-tuning dataset auto-generated by Llama2 (Touvron et al., 2023). Additionally, judgments from the past ten years were used during the continued pretraining stage.

Evaluations on Traditional Chinese benchmarks indicate that current state-of-the-art (SOTA) models, such as GPT-4o (OpenAI, 2024) and Claude (Anthropic, 2024), also demonstrate a strong understanding of Taiwan-specific knowledge and Traditional Chinese (Tam et al., 2024).

On top of the models trained in Traditional Chinese, those primarily trained in Simplified Chinese, such as Qwen (Team, 2024), also achieve high scores, suggesting that SOTA Simplified Chinese models should be considered as well (Hsu et al., 2024). Consequently, we include InternLM (Cai et al., 2024) in our considerations, as it outperforms Qwen in benchmarks designed for multitask Simplified Chinese understanding, such as CMMLU (Li et al., 2024).

[1] https://laic.cjbdi.com/

[2] Data tokens are not counted in the report.

3 TMCL Dataset

3.1 Dataset Representation

TMCL is a set $\{d, q, p, c, a\}$ composed of a description of the task d, a question q, a passage excerpted from a criminal judgment p, a set of choices $c = \{c_1, ..., c_i\}, i \in \{3, 4\}$, and a single answer a. The size of c is four for every EJP task and varies for CFFJ tasks based on the feature of the task. For binary features, c is fixed as $\{true, false, uncertain\}$ with the size as three. For other non-binary features, the size is fixed as four, where the choices are not guaranteed same. Appendix A.3 provides the descriptive statistics of the dataset.

3.2 Data Source

We compile our dataset from two publicly available websites, as detailed below:

1. Ministry of Examination (MOEX)[3]: We collect recent ten years of multi-choice criminal examinations for judges, lawyers and judicial personnel. Some positions require a second or third examination, such as physical and oral examinations. Details for examinations and corresponding positions are listed in Appendix A.4.

2. Judicial Yuan Judgment System (JYJS)[4]: Judicial Yuan provides open-source judgments from 1990s. We collect criminal judgments from January 2013 to June 2024, where most cases are public prosecution at the first instance. These judgments contain full names of defendants and persons involved in the case. To mitigate privacy concerns, personal information except for the full name, such as address and personal identification numbers, is substituted by placeholders for every individual in the judgments. In addition, to protect minors and witnesses, their full-names would be replaced by IDs[5].

3.3 Data Processing

The examination data from MOEX is provided in PDF format. We transcribe PDF files into text using pdfplumber[6] and categorize as the EJP group.

For the CFFJ group, we construct examples by segmenting the texts and extracting the syllabus, facts, and reasoning part. We categorize features into binary and non-binary types. Binary features are transcribed into multi-choice format, with A being *true*, B being *false*, and C being *uncertain*. In contrast, non-binary features encompass more complex or conceptual attributes, such as the weapon used by the defendant, or the debate regarding the evidentiary capabilities. The correct label for such features is first extracted and coded as the correct answer for evaluation, after which other three artificial (incorrect) options are generated. Finally, to ensure the correctness of the dataset, manual data annotation and verification are conducted. The human-annotated questions are curated by law-school students with a minimum of three years of legal training and Traditional Chinese as their native language, ensuring domain relevance and rigor, with each student being assigned a specific annotation task that is subsequently reviewed by another student. Accordingly, each task is doubly-checked by two different assistants.

For each task, five development questions are set aside for few-shot inference. Table 1 compares our dataset with other datasets, considering only law-related questions. In addition, we keep our data length below $8,192$ characters to fit the context window of local models, whose tokenizer supports Traditional Chinese with a token length smaller than the character size.

4 Experiments

4.1 Experiment Setup

We test our dataset on ten models, six of which are downloadable and four are callable via API. For the six local checkpoints, we use lm-evaluation-harness (Gao et al., 2024) to perform the log-likelihood evaluation. Log-likelihood evaluation, first introduced in the MMLU (Hendrycks et al., 2021), measures the confidence with which a language model generates an output. For API-based models not supporting log likelihood evaluation, we use a generation-based evaluation with an additional JSON prompt, requiring the LLM to output the results in JSON format to improve consistency. The accuracy is computed based on the exact match between the value in the output JSON object and the answer. For hyperparameter settings, we set tempurature $= 0.2$ and top p $= 0.7$ as the default setting of NVIDIA NIM[7]. In addition, We run our experiments on 3 NVIDIA RTX A6000 GPUs.

[3] https://www.moex.gov.tw/

[4] https://judgment.judicial.gov.tw/FJUD/defaulte.aspx

[5] In judgments, placeholders such as "ＯＯ" are often used to substitute personal information. Common IDs are 甲ＯＯ, 乙ＯＯ.

[6] https://github.com/jsvine/pdfplumber

[7] https://build.nvidia.com/

n-shot	Checkpoints	CFFJ	EJP	Average
five-shot	Claude-3-5-Sonnet-20241022 (Anthropic, 2024)	0.7368	**0.7875**	**0.7622**
	GPT-4o-2024-08-06 (OpenAI, 2024)	0.7047	0.7509	0.7278
	Llama-3.1-405b-Instruct (Llama Team, 2024)	0.7110	0.7111	0.7111
	Llama-3.3-70b-Instruct (Llama Team, 2024)	**0.7725**	0.6433	0.6964
	Qwen2.5-7B-Instruct (Team, 2024)	**0.5968**	**0.6049**	**0.6009**
	Llama-3-Taiwan-8B-Instruct-128k (Lin and Chen, 2023)	0.4928	0.4955	0.4942
zero-shot	Llama-3.1-405b-Instruct (Llama Team, 2024)	**0.7102**	0.7289	**0.7196**
	GPT-4o-2024-08-06 (OpenAI, 2024)	0.6449	0.7133	0.6791
	Claude-3-5-Sonnet-20241022 (Anthropic, 2024)	0.5860	**0.7661**	0.6761
	Llama-3.3-70b-Instruct (Llama Team, 2024)	0.6780	0.6546	0.6663
	Internlm3-8b-Instruct (Cai et al., 2024)	**0.5747**	**0.6117**	**0.5932**
	Llama-3-Taiwan-8B-Instruct (Lin and Chen, 2023)	0.4776	0.5535	0.5156
	Qwen2.5-7B-Instruct (Team, 2024)	0.4182	0.5616	0.4899
	Llama-3-Taiwan-8B-Instruct-128k (Lin and Chen, 2023)	0.4364	0.5194	0.4779
	Breeze-7B-Instruct-v1_0 (Hsu et al., 2024)	0.4680	0.4870	0.4775
	Llama3-TAIDE-LX-8B-Chat-Alpha1 (TAIDE, 2024)	0.5148	0.3063	0.4106

Table 2: Evaluations for zero-shot and five-shot settings. Dashed lines are used to separate API-based models and local models. The highest score in the group is shown in bold.

4.2 Zero-shot and Five-shot Evaluation

We run zero-shot inference on ten models and five-shot inference on seven models, and the results are shown in Table 2.

The CFFJ group and the EJP group require different abilities to achieve high scores. Since the CFFJ group is based on criminal judgments, these documents contain longer contexts with a lot of unrelated and misleading information. Therefore, the ability to understand instructions and Chinese language understanding is more important in this group compared to knowledge and understanding of Taiwanese legal terms. As a result, InternLM (Cai et al., 2024) achieves the highest accuracy, which is close to that of Claude-3.5-Sonnet (Anthropic, 2024). In addition, We attribute Claude-3.5-Sonnet's (Anthropic, 2024) weak performance in the zero-shot setting to its failure to follow instructions rather than a lack of understanding, as its performance improves significantly in the five-shot setting.

On the contrary, the EJP group requires a deeper knowledge and understanding of Taiwanese legal terms. While Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o (OpenAI, 2024), and Qwen2.5-7B-Instruct (Team, 2024) show an improvement of approximately four to six percentage points compared to the zero-shot and five-shot settings, we find that the results of Llama-3 variants (Llama Team, 2024) show no significant improvement. We believe this is due to the insufficient proportion of Chinese documents in the models' pre-training data, including models primarily trained on Traditional Chinese data. The scale of continued pretraining data for these Traditional Chinese models are hundreds of billions of tokens, whereas Simplified Chinese models are pretrained from trillions of tokens. This also aligns with previous studies, which have shown that Simplified Chinese models can outperform Traditional Chinese models across various benchmarks (Hsu et al., 2024).

4.3 Importance of Instruction-Tuning Dataset

The results of Traditional Chinese models also show significant differences. While TAIDE (TAIDE, 2024) includes ten years of recent Taiwanese legal judgments in its continued pretraining dataset (the same data source as our dataset), its performance falls short compared to other models due to the scale and quality of instruction tuning. Breeze (Hsu et al., 2024) performs better than TAIDE but worse than Taiwan-Llama (Lin and Chen, 2023) due to the lack of human-instructed SFT. This suggests that not only does pretraining data from a closely related domain matter, but a well-designed SFT stage is also crucial.

5 Conclusion

We introduce TMCL, a novel benchmark comprising 1,914 multiple-choice questions across 11 tasks and two groups derived from Taiwan's criminal judgments and official examinations. Our evaluation across ten recent LLMs reveals that despite their strong performance on general tasks, current models struggle to handle tasks specialized in Taiwan criminal law.

Limitation

Due to limited local GPU resources, we could not evaluate Llama-3-Taiwan-70B-Instruct (Lin and Chen, 2023) which is a checktpoint based on Meta-Llama-3-70B (Llama Team, 2024).

Ethical Statement

The judgments in our dataset are a revised version without personal information, lists of evidence, or crime details. This version differs from the internal version used by judges and prosecutors and is accessible to the public without any maturity ratings. The corpus is also available on the Judicial Yuan OpenData Platform[8].

Six student annotators in our study were recruited as part-time research assistants. They were paid at least the minimum wage and worked fewer than 20 hours per month.

References

Anthropic. 2024. Introducing claude 3.5 sonnet.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring taiwanese mandarin language understanding. In *First Conference on Language Modeling*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

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.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. 2023. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite. *Preprint*, arXiv:2309.08448.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7b technical report. *Preprint*, arXiv:2403.02712.

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. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan LLM: bridging the linguistic divide with a culturally aligned language model. *CoRR*, abs/2311.17487.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

[8] https://opendata.judicial.gov.tw/

OpenAI. 2024. Hello gpt-4o.

TAIDE. 2024. Taide taiwanese native large language model.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. 2024. TMMLU+: An improved traditional chinese evaluation suite for foundation models. In *First Conference on Language Modeling*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *Preprint*, arXiv:1807.02478.

A Appendix

A.1 Task Description

The **Examination for Judicial Personnel (EJP)** group includes three evaluation tasks: *Bar Judge Examination*, *Judicial Grade Four*, and *Judicial Grade Five*. These tasks are compiled from historical examination questions for judicial personnel, comprising a total of 828 questions.

The **Conceptual Features from Judgments (CFFJ)** group includes eight evaluation tasks: *Public Safety*, *Public Authority*, *Mastermind*, *Evidence Debate*, *Lineal Relative*, *Weapon*, *Cruel Mean*, and *Plan*, comprising a total of 1,086 questions:

1. **Public Safety** assesses the model's capability to identify whether the given case has significantly harmed public safety.
2. **Public Authority** involves cases where the victims are public servants, evaluating a model's ability to correctly opt the type of damages the victim suffered while exercising public authority.
3. **Mastermind** asks whether the specified defendant was the mastermind given a case of organized crime.
4. **Evidence Debate** examines whether the model can choose the optimal summary of the debate regarding the evidentiary capabilities.
5. **Lineal Relative** evaluates a model's ability to distinguish whether the specified defendant and victim are lineal relatives given a family case.
6. **Cruel Mean** evaluates a model's ability to judge whether the means employed in the case are considered cruel given a violent case.
7. **Weapon** assesses a model's ability to determine the specific weapon used by the defendant to commit the crime given a violent case.
8. **Plan** assesses a model's ability to conclude whether the case was rigorously planned given a case of organized crime.

A.2 Column Description

All evaluation tasks in the CFFJ group include the following columns: *question*, *data*, *A*, *B*, *C*, *D*, *answer*, *case label*, *case label short*, *main result, fact*, and *reasoning*. For tasks in the EJP group, only the columns *question*, *A*, *B*, *C*, *D*, and *answer* are available, as the data are compiled from the examination questions, which do not rely on judgment texts. Detailed description for each column is presented in Table A.2.

A.3 Descriptive Statistics

We present the descriptive statistics for the TMCL benchmark data, including the length of the complete prompt (which consists of instructions, questions, excerpts from judgments, and choices), the number of examples, and the label distribution, as shown in Table A.3. Each sub-column in the Distribution of Label section represents the proportion of the corresponding option within a given task. For certain tasks in the CFFJ group, column D is left empty (represented by a dash), indicating that only options A, B, and C are available. These tasks essentially consist of true/false questions, where A represents *true*, B represents *false*, and C represents *uncertain*.

A.4 Examinations

We present a summary of the Examinations for Judges and Prosecutors (EJP) in Table A.4. As shown in the table, each evaluation task under Examination for Judicial Personnel category corresponds to a specific examination category, with detailed civil service positions and passing scores outlined. The passing scores are calculated as the average values over the period from 2019 to 2024 and are rounded to two decimal places. Only the multiple-choice questions from the first exam are included, as the second and third exams consist of physical or oral examinations. These formats do not have definitive answers and may be challenging to use for evaluating language models.

A.5 Prompt Format

Given a row in our dataset and a description for the task, the prompt format is as follows:

```
{task description}
({FIXED JSON FORMAT PROMPT for API-based
models})
( {Example1}
{Example2}...)
Question: {row["question"]}
A. {row["A"]}
B. {row["B"]}
C. {row["C"]}
(D. {row["D"]} if "D" in row)
{row["data"]}
Answer:
```

The fixed JSON template is written in Traditional Chinese.

```
輸出請使用JSON格式，且*只需輸出JSON格式的
答案*，不要有任何解釋，如下：
output = {"LLM_answer": str}
---
Please output in JSON format, and *only
return the answer in JSON format* without
any explanation, as follows:
output = {"LLM_answer": str}
```

A.6 Consistency of Structural Outputs

Since only GPT-4o (OpenAI, 2024) supports structural outputs, there are errors in the results of other API-based models. We consider these errors as wrong answers in Table 2. We find that most errors occur in the CFFJ group, especially *Weapon* and *Evidence Debate*, which could be improved by few-shot learning. Details are shown in Table A.6. We also find that the result of Llama-3.1-405b-Instruct (Llama Team, 2024) is not stable compared with other two models.

Column Name	Description
question	The question component of the prompt.
data	Necessary contextual information that the language model relies on to answer the question, which is an excerpt from judgment texts, comprising of the main result column and the fact column.
A	Content of each choice. The answer column only contains the choice code (A, B, C, or D). For tasks based on binary features, only A (*true*), B (*false*), and C (*uncertain*) are available.
B	
C	
D	
answer	Correct answer for the row, which would only be one of A, B, C, and D (if available).
case label [1]	Officially used judgment case labels, mostly composed of Mandarin characters.
case label short [1]	Officially used judgment case labels, mostly composed of digits and alphabets.
main result [1]	The syllabus of the judgment text (a component of the data column).
fact [1]	The criminal fact paragraph of the judgment text (a component of the data column).
reasoning [1]	The reasons for judgment.

1. These columns are not available for tasks under Examination for Judicial Personnel group.

Table A.2 Column Description

Evaluation Task	Descriptive Statistics for Prompt Length							# Examples			Distribution of Label			
	mean	std	min	Q_{25}	Q_{50}	Q_{75}	max	total	dev	test	A	B	C	D
Examination for Judicial Personnel														
Bar Judge Examination	309.45	81.15	151	250.2	303	358.0	610	474	5	469	0.23	0.24	0.26	0.27
Judicial Grade Four	165.64	41.02	91	138.5	162	186.5	356	247	5	242	0.24	0.20	0.30	0.26
Judicial Grade Five	166.12	47.75	94	134.5	159	194.5	321	107	5	102	0.28	0.21	0.32	0.19
Conceptual Features from Judgments														
Public Safety	1667.46	1351.22	375	737.2	1228	2000.0	7294	172	5	167	0.69	0.3	0.01	-
Public Authority [1]	884.60	706.16	371	506.5	702	919.0	5807	130	5	125	**0.00**	**0.30**	**0.45**	**0.25** [1]
Mastermind	1350.43	1190.54	289	563.0	948	1701.5	6381	167	5	162	0.61	0.33	0.06	-
Evidence Debate	1337.25	526.33	751	944.0	1110	1717.8	2466	24	5	19	0.33	0.21	0.25	0.21
Lineal Relative	813.38	762.76	135	427.8	627	889.5	5923	104	5	99	0.22	0.64	0.14	-
Weapon	1288.57	1071.58	252	547.0	930	1547.5	6220	139	5	134	0.22	0.29	0.26	0.22
Cruel Mean	1797.63	1381.55	362	775.0	1309	2426.5	7109	174	5	169	0.31	0.65	0.04	-
Plan	2144.09	1404.57	443	1003.8	1837	2895	7114	176	5	171	0.51	0.31	0.18	-

1. For the Public Authority task, option A represents loss of life, indicating that the victim suffered fatal harm while exercising public authority. However, the selected examples do not include such cases. Consequently, the proportion of option A for the Public Authority task is zero.

Table A.3 Descriptive Statistics

Evaluation Task	Examination Category	Civil Service Position	Passing Score		
			First	Second	Third
Bar Judge Examination	Examination for Judges and Prosecutors	Judges and Prosecutors	381.33	512.92	599.61
		Intellectual Property Lawyer	372.33	501.42	-
		Labor and Society Lawyer		505.33	-
		Finance and Tax Lawyer		508.58	-
		Maritime Lawyer		509.53	-
Judicial Grade Four	Grade Four Special Examination for Judicial Personnel	Bailiff - Female	68.53	68.03	-
		Bailiff - Male	67.41	67.58	-
		Process Server	64.72	-	-
		Correctional Facility Custodial Personnel - Female	67.5	66.36	-
		Correctional Facility Custodial Personnel - Male	62.64	61.83	-
		Court Clerk - Grade Four	57.25	-	-
Judicial Grade Five	Grade Five Special Examination for Judicial Personnel	Clerk Assistant	80.75	-	-

Table A.4 Taiwanese Judicial Examination Information

Checkpoints	n-shot	Error Counts	CFFJ	EJP
Llama-3.3-70b-Instruct (Llama Team, 2024)	0-shot	5	5	0
	5-shot	0	0	0
Llama-3.1-405b-Instruct (Llama Team, 2024)	0-shot	7	7	0
	5-shot	4	0	4
claude-3-5-Sonnet-20241022 (Anthropic, 2024)	0-shot	11	11	0
	5-shot	0	0	0

Table A.6 Error counts of API-models by group.