TaxPraBen: A Scalable Benchmark for Structured Evaluation of LLMs in Chinese Real-World Tax Practice

Gang Hu[a], **Yating Chen**[a], **Haiyan Ding**[a], **Wang Gao**[c], **Jiajia Huang**[d]
Min Peng[b], **Qianqian Xie**[b], **Kun Yue**[a,*]
[a]Yunnan University, Yunnan, China. [b]Wuhan University, Wuhan, China.
[c]Jianghan University, Wuhan, China. [d]Nanjing Audit University, Nanjing, China.
hugang@ynu.edu.cn, yating_chen@stu.ynu.edu.cn, dinghaiyan@ynu.edu.cn,
gaow@jhun.edu.cn, huangjj@nau.edu.cn, pengm@whu.edu.cn
xqq.sincere@gmail.com, kyue@ynu.edu.cn

Abstract

While Large Language Models (LLMs) excel in various general domains, they exhibit notable gaps in the highly specialized, knowledge-intensive, and legally regulated Chinese tax domain. Consequently, while tax-related benchmarks are gaining attention, many focus on isolated NLP tasks, neglecting real-world practical capabilities. To address this issue, we introduce TaxPraBen, the first dedicated benchmark for Chinese taxation practice. It combines 10 traditional application tasks, along with 3 pioneering real-world scenarios: tax risk prevention, tax inspection analysis, and tax strategy planning, sourced from 14 datasets totaling 7.3K instances. TaxPraBen features a scalable structured evaluation paradigm designed through process of "structured parsing—field alignment extraction—numerical and textual matching", enabling end-to-end tax practice assessment while being extensible to other domains. We evaluate 19 LLMs based on Bloom's taxonomy. The results indicate significant performance disparities: all closed-source large-parameter LLMs excel, and Chinese LLMs like Qwen2.5 generally exceed multilingual LLMs, while the YaYi2 LLM, fine-tuned with some tax data, shows only limited improvement. TaxPraBen[1] serves as a vital resource for advancing evaluations of LLMs in practical applications.

1 Introduction

Taxation is vital for governance and economic regulation, demanding precise and timely expertise due to its complex, evolving policies (Rixen and Unger, 2022). Recent advancements in Large Language Models (LLMs) in natural language processing (NLP) offer new ways to interpret tax laws and enhance tax management, indicating a significant technological trend (Nay et al., 2024).

However, general LLMs such as OpenAI's ChatGPT (Brown et al., 2020), Google's Gemini (Team et al., 2024), and open-source DeepSeek (Guo et al., 2025a), while capable in broad domains, still underperform in specialized fields like finance (Hu et al., 2024) and auditing (Jiajia et al., 2024) that demand deep semantic understanding and complex numerical reasoning. This has spurred the development of various evaluation benchmarks for domain-specific LLMs. General benchmarks like GLUE (Wang et al., 2018), and MMLU (Hendrycks et al., 2009) focus on tasks like reading comprehension, making them inadequate for specialized assessments. This has led to the rise of domain-specific benchmarks, such as FinBen (Xie et al., 2024) for finance, MedBench (Cai et al., 2024a) for Medicine, and LAiW (Dai et al., 2025) for law. However, there is a relative scarcity of models and benchmarks specific to the tax scenarios, primarily due to a lack of large high-quality annotated data (Choi et al., 2025). Existing overseas tax data studies (Steinigen et al., 2023; Naudet, 2023) often differ significantly from China's tax management realities, complicating adaptation efforts. Moreover, some LLM studies (Luo et al., 2023) incorporate tax data that is not open source and lacks real-world applicability. Consequently, the proficiency of these LLMs in tax knowledge remains unclear, underscoring the need for custom tax-specific benchmarks.

Even the recently proposed TaxBen (Chen et al., 2025), like other domain-specific benchmarks, focuses mainly on isolated NLP tasks such as text classification, generation, and reasoning; comparisons are provided in Appendix A.1. General benchmarks have similar limitations. For example, MATH (Hendrycks et al., 2021) emphasizes mathematical correctness over explanation quality, while GAOKAO (Zhang et al., 2023a) and AGIEval (Zhong et al., 2024) focus on exam-style reasoning but overlook mixed-format outputs in professional settings. However, many real-world tax and financial auditing tasks require

[1]https://github.com/Yating-Chen/TaxPraBen

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38061–38104

both semantic reasoning and quantitative calculation (Krumdick et al., 2024). Existing domain-specific benchmarks, including TaxBen, overemphasize language understanding and neglect structured outputs, thus overstating the practical capabilities of current LLMs. As shown in Appendix A.2, some LLM rankings on TaxBen are inflated: models perform well on isolated numerical reasoning or semantic matching tasks, but struggle in real scenarios requiring both. Overall, existing benchmarks, including TaxBen, cannot reliably assess holistic practical capabilities by averaging NLP tasks. This suggests that real-world tax scenarios require LLMs to integrate semantic understanding and numerical reasoning for tax law interpretation, calculation, and compliance assessment. Therefore, traditional NLP benchmarks fail to capture the real demands of Chinese tax practice, posing challenges for LLM evaluation in this domain.

With this in mind, we introduce TaxPraBen, the first benchmark for evaluating LLMs in Chinese tax practice. Developed with tax-related experts, TaxPraBen contains 14 datasets with 7.3K samples, covering carefully annotated tax data and guiding prompts. All data are manually collected and annotated with model assistance rather than taken from public datasets. We also build a complete workflow integrating data annotation, evaluation metrics, and structured assessment, making TaxPraBen scalable. Following Bloom's taxonomy of cognitive skills (Fei et al., 2024; Chen et al., 2025), we divide the tasks into three groups: (1) *Knowledge Memorization*, (2) *Knowledge Understanding*, and (3) *Knowledge Application*.

Evaluating 19 representative general LLMs on TaxPraBen yields the following results: (1) **Overall Performance Variation**: ERNIE-3.5, Grok3, and ChatGPT perform best in both zero-shot and one-shot settings, highlighting the value of large-scale parameters and knowledge enhancement for tax tasks. (2) **Task Performance Bias**: Tax knowledge application is more challenging than understanding and memorization, reflecting the domain's reliance on economic activities and contextual scenarios. (3) **Language Background Advantage**: Chinese models perform better on tasks involving Chinese tax terminology and policies, showing the benefit of language-specific optimization. (4) **Data Coverage Gaps**: Tax-data fine-tuning still brings limited gains, likely due to insufficient data coverage or task mismatch. (5) **Reasoning Task Challenges**: All models struggle with reasoning tasks, revealing weaknesses in numerical computation and tax-related logic understanding. (6) **Example Introduction Pitfall**: One-shot examples may cause over-reliance on exemplars, hurting generalization across diverse reasoning tasks.

Our contributions are as follows: 1) **Introducing TaxPraBen**, the first benchmark specifically designed for Chinese tax practice, incorporating real-world scenarios such as risk prevention, inspection analysis, and strategy planning. 2) **Filling the gap in scarce tax datasets**, curated by domain experts and enhanced with ChatGPT-assisted annotations, providing high-quality data and well-designed prompts. 3) **Establishing a taxonomy for tax tasks**, organizing the dataset according to Bloom's cognitive taxonomy to assess capabilities in memorization, understanding, and application. 4) **Conducting in-depth evaluations of 19 popular LLMs**, revealing and discussing their strengths and weaknesses in tax-related tasks. 5) **Providing multi-evaluations across tax and NLP tasks**, helping analyze LLMs' shortcomings and potential strengths. 6) **Proposing a structured evaluation method** that extends beyond tax, combining semantic reasoning and numerical accuracy for practical assessment in law, healthcare, finance, and more.

2 Related Work

(1) Domain-specific LLMs. LLMs like OpenAI's GPT (Brown et al., 2020) and Meta's LLaMA (Touvron et al., 2023) excel in various NLP tasks but struggle with Chinese-specific applications due to their English-centric training. To address this, models like Gemini (Team et al., 2024), DeepSeek (Guo et al., 2025a), and Qwen (Bai et al., 2023) have been developed to better capture Chinese linguistic and cultural nuances. Domain-specific models such as PIXIU (Xie et al., 2023) and HuaTuo (Wang et al., 2023) show promise in fields like auditing (Jiajia et al., 2024), education (Yu et al., 2023), healthcare (Chen et al., 2023a), law (Zhou et al., 2024), finance (Hu et al., 2024), and psychological counseling (Chen et al., 2023b). Research has also examined integrating tax data into the pretraining or fine-tuning of LLMs (Luo et al., 2023) in legal and financial contexts. However, it remains uncertain how effectively these models can handle nuanced tax-related knowledge.

(2) Chinese LLM Benchmarks. Several benchmarks have been developed to evaluate Chinese LLMs. CLEVA (Li et al., 2023), C-Eval (Huang et al., 2023), and GAOKAO (Zhang et al.,

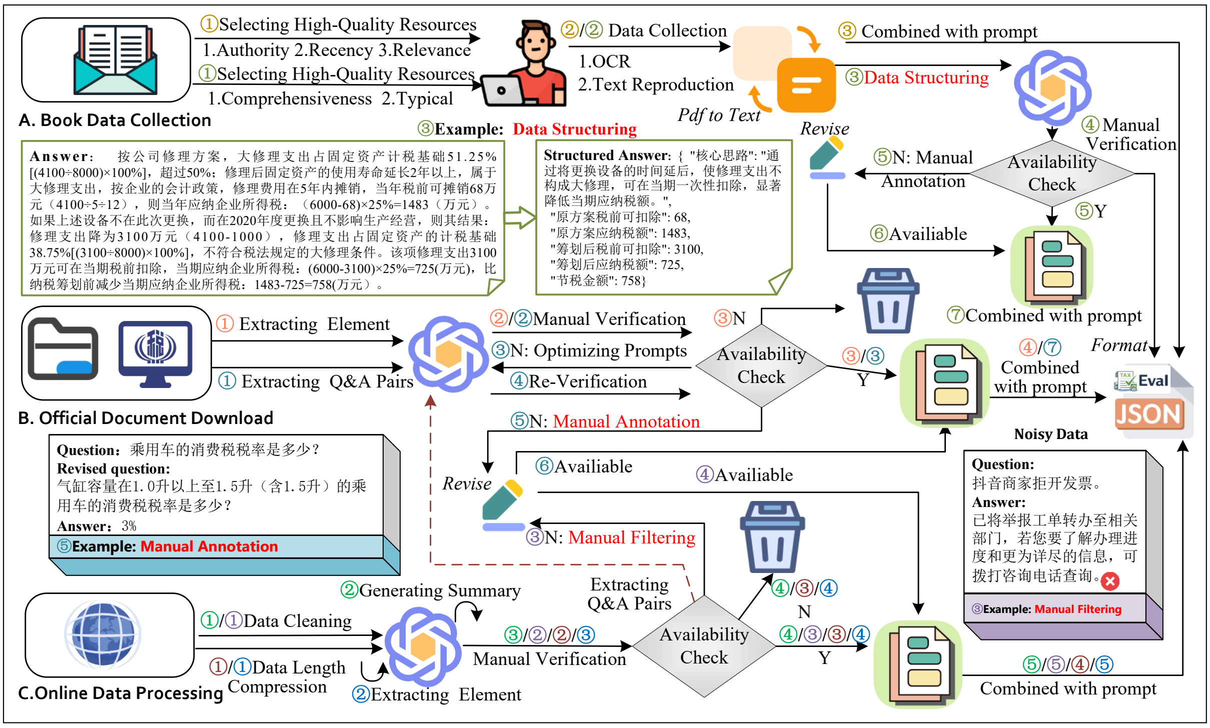


Figure 1: TaxPraBen's data construction workflow uses 3 methods: (A) Book Data Collection, (B) Official Document Download and (C) Online Data Processing. All data is annotated via the manual "Availability Check".

2023a) assess general knowledge and reasoning. MMCU (Zeng, 2023) and Xiezhi (Gu et al., 2024) focus on fields like medicine and law, while CG-Eval (Zeng et al., 2024) covers multiple disciplines. Domain-specific benchmarks include FinBen (Xie et al., 2024) for finance and CMB (Wang et al., 2024) for medicine. However, the tax domain is underrepresented, with datasets like LawBench (Fei et al., 2024) and LAiW (Dai et al., 2025) offering limited tax datasets, hindering the evaluation of tax reasoning abilities. Although TaxBen (Chen et al., 2025) addresses tax-related NLP tasks, it lacks the integration of semantic interpretation and numerical analysis. Thus, it is essential to refine the assessment framework for taxation competence in real-world application scenarios.

3 Overview design of TaxPraBen

3.1 Task Categorization

Establishing a scientific and systematic evaluation framework is crucial for assessing LLMs in the tax domain. We develop an evaluation system based on Bloom's Cognitive Taxonomy (Fei et al., 2024), which includes six cognitive levels. To simplify the evaluation, we draw from existing task taxonomies (Xie et al., 2024), concentrating on the first three levels to create three core tax capability assessments: 1) Knowledge Memorization (KM): Examines models' ability to accurately recall and recite tax legal provisions, policy regulations, and institutional documents; 2) Knowledge Understanding (KU): Tests models' capacity to identify key information from tax materials and comprehend policy implications; 3) Knowledge Application (KA): Evaluates models' comprehensive ability to apply tax regulations, policy clauses, and computational methods in addressing practical tax issues.

3.2 Data Processing

We construct the TaxPraBen benchmark using a multi-source data fusion strategy, comprising three acquisition pipelines: (A) Book Data Collection, (B) Official Document Download, and (C) Online Data Processing, as illustrated in Figure 1. These pipelines integrate tax-specific sources, ChatGPT-based processing, and rigorous human validation to ensure data quality, structural consistency, and alignment with real-world tax applications. The manual check process is shown in Appendix C.2.

(A) Book Data Collection: We select tax exam guidebooks and tax planning casebooks as data sources. The guidebooks are filtered based on authority, timeliness, and relevance, while the casebooks are selected according to comprehensiveness and typicality. Using OCR and text reproduction techniques, we extract various types of tax exam questions (*TaxSCQ*, *TaxMCQ* and *TaxCalc*) from

Tax Ability-Hierarchy Task	Dataset Name	Specific Task	Data Scale	Evaluation Metric	NLP- Task Type	Method	
						HG	MC
Tax Knowledge Memorization (KM)	TaxRecite	Tax Law Recitation	200	BERTScore, BARTScore	Generation (GEN)	✓	
Tax Knowledge Understanding (KU)	TaxSum	Tax News Summarization	1000	BERTScore, BARTScore	Generation (GEN)	✓	
	TaxTopic	Tax Topic Classification	1000	Accuracy, F1, Macro F1	Classification (CLS)		✓
	TaxRead	Tax Reading Comprehension	1000	EM Accuracy	Generation (GEN)	✓	
Tax Knowledge Application (KA)	TaxCalc	Tax Payment Calculation	500	EM Accuracy	Reasoning (REA)		✓
	TaxSCQ	Tax Single-Choice Exam	700	Accuracy, F1, Macro F1	Classification (CLS)		✓
	TaxMCQ	Tax Multiple-Choice Exam	400	EM Accuracy	Classification (CLS)		✓
	TaxQA	Tax Knowledge Q&A	700	BERTScore, BARTScore	Generation (GEN)	✓	
	TaxBoard	Tax Board Q&A	500	BERTScore, BARTScore	Generation (GEN)	✓	
	TaxCrime	Tax Law Identification	200	Accuracy, F1, Macro F1	Classification (CLS)	✓	
	TaxOpinion	Tax Opinion Summarization	500	BERTScore, BARTScore	Generation (GEN)	✓	
	TaxRisk	Tax Risk Prevention	200	BERTScore	Generation (GEN)	✓	
	TaxInspect	Tax Inspection Analysis	200	BERTScore, EM Accuracy	Generation (GEN)	✓	
	TaxPlan	Tax Strategy Planning	200	BERTScore, EM Accuracy	Reasoning (REA)	✓	

Table 1: Overview of tax tasks, datasets, specific tasks, data scale, evaluation metrics, NLP types, and creation methods in TaxPraBen. "MC" and "HC" refer to "Manually Created" and "Human-ChatGPT Collaborative".

PDF files. For tax planning cases, we employ ChatGPT to perform structured information extraction, followed by manual verification (*TaxPlan*).

(B) Official Document Download: We collect tax-related policy documents, regulations, and official announcements from the State Administration of Taxation. By integrating prompt engineering with ChatGPT, we generate question–answer (Q&A) pairs (*TaxRecite* and *TaxQA*). Each round of generation is followed by manual review, refinement of prompts based on identified issues, and expert validation for factual correction and completion. In addition, we retrieve criminal case verdicts from China Judgments Online and use ChatGPT to extract structured elements related to offense behavior and violated legal articles (*TaxCrime*).

(C) Online Data Processing: First, we crawl tax-related news, risk control reports, and audit cases from the Tax House website. Long-form texts are compressed using ChatGPT to meet input length constraints. For tax news, we extract title-topic pairs (TaxTopic) and article-summary pairs (*TaxSum*). For risk control and inspection cases, we extract structured key elements using ChatGPT (*TaxRisk* and *TaxInspect*). Second, we retrieve public tax-related opinion posts involving tax evasion from an online sentiment monitoring platform. A two-stage ChatGPT process is applied: first to assess tax relevance and clean noise, and second to compress the input and generate abstractive summaries (*TaxOpinion*). Third, we scrape user interactions with officials from the State Administration of Taxation's message board, using ChatGPT to filter out irrelevant or redundant content (*TaxBoard*). All generated content is manually validated for usability and alignment with tax-specific semantics.

Overall, the data definition and collection details for the three-level tax tasks in the evaluation benchmark TaxPraBen are shown in Appendix C.1.

3.3 Instruction Construction

To evaluate the performance of LLMs in various tax tasks, we collaborate with tax-related experts to annotate instructions for standardized outputs. We form a professional prompt annotation and evaluation team consisting of teachers and students. Evaluators score prompts on a 4-point scale ("Strongly Disagree (0)" to "Strongly Agree (3)"), focusing on accuracy (ACC), naturalness (NAT), and informativeness (INF). We test these prompts on the ChatALL platform. To ensure scoring reliability, we calculate consistency scores using Fleiss' Kappa and Krippendorff's Alpha, retaining only high-quality prompts with an average score exceeding 2. Detailed annotation processes and consistency analysis results are available in Appendix C.5. The final instructions combine input texts, labels, and validated prompts, evaluated in JSON format.

As shown in Appendix C.3, TaxPraBen features a three-tier taxonomy of tax tasks, with a balanced distribution of 200 to 1000 instances per type. This design enables fair evaluation across cognitive levels, while diverse inputs and prompt lengths enhance task complexity. Overall, TaxPraBen provides a comprehensive and challenging benchmark for assessing LLMs in the Chinese tax practice domain. Unlike most domain-specific benchmarks that rely solely on public datasets (Dai et al., 2025; Fei et al., 2024), TaxPraBen integrates expert design and manual annotation to ensure datasets reflect real tax scenarios and meet specific evaluation goals. Table 1 summarizes the task types, dataset statistics, and evaluation metrics in TaxPraBen.

3.4 Practical Relevance

To show the practical relevance of the Knowledge Application (KA) task in TaxPraBen, we map each tax dataset to the real-world scenario in Table 2. These tasks not only assess model performance

Dataset	Real-World Scenarios	Application Value
TaxCalc	Pre-filing tax calculation for individuals and businesses.	Supports the underlying logic for intelligent tax filing tools.
TaxSCQ	Exams for tax system staff or entry qualifications.	Enables training/testing for tax Q&A systems or AI tutors.
TaxMCQ	Complex policy applicability and compliance judgment.	Trains models for regulatory perception and decision-making.
TaxQA	Inquiries to tax bureaus (e.g., 12366), tax advisory.	Builds intelligent tax consultation systems.
TaxBoard	Message boards and online government services.	Enhances automatic response and assistant capabilities in e-government.
TaxCrime	Legal article application in law enforcement.	Supports case legality judgment and transparent law enforcement.
TaxOpinion	Public opinion monitoring and sentiment analysis.	Serves risk warning and government response mechanisms.
TaxRisk	Financial audits and tax risk control.	Builds AI systems for risk identification and response strategies.
TaxInspect	Extraction of key elements in enforcement cases.	Supports casebase development for training and enforcement.
TaxPlan	Tax-saving planning in enterprise operations.	Enables AI-assisted tax planning and optimization decisions.

Table 2: TaxPraBen subsets for Knowledge Application (KA) task: real-world scenarios and application value.

***Case 1*: TaxRisk (Tax Risk Prevention)**

{"面临的风险": 买受人无法取得增值税专用发票，影响进项税金抵扣权益，进而引发纠纷，浪费大量行政和司法资源。

"对应的解决措施": 加强税法和行政制度供给，完善增值税抵扣权制度，畅通司法诉讼救济途径，加强司法和行政部门的协作机制。}

***Case 2*: TaxInspect (Tax Inspection Analysis)**

{"犯罪行为": 虚开增值税专用发票，非法抵扣税款。

"所犯罪名": 虚开发票罪。

"处罚结果": 判处有期徒刑一年六个月，缓刑二年，并处罚金人民币13万元。}

***Case 3*: TaxPlan (Tax Strategy Planning)**

{"核心思路": 时间延后，使修理支出不构成大修理可在当期一次性扣除，显著降低当期应纳税额。

"原方案税前可扣除": 68, "原方案应纳税额": 1483

"筹划后税前可扣除": 3100, "筹划后应纳税额": 725

节税金额: 758}

Figure 2: A unified output format protocol for 3 typical cases of the tax practice scenarios.

but also simulate essential functions in modern tax workflows, including calculating tax liabilities, determining applicable policies, and developing tax-saving strategies. They support intelligent government systems through automated consultations and risk identification, while assisting regulatory enforcement with compliance assessments. This alignment underscores the value of the TaxPraBen benchmark in model performance evaluation and the development of reliable tax AI systems.

4 Structured Evaluation for TaxPraBen

4.1 Unified Output Protocol

In tax practice domain, assessing LLM outputs requires semantic accuracy and structural alignment with real-world use cases. TaxPraBen introduces a unified output protocol for various practice areas, enabling structured responses and automatic evaluation across LLMs. As shown in Figure 2, we define standardized JSON output schemas for 3 classic cases: (1) **TaxRisk**: Extract key elements from case narratives, including identified risks and solutions, and compute the average BERTScore for all fields. (2) **TaxInspect**: Perform a mixed matching using EM Accuracy for criminal charges and BERTScore for violation descriptions and penalties to ensure semantic precision. (3) **TaxPlan**: Create a tax planning strategy that integrates text generation and numerical reasoning, and assess semantic similarity and numerical accuracy. The unified output formatting protocol enhances the evaluability and reliability of model outputs by transforming free-form generations into structured data, enabling automatic scoring, cross-model comparisons, and batch evaluations, while also facilitating integration with real-world practice applications.

Task	Prompt Description
TaxRisk	你会收到一段模型的输出内容， 为了精准地匹配，你需要做的是判断模型有没有提取出内容。如果有，则从中提取模型输出的这两部分内容，如果模型本身就没有总结出来，你就留空，不需要替它总结，严格按照以下格式输出，留空的也需要保持格式：{"面临的风险": , "风险对应的解决方案": }。模型输出如下：{raw_output}。请只返回JSON对象，不要任何额外解释。
TaxInspect	你会收到一段模型的输出内容， 为了精准地匹配，你需要做的是判断模型有没有提取出内容。如果有，则从中提取模型输出的这三部分内容，如果模型本身就没有总结出来，你就留空，不需要替它总结，严格按照以下格式输出，留空的也需要保持格式：{"犯罪行为": , "所犯罪名": , "处罚结果": }。模型输出如下：{raw_output}。请只返回JSON对象，不要任何额外解释。
TaxPlan	请你从模型输出中提取填空答案，若某一字段没有找到答案，则该字段留空，不要凭空生成答案，并严格按照以下格式输出：{"核心思路": , "会计利润": , "第一方案所得税": , "第二方案所得税": , "分年捐赠筹划后所得税": , "节税额": }，除核心思路外，其余字段都只返回数值，不要有任何符号和单位。模型输出如下：{raw_output}。请只返回JSON对象，不要任何额外解释。

Table 3: Structured output guidance prompts.

4.2 Aligned Field Evaluation

Despite having explicit output templates, most open-source LLMs often still produce inconsistent outputs, such as missing fields, malformed JSON, or mixed text. This hinders automated evaluation for structure-sensitive tax practice scenarios: TaxRisk, TaxInspection, and TaxPlan, as well as cross-model comparisons and batch assessments.

To tackle this issue and reduce API costs, we leverage ChatGPT-3.5 as a text structure-aware parser. It outperforms fragile regex with strong semantic comprehension and tolerance for formatting noise, requiring no fine-tuning unlike supervised structure predictors. This makes it a lightweight, adaptive solution for normalizing model outputs. Structured output guidance prompts for the 3 practice tax scenarios are shown in Table 3. As seen in Appendix F.1, targeted prompting allows ChatGPT to fully parse unstructured content, strip redundant text, and extract all critical fields. We then conduct automated evaluations following the metric calculation methodology in Section 6.2.

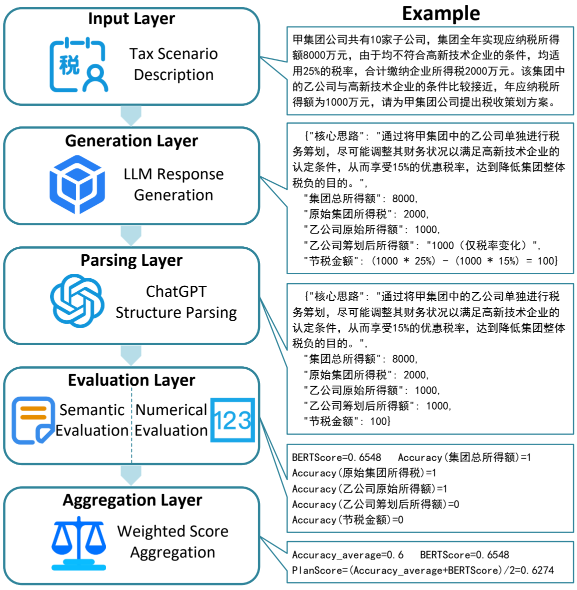


Figure 3: Structured evaluation pipeline for TaxPlan.

This strategy reduces evaluation sensitivity to output structure, ensuring that minor deviations do not interfere with results. It is model-agnostic, compatible with both commercial and open-source LLMs, enabling unified cross-system assessments while reducing reliance on brittle rule-based scripts and manual templates, thus making the evaluation pipeline more scalable and maintainable.

4.3 Real-World Practice Case

We take TaxPlan as an example to present our structured evaluation pipeline in Figure 3: Feed tax scenario descriptions to target LLMs, with unified output formatting prompts to guide them to generate predefined JSON-style responses. Since open-source models often fail to follow formatting rules, we use ChatGPT as a structure parser to clean noisy outputs, convert them into valid JSON and extract key fields. Conduct two-dimensional evaluation: use BERTScore to assess text description quality, and EM Accuracy to check numerical matching precision. Calculate the final score via weighted average to fully reflect both reasoning rationality and quantitative accuracy of the model outputs.

5 Scalable Promotion of TaxPraBen

Our structured evaluation framework, though originally developed for tax planning tasks, demonstrates significant extensibility across domains. Many real-world scenarios require models to generate hybrid outputs that integrate free-text reasoning with critical numerical data. Traditional evaluation methods, which typically focus solely on textual fluency, lexical overlap, or single-token accuracy, often fail to comprehensively assess such responses in terms of both semantic completeness and numerical correctness. Our proposed structured tax evaluation methodology is, in fact, generalizable to numerous professional domains such as law, finance, and healthcare (as demonstrated in Appendix D)—all fields that similarly require the generation of hybrid content combining combining "explanatory text + critical numbers". This approach effectively addresses significant gaps in conventional single-metric evaluation pipelines.

6 Experiments

6.1 Baseline Modes

We assess representative LLMs based on their availability, popularity, and performance across various benchmarks. They are listed in Table 4, with detailed descriptions provided in Appendix F.2.

Model	P	Pre	Fin	Access	Base LLM	Release
Multilingual General LLMs						
ChatGPT	175B	✓	×	API	–	11/2022
GPT-4o	200B	✓	×	API	–	05/2024
MistralV0.3	7B	✓	×	Weights	–	07/2024
Gemma	7B	✓	×	Weights	–	07/2024
LLaMA3	8B	✓	×	Weights	–	09/2024
Bayling2	7B	✓	×	Weights	LLaMA2	11/2024
Grok3	1200B	✓	×	API	–	02/2025
Chinese-oriented LLMs						
$\text{DeepSeek}_{\text{llm}}$	7B	✓	×	Weights	–	11/2023
Baichuan2	7B	✓	×	Weights	–	01/2024
Atom	7B	×	✓	Weights	LLaMA2-7B	02/2024
Qwen2.5	7B	✓	×	Weights	–	02/2024
ChnLLaMA3	8B	×	✓	Weights	LLaMA3-7B	04/2024
ERNIE-3.5	∼ 1000B	✓	×	API	–	07/2024
ChatGLM3	6B	✓	×	Weights	–	08/2024
Yi	6B	✓	×	Weights	–	11/2024
GLM4	9B	✓	×	Weights	–	11/2024
$\text{DeepSeek}_{\text{R1}}$	7B	✓	×	Weights	–	02/2025
InternLM2.5	7B	✓	×	Weights	–	03/2025
Tax-related LLMs						
YaYi2	30B	✓	×	Weights	–	03/2024

Table 4: Various baseline LLMs, including parameters (P), pre-trained (Pre) or fine-tuned (Fin), access methods (public weights/API calls), base LLM, and release date.

6.2 Evaluation Metrics

In accordance with (Xie et al., 2023; Hu et al., 2024), we adopt a unified JSON template (see Table 5) to organize task-specific evaluation metrics for the 5 output types in TaxPraBen, as detailed below: **(a)** *Text Classification Tasks*: Datasets like TaxTopic, TaxSCQ, and TaxCrime involve discrete label prediction. We use Accuracy (ACC), F1 score, and Macro-F1 to measure performance. **(b)** *Text Generation Tasks*: Datasets like TaxRecite, TaxSum, TaxQA, TaxBoard, TaxOpinion, and TaxRisk require free-form or structured textual responses. For these, we apply BERTScore and BARTScore to assess semantic similarity. In TaxRisk and TaxInspect, which generate multiple semantic aspects, we compute the average BERTScore. **(c)** *Struc-*

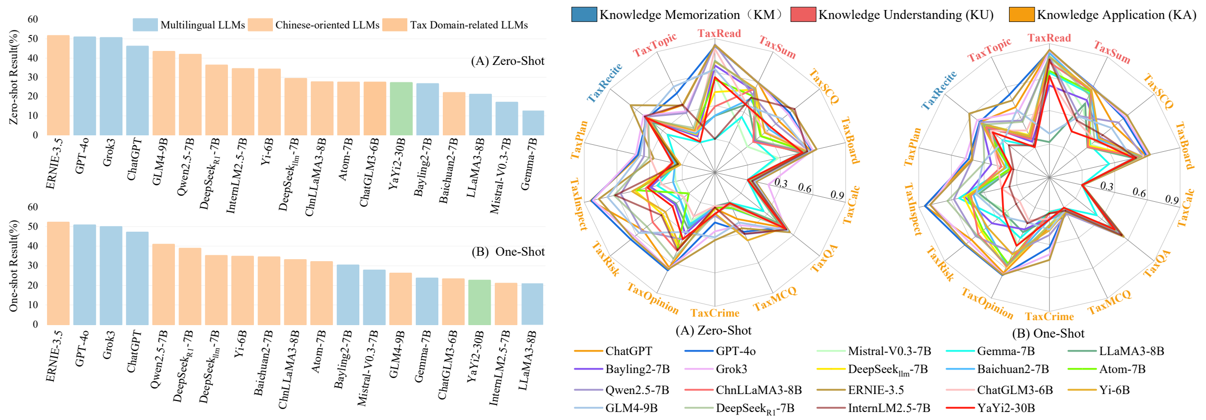


Figure 4: The zero-shot and one-shot overall performance of the 19 popular LLMs evaluated on TaxPraBen.

tured Prediction Tasks: Datasets like TaxRead, TaxCalc, and TaxMCQ require deterministic outputs in fixed formats (e.g., numbers or texts). We evaluate them using Exact Match (EM) Accuracy. **(d)** *Mixed Matching Task*: TaxInspect and TaxPlan combine absolute matches for numerical/text fields and semantic matches for explanatory rationale, averaging EM Accuracy and BERTScore for the final score. To ensure fair assessment across tasks and data types, we compute the overall average of metrics from 5 runs, as defined in Appendix F.3. All overall average metrics range from 0 to 1, with higher values indicating better performance.

{**Id**:[data_id] **query**:[prompt] **text**:[text] **answer**:[response] (a) **choices**:[...] **gold**:[...] / (b) — / (c) — / (d) — }

Table 5: The JSON template for each task evaluation

6.3 Inference Settings

We use the LLM Evaluation Harness (Gao et al., 2021) to build our inference and evaluation pipelines. For semantic similarity evaluation, we employ the Chinese-supporting models: chinese-xlnet-base for BERTScore and bart-large-cnn for BARTScore. API-based LLMs like ChatGPT, GPT-4o, Grok3, and ERNIE-3.5 are evaluated via the YunWu platform, while open-source LLMs are assessed on a cluster with 4 A100 GPUs (80GB). We set the input length limit to 2048 tokens. The maximum generation length is 300 tokens for Q&A and summarization datasets, and 200 tokens for others.

7 Results

7.1 Model Overall Performance

As illustrated in Figure 4, the following findings can be drawn: **(1) Closed-source LLMs Show Superior Performance:** In all settings, closed-source LLMs like ERNIE-3.5, GPT-4o, Grok3, and ChatGPT consistently rank in the Top 3, with ERNIE-3.5 particularly excelling. This is attributed to their large parameter sizes and extensive training datasets. Notably, ERNIE-3.5 benefits from the integration of domain-specific knowledge, which further enhances its performance on tax-related tasks. **(2) Chinese-oriented LLMs Outperform Multilingual LLMs:** Within the open-source group, Chinese-oriented LLMs pretrained predominantly on Chinese corpora consistently outperform multilingual LLMs. This trend reflects the language-specific nature of tax-related tasks, which often involve Chinese regulatory terminology, institutional jargon, and culturally contextual reasoning. As a result, LLMs with strong Chinese pretraining demonstrate superior alignment with task semantics and domain expressions. **(3) Domain Fine-tuning Doesn't Guarantee Strong Results:** Surprisingly, the only tax-related LLM, YaYi2, performs poorly compared to general LLMs. Despite fine-tuning on tax data, it lags behind some smaller open-source LLMs. We believe this is mainly due to two factors: limited quantity and diversity of tax training data, and significant differences between the fine-tuning content and evaluation tasks. Without sufficient data variety and task alignment, domain-specific fine-tuning may still fail. **(4) Zero-Shot and One-Shot Performance Varies Significantly:** Among the 19 models studied, 11 show improvement in the one-shot setting, effectively utilizing in-context learning with task-specific examples. However, LLMs like GLM4 and InternLM2.5 show performance declines, likely due to longer prompts exceeding attention limits and weaker instruction alignment. These results underscore the impact of prompt design and model architecture on few-shot learning, suggesting that one-shot improvements are not universally applicable.

Type	Settings	Name	ChatGPT	GPT-4o	MistralV0.3	Gemma	LLaMA3	Bayling2	Grok3	$\text{DeepSeek}_{\text{llm}}$	Baichuan2	Atom	Qwen2.5	ChnLLaMA3	ERNIE-3.5	ChatGLM3	Yi	GLM4	$\text{DeepSeek}_{\text{R1}}$	InternLM2.5	YaYi2
Tax Task	Zero-Shot	KM	0.488	0.478	0.400	0.244	0.412	0.466	0.519	0.494	0.470	0.474	0.499	0.445	**0.667**	0.487	0.482	0.480	0.461	0.505	0.485
		KU	0.602	**0.637**	0.277	0.085	0.236	0.409	0.579	0.341	0.243	0.370	0.538	0.312	0.599	0.427	0.449	0.515	0.455	0.273	0.307
		KA	0.415	0.472	0.114	0.127	0.183	0.202	**0.482**	0.258	0.188	0.225	0.375	0.247	0.475	0.206	0.295	0.404	0.324	0.349	0.239
	One-Shot	KM	0.497	0.499	0.413	0.239	0.407	0.475	0.527	0.474	0.480	0.467	0.494	0.426	**0.619**	0.372	0.476	0.225	0.461	0.165	0.350
		KU	0.611	**0.653**	0.429	0.393	0.182	0.350	0.591	0.488	0.491	0.439	0.453	0.460	0.633	0.341	0.485	0.257	0.520	0.348	0.257
		KA	0.425	0.464	0.217	0.188	0.197	0.271	0.468	0.297	0.286	0.269	0.386	0.281	**0.477**	0.186	0.293	0.265	0.340	0.172	0.203
NLP Task	Zero-Shot	CLS	0.248	0.370	0.042	0.042	0.072	0.055	0.325	0.088	0.120	0.100	0.273	0.067	**0.379**	0.104	0.207	0.375	0.100	0.327	0.107
		GEN	0.626	0.644	0.266	0.193	0.327	0.426	0.645	0.458	0.316	0.423	0.547	0.436	**0.670**	0.388	0.489	0.515	0.542	0.428	0.410
		REA	0.220	0.239	0.037	0.025	0.025	0.041	**0.308**	0.041	0.033	0.023	0.194	0.048	0.169	0.156	0.020	0.223	0.164	0.041	0.045
	One-Shot	CLS	0.261	0.359	0.094	0.047	0.077	0.106	0.335	0.114	0.122	0.076	0.258	0.065	**0.383**	0.118	0.146	0.244	0.152	0.095	0.078
		GEN	0.639	0.652	0.428	0.377	0.321	0.447	0.646	0.550	0.533	0.514	0.533	0.526	**0.668**	0.325	0.522	0.295	0.565	0.315	0.348
		REA	0.210	0.223	0.034	0.048	0.023	0.117	**0.236**	0.027	0.033	0.030	0.205	0.071	0.207	0.088	0.053	0.155	0.149	0.015	0.025

Table 6: The performance of the 19 LLMs on Tax task (KM, KU, KA) and NLP main task (CLS, GEN, REA).

7.2 Tax Task Discrepancy

We observe the following insights from Table 6: (1) **In KM Task:** ERNIE-3.5 performs best due to its knowledge-enhanced pre-training strategy, while other models face significant bottlenecks. This is largely attributed to the high specialization and frequent updates of tax law knowledge, which means most LLMs are trained on static corpora, potentially leading to outdated knowledge. Additionally, many models show decreased performance in one-shot settings, indicating that a single example may disrupt the model's inherent memory capabilities. (2) **In KU Task:** GPT-4o, ChatGPT, and ERNIE-3.5 excel, particularly in comprehending tax-related texts. The locally deployed Qwen2.5 performs relatively well, but other models underperform, highlighting challenges in handling complex semantics and specialized terminology. When comparing zero-shot and one-shot settings, most LLMs show improved performance in one-shot scenarios, as this approach helps them focus on key features. (3) **In KA task:** Closed-source LLMs perform well but show the weakest overall results. This is because applying tax knowledge requires LLMs to possess stronger reasoning abilities, which poses greater challenges. The complexity of the tax domain necessitates that models accurately apply policies based on specific scenarios and taxpayer identities. This case-by-case application significantly increases difficulty. Compared to zero-shot setting, most LLMs perform better in one-shot scenarios, as examples help clarify output formats and task objectives, thereby enhancing the stability and accuracy of their responses.

7.3 NLP Task Discrepancy

Table 6 presents the following conclusions: **(1) CLS Task Highlights Specialized Demands:** The CLS task shows a strong need for expertise, underperforming compared to (Hu et al., 2024; Jiajia et al., 2024) on expected simple multiple-choice questions. This is due to complex tax case analyses and calculations. Models like GPT-4o, Grok3, and ERNIE-3.5 excel, with Chinese-focused LLMs such as InternLM2.5 and GLM4 following closely. **(2) GEN Task Showcases Proficiency:** LLMs demonstrate a relatively strong capability in the GEN task, where ERNIE-3.5 takes the lead, supported by Grok3 and GPT-4o. Among open-source LLMs deployed locally, Qwen2.5 and $\text{DeepSeek}_{\text{R1}}$ stand out. Their superior performance is attributed to their larger parameter sets and enhanced support for Chinese, making them well-suited for producing tax-related textual content. **(3) REA Task Reveals Reasoning Limitations:** The REA task presents significant challenges requiring deep understanding, common sense, and multi-step reasoning. Most LLMs perform poorly, with little improvement in one-shot scenarios due to difficulties in adapting calculation patterns. The complexity of multi-step reasoning, especially with tax formulas, remains a hurdle, even for closed-source large-parameter LLMs like Grok3 and ERNIE-3.5.

8 Conclusion

In this study, we introduce TaxPraBen, a pioneering practical benchmark for evaluating LLMs in the Chinese tax domain across 3 cognitive levels: Knowledge Memorization, Understanding, and Application. By incorporating 10 diverse real-world practice tasks derived from 14 datasets, TaxPraBen assesses 19 leading LLMs, revealing notable performance gaps. TaxPraBen's contributions include a rare practical dataset, the establishment of a taxonomy, the design of comprehensive metrics, and a structured evaluation approach, positioning it as a vital resource for advancing research on LLM applications in the Chinese tax domain.

Limitations

While TaxPraBen offers a valuable benchmark for LLMs in Chinese tax practice, it has limitations: (1) **Risk of Data Leakage**: Most datasets are sourced from the internet and public books. While this data collection method is convenient, it also increases the risk of test data leakage. Existing LLMs have been trained on vast amounts of internet data, which means they may inadvertently memorize and reproduce content from these datasets. This situation can lead to models achieving exceptionally high scores during evaluations, but such scores do not accurately reflect the models' reasoning abilities or their effectiveness in real-world applications. (2) **Limitations of Model Parameters**: Current evaluations primarily focus on mainstream LLMs with similar parameters, which introduces limitations when comparing the performance of different models. The lack of comparisons with larger open-source LLMs restricts deeper insights into the impact of various parameters. Models of different scales and architectures may exhibit varying capabilities when handling specific tasks, yet these differences have not been thoroughly explored within the current evaluation framework. (3) **Shortcomings of Automated Evaluation**: Although automated semantic similarity assessment metrics for generative tasks offer convenience, they may not accurately reflect human judgments regarding answer quality. Human evaluators consider a broader range of factors, including relevance, accuracy, and contextual suitability, when assessing answers. Therefore, relying solely on these automated metrics can lead to misjudgments about model performance.

Ethics Statement

Given the sensitivity of personal privacy and commercial confidentiality in the tax domain, we conduct a comprehensive data review and ethical considerations for the TaxPraBen benchmark study. We implement strict anonymization measures for the collected data, including the removal of company names and personal identifiers to protect privacy. Additionally, we utilize relevant open-source resources to ensure transparency and ethical compliance. Legal experts evaluate our research methods to ensure adherence to data protection standards. We are committed to maintaining the confidentiality of all stakeholders, ensuring that our research respects personal privacy while making a positive contribution to the field of tax practice. Furthermore, the annotation work is completed voluntarily by individuals who receive no compensation. All annotators are informed of the intended use of the data and give consent to participate.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024a. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and others. 2024b. Internlm2 technical report. https://arxiv.org/abs/2403.17297.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4310–4330.

Yating Chen, Siqi Lv, Peiyuan Xia, Zhenxu Wang, Yiming Qin, Qingqing Wang, and Gang Hu. 2025. Taxben: Benchmarking the chinese tax knowledge of large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 307–321. Springer.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, and others. 2023a. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273*.

Eunkyung Choi, Young Jin Suh, Hun Park, and Wonseok Hwang. 2025. Taxation perspectives from large language models: A case study on additional tax penalties. *arXiv preprint arXiv:2503.03444*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. Laiw: A chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, pages 10738–10766.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7933–7962.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and others. 2021. A framework for few-shot language model evaluation. *Zenodo*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, and others. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18099–18107.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, and others. 2025b. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6258–6292.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2009. Measuring massive multitask language understanding, 2021. *URL https://arxiv. org/abs*, page 20.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Gang Hu, Ke Qin, Chenhan Yuan, Min Peng, Alejandro Lopez-Lira, Benyou Wang, Sophia Ananiadou, Jimin Huang, and Qianqian Xie. 2024. No language is an island: Unifying chinese and english in financial large language models, instruction data, and benchmarks. *arXiv preprint arXiv:2403.06249*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Huang Jiajia, Zhu Haoran, Xu Chao, Zhan Tianming, Xie Qianqian, and Huang Jimin. 2024. Auditwen: An open-source large language model for audit. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1351–1365.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. Bizbench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8309–8332.

Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael Lyu, and others. 2023. Cleva: Chinese language models evaluation platform. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 186–217.

Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, and others. 2024. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*, 7(4):1116–1128.

Yin Luo, Qingchao Kong, Nan Xu, Jia Cao, Bao Hao, Baoyu Qu, Bo Chen, Chao Zhu, Chenyang Zhao, Donglei Zhang, and others. 2023. Yayi 2: Multilingual open-source large language models. *arXiv preprint arXiv:2312.14862*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Ziyang Miao, Qiyu Sun, Jingyuan Wang, Yuchen Gong, Yaowei Zheng, Shiqi Li, and Richong Zhang. 2025. Easy dataset: A unified and extensible framework for synthesizing llm fine-tuning data from unstructured documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 960–968.

Louis Brulé Naudet. 2023. Livre des procédures fiscales, non-instruct (11-12-2023). https://hf-mirror.com/datasets/louisbrulenaudet/lpf.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource greek finance. arXiv preprint arXiv:2502.18772. https://arxiv.org/abs/2502.18772.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Thomas Rixen and Brigitte Unger. 2022. Taxation: A regulatory multilevel governance perspective. *Regulation & Governance*, 16(3):621–633.

Daniel Steinigen, Marcin Namysl, Markus Hepperle, Jan Krekeler, and Susanne Landgraf. 2023. Semantic extraction of key figures and their properties from tax legal texts using neural models. In *ASAIL@ ICAIL*, pages 60–71.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and others. 2024. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6184–6205.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, and others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36:33469–33484.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng

Li, Jiangcheng Zhu, Jianqun Chen, and others. 2024. Yi: Open foundation models by 01. ai. https://arxiv.org/abs/2403.04652.

Jingsi Yu, Junhui Zhu, Yujie Wang, Yang Liu, Hongxiang Chang, Jinran Nie, Cunliang Kong, R Chong, Xin Liu, Jiyuan An, and others. 2023. Taoli llama.

Wanlong Yu, Wei Wan, Zhenxu Wang, Feng Li, Kang Wang, and Gang Hu. 2025. Open bilingual benchmark and leaderboard for large language models in cybersecurity. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 456–470. Springer.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, and others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2024. Withdrawn: Evaluating the generation capabilities of large chinese language models.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, Shoutao Guo, Yan Zhou, Xiaodong Liu, and Yang Feng. 2024. Bayling 2: A multilingual large language model with efficient language alignment. *arXiv preprint arXiv:2411.16300*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023a. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Xuanyu Zhang, Bingbing Li, and Qing Yang. 2023b. Cgce: A chinese generative chat evaluation benchmark for general and financial domains. *arXiv preprint arXiv:2305.14471*.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. 2024. Atom: Low-bit quantization for efficient and accurate llm serving. *Proceedings of Machine Learning and Systems*, 6:196–209.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2406.04614*.

A Why important

A.1 Benchmark Comparison

In recent years, numerous LLM benchmarks have been introduced for vertical domains like law (LexGLUE (Chalkidis et al., 2022), DISC-LawLLM (Yue et al., 2023), LawBench (Fei et al., 2024), LAiW (Dai et al., 2025)), finance (Fineval (Guo et al., 2025b), CGCE (Zhang et al., 2023b), FLARE (Xie et al., 2023), FinBen (Xie et al., 2024), ICE-FLARE (Hu et al., 2024)), and healthcare (DISC-MedLLM (Bao et al., 2023), MedMCQA (Pal et al., 2022), CMB (Wang et al., 2024), MedBench (Liu et al., 2024)) (see Table A1). Most of the focus on traditional NLP tasks (including MCQ, Q&A, REA, NR) to assess the transfer of general language understanding to specialized fields. However, they primarily evaluate whether LLMs can answer general questions, often neglecting the assessment of real-world problem-solving and complex decision-making abilities.

Name	Method		NLP Specific Task					PRA	CoMe
	HC	IPA	MCQ	Q&A	CLS	REA	NR		
Law									
LexGLUE	×	✓	✓	×	✓	×	×	×	×
DISC-LawLLM	✓	✓	✓	×	×	×	×	×	×
LawBench	×	✓	✓	✓	✓	✓	×	×	×
LAiW	×	✓	×	✓	✓	✓	×	×	×
Finance									
Fineval	×	✓	✓	✓	✓	✓	×	×	×
CGCE	×	✓	×	✓	×	×	✓	×	×
FLARE	×	✓	×	✓	✓	✓	×	×	×
FinBen	×	✓	✓	✓	✓	✓	×	×	×
ICE-FLARE	×	✓	✓	✓	✓	✓	×	×	×
Medicine									
DISC-MedLLM	✓	✓	✓	✓	×	×	×	×	×
MedMCQA	✓	×	✓	×	×	×	×	×	×
CMB	✓	✓	✓	✓	×	✓	×	×	×
MedBench	✓	✓	✓	✓	✓	✓	×	✓	×
Taxation									
TaxBen	✓	×	✓	✓	✓	✓	✓	×	×
TaxPraBen	✓	×	✓	✓	✓	✓	✓	✓	✓

Table A1: Comparing different domain-specific benchmarks, including data construction methods (human construction, HC; improved public data, IPA), task coverage across NLP specific tasks (multiple-choice questions, MCQ; question and answer, Q&A; text reasoning, REA; numerical reasoning, NR), practical application task (PRA) and the use of combined metrics (CoMe). "✓" indicates presence, and "×" indicates absence.

In contrast, TaxPraBen is the first benchmark designed for Chinese tax practice, marking a fundamental shift from task-oriented to capability-oriented evaluation. Its main innovation is expanding the focus from isolated language tasks to practice-oriented, integrated tasks that cover key skills such as legal interpretation, numerical reasoning, and risk assessment, making it much more aligned with real-world professional needs. For instance, traditional TaxBen's NR task is limited to basic questions like "*What is the tax rate for income bracket X*?". In contrast, TaxPraBen's PRA task require a comprehensive question: "*Calculating a taxpayer's final tax liability and identifying potential audit risks based on income sources, deductions, and regional regulations*", thus replicating the full professional workflow and greatly enhancing practical relevance. It highlights TaxPraBen's innovative use of a combined metric to jointly assess the accuracy of tax amount calculations and the semantic similarity of audit risk identification. Moreover, TaxPraBen introduces a breakthrough in data construction. Unlike most benchmarks that rely on public datasets, TaxPraBen is built entirely on authentic cases from tax professionals, combining manual curation with LLM-based prompt engineering. This ensures high-quality, realistic scenarios and strengthens its practical relevance.

In summary, as the first truly comprehensive benchmark for Chinese tax practice, TaxPraBen stands out by shifting from task- to capability-oriented evaluation, filling the gap in high-quality tax data, and redefining structured assessment systems. This not only advances the use of LLMs in tax but also provides valuable insights for developing benchmarks in other specialized domains.

A.2 TaxBen Contrast

In the field of tax artificial intelligence, the application of LLMs is gaining attention. However, due to their inherent opacity, tax experts remain cautious about their real-world applicability. While TaxBen advances benchmarking in this domain, it focuses narrowly on isolated NLP tasks-such as Q&A, text generation & classification, and numerical reasoning—neglecting to evaluate models' practical capabilities in real-world scenarios.

As shown in Table A2, apart from closed-source large-parameter LLMs (including ERNIE-3.5, Grok-3, and ChatGPT), several LLMs like ChatGLM3 and Atom—exhibit inflated rankings on the simpler NLP tasks covered by TaxBen, but see significant declines in TaxPraBen, which includes real-world application tasks. The notable discrepancy in model rankings between TaxPraBen and

LLM Type	Name	Zero-Shot		One-Shot	
		TaxBen	TaxPraBen	TaxBen	TaxPraBen
GLS_{LLM}	ERNIE-3.5	1	1	1	1
	Grok-3	2	2	2	2
	ChatGPT	3	3	3	3
COS_{LLM}	ChatGLM3	8	12	13	15
	$DeepSeek_{llm}$	12	9	8	6
	Atom	10	11	9	10
Tax_{LLM}	YaYi2	13	13	16	16

***Note**: To ensure fair model rankings, we consider only jointly evaluated LLMs.*

Table A2: Comparison of the evaluation ranking results of various LLMs between TaxPraBen and TaxBen. General closed-source LLMs: GLS_{LLM}, Chinese open-source LLMs: COS_{LLM}, Tax-related LLM: TAX_{LLM}.

TaxBen highlights the limitations of TaxBen in addressing genuine practical tasks. While TaxBen does include numerical reasoning and question-answering tasks, can the practical capabilities of the models be adequately reflected solely by averaging the metrics from these two tasks? Considering that real-world practical tasks require both semantic matching and computational reasoning abilities, the answer is clearly no. Further analysis, as illustrated in Table A3, supports this conclusion. For instance, models like InternLM2.5, Yi, and ERNIE-3.5 perform excellently and rank highly on the numerical-reasoning (NR data) and semantic-matching (Q&A data) tasks in TaxBen. Yet, on the TaxPlan task in TaxPraBen—which demands the integration of both capabilities—their metrics decline markedly, with most rankings dropping to 16th-18th ranking positions, showing an average decline of 5–11 positions (a total of 18). This clearly demonstrates that the practical ability of LLMs cannot be represented simply by weighting the evaluation metrics of individual NLP subtasks.

These findings further reveal that existing domain-specific benchmarks, with their emphasis on NLP-task evaluation, exhibit significant shortcomings in assessing real-world applicability. Thus, we propose TaxPraBen, the first benchmark designed to evaluate LLMs in Chinese tax-practice scenarios. TaxPraBen incorporates 10 datasets (see Table 2) covering real-world tasks such as tax-risk prevention, tax-inspection analysis, and tax-strategy planning, thereby enabling a systematic assessment of models' comprehensive applied ability in complex tax-practice settings.

B Taxonomy Motivation

TaxPraBen employs Bloom's Taxonomy as its evaluation system, systematically assessing the

LLM	TaxBen					TaxPraBen		
	NR Data Avg.Val	R	Q&A Data Avg.Val	R	Avg .R	TaxPlan Avg.Val	R	R(↓)
Zero-Shot								
InternLM2.5	0.042	5	0.520	4	5	0.039	16	11
Yi	0.024	7	0.495	9	8	0.016	18	10
ERNIE-3.5	0.076	2	0.559	1	2	0.263	7	5
One-Shot								
InternLM2.5	0.000	7	0.529	4	6	0.031	18	12
ERNIE-3.5	0.002	2	0.487	9	6	0.052	15	9
Qwen2.5	0.000	7	0.421	17	12	0.045	17	5

***Note**: To ensure fair model rankings, we consider only jointly evaluated LLMs and round the averag of the metric rankings (Avg.R) for NR and Q&A data.*

Table A3: Comparison of metric rankings on TaxBen (numerical reasoning (NR data)+semantic matching (Q&A data)) and TaxPraBen (tax practice (TaxPlan)). Metric Ranking: R, Average Value. "Avg.R" denotes average of the metric rankings for NR and Q&A data.

mastery of tax knowledge in LLMs through a hierarchical cognitive ability structure. This framework categorizes tax competencies into 3 progressive dimensions: *Knowledge Memorization (KM)*, K*nowledge Understanding (KU)*, and *Knowledge Application (KA)*. In the specialized tax domain, this tiered approach first requires models to accurately recall legal provisions (e.g., TaxRecite tests verbatim clause reproduction), then interpret policy implications (e.g., TaxRead evaluates reading comprehension of regulations), and finally apply knowledge practically (e.g., TaxPlan examines tax planning strategy design). Bloom's Taxonomy offers 3 key advantages: (1) Clear hierarchical standards enable precise quantification of knowledge depth; (2) Reveal inter-level relationships helps diagnose fundamental skill gaps; (3) Adapt to tax practice complexity, it identifies cross-level weaknesses. Compared to traditional one-dimensional difficulty grading, this structured method provides targeted optimization guidance (e.g., enhancing legal memorization or tax reasoning abilities). Its adoption marks a paradigm shift from evaluating task completion to knowledge mastery.

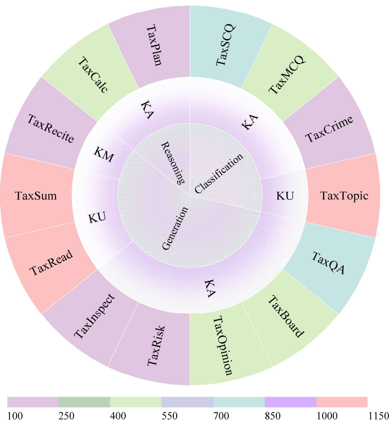
C Instruction Details

C.1 Collection Process

The definition and data collection details of the datasets for the three-level tax tasks in our proposed TaxPraBen benchmark are shown in Table C1. These datasets span three core tax task levels (KM: Knowledge Memorization, KU: Knowledge Understanding, KA: Knowledge Application) with 14

Dataset Source	URL
Tax House website	`https://www.shui5.cn/`
Tax Practice Cases	`https://wxredian.com/art?id=ecf5c4f6a9986a6eda47f48ba7ed5e74`
Certified Tax Agent Exam	`http://www.canet.com.cn/sws/zhinan/kaoshijiaocai.html`
Opinion Monitoring System	`https://open-yuqing.stonedt.com/`
Taxation Administration Website	`https://www.chinatax.gov.cn/`

Table C2: The original source along with the URL of our used datasets.



Dataset	Zero-Shot (Prompt)			One-Shot (Prompt)			Input Text			Label		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
TaxRecite	42	37	39.06	592	109	206.80	69	33	40.11	514	17	115.05
TaxSum	38	35	36.37	1469	273	573.60	1335	130	405.50	236	14	115.12
TaxTopic	130	125	128.02	207	120	150.50	100	6	37.26	1	1	1.00
TaxRead	56	49	53.82	1812	221	566.30	2066	125	521.85	286	1	15.44
TaxCalc	86	65	69.80	1383	327	636.40	1266	216	535.48	9	4	5.84
TaxSCQ	68	55	59.93	494	111	218.50	328	35	129.10	1	1	1.00
TaxMCQ	146	136	141.14	587	153	319.50	469	42	140.94	4	2	3.15
TaxQA	38	29	34.75	724	60	126.40	149	7	30.03	532	2	46.78
TaxBoard	48	41	45.28	1081	109	299.10	474	23	99.58	814	11	143.32
TaxCrime	219	201	208.72	521	303	387.13	299	24	140.60	2	1	1.13
TaxOpinion	53	38	43.82	2235	825	1386.05	1646	502	987.79	445	163	298.24
TaxRisk	112	68	84.88	1873	329	1110.60	1597	153	923.00	302	40	110.76
TaxInspect	198	171	179.13	1603	411	930.92	1149	156	627.77	213	32	105.22
TaxPlan	354	83	150.65	1071	590	719.80	432	57	150.15	394	65	141.91

Table C3: TaxPraBen includes 3 NLP and tax tasks across 14 datasets ranging from 200 to 1,000 entries, along with statistics on the maximum (Max), minimum (Min), and average (Avg) lengths of prompts and input/labels.

specific datasets. Each dataset details its task definition and collection sources, official tax documents, the website of the State Taxation Administration, Tax House articles, certified tax agent exam materials, and real-world tax practice cases, with some leveraging ChatGPT-3.5 as an auxiliary tool for text compression or question-answer generation. For the convenience of researchers, Table C2 lists the original sources of our reconstructed dataset.

C.2 Dataset Refinement

In Table 1, aside from the TaxTopic, TaxCalc, TaxSCQ, and TaxMCQ datasets—which have definitive answers—all other tax-related datasets constructed via a human–ChatGPT collaborative approach require manual verification. Unlike other benchmarks that directly use data generated with LLM assistance, we recognize that LLM-based prompt engineering during the construction phase can introduce hallucinations (Farquhar et al., 2024), such as fabricated figures or descriptions, which may add noise or even incorrect labels to the data. This issue is particularly evident in the TaxRisk, TaxInSpect, and TaxPlan datasets. Despite multiple attempts to refine the prompts, obtaining structured outputs that meet our requirements remains challenging. Consequently, we manually review and correct all data generated with LLM involvement.

The refinement process is carried out by a dedicated annotation team of 50 students. While similar in composition to the prompt design team—which consists mainly of graduate students—this group also includes a significant number of undergraduates due to the scale of the annotation task. All team members are native Chinese speakers with diverse academic backgrounds in computer science, finance, taxation, statistics, and related fields. Computer science specialists are responsible for explaining annotation guidelines and conducting random sampling checks, while other professionals, leveraging their domain expertise in taxation, perform the actual data labeling. This collaborative approach ensures both technical rigor and domain accuracy, thereby enhancing the robustness and reliability of the dataset. Each annotator initially contributes 20 distinct data entries. After the first round, the computer science team randomly samples 5 entries from each annotator to verify compliance with the requirements. Through several rounds of feedback and guidance, all annotators eventually meet the quality standards. For annotation, we use the EasyDataSet [2] (Miao et al., 2025) platform, which is designed for constructing instruction sets for LLMs. The platform supports manual correction, allows precise Q&A labeling, and helps reduce label duplication. We provide the

[2] `https://docs.easy-dataset.com/`

annotation demonstration examples in Figure C1.

C.3 Dataset Statistics

We present detailed statistical information of our TaxPraBen dataset in Table C3. Specifically, we analyze the data length distribution. The maximum input is 2,066 characters (≈510–520 tokens), and the longest target is 814 characters (≈200–300 tokens)—values that closely align with our experimental settings. Accordingly, we cap the maximum input at 2,048 tokens. For Q&A and summarization datasets, we allow up to 300 generated tokens; for all others, the limit is 200 tokens, with prompts explicitly requesting concise outputs. These choices are well-founded: (1) virtually every sample fits within the model's context window, so no information is lost to truncation; (2) for generative tasks, the required outputs rarely exceed 300 tokens, confirming that the limits strike a balance between informativeness and brevity. In short, the statistics corroborate the soundness of our configuration.

C.4 Dataset Instance

To enhance transparency and ensure reproducibility, we present the instance formats and representative prompts utilized in TaxPraBen. For each dataset, we provide an example table that illustrates the evaluation input template—such as zero-shot or one-shot settings—along with a concrete example. Each example includes the input text, candidate options (or output constraints), and the gold label. To maintain clarity, examples are grouped by dataset and presented in separate tables, as shown in Tables C4–C17. These datasets are carefully curated by domain experts, are cost-effective, and have no commercial usage restrictions.

Knowledge Memorization (KM): contains 1 dataset. (1) **TaxRecite**: consists of "provision cue–statutory text" pairs, designed to evaluate the model's ability to reproduce tax-law provisions verbatim and maintain clause-level consistency. ***Knowledge Understanding (KU)***: contains 3 datasets. (1) **TaxSum**: includes tax-related news articles and reference summaries; the model is required to generate a summary of no more than 300 words. (2) **TaxTopic**: a single-label classification task based on news titles. The label space contains 10 categories: A. Financial literacy news; B. Tax and finance forms; C. Local regulations; D. Regulation interpretation; E. Taxable income adjustments; F. Tax assessment; G. Tax incentives; H. Tax planning; I. Tax Q&A; J. National regulations. (3) **TaxRead**: comprises article–question–reference answer triples, focusing on extractive reading comprehension grounded in the provided text. ***Knowledge Application (KA)***: contains 10 datasets. (1) **TaxCalc**: features tax computation scenarios and requires outputs with constrained key numerical fields. (2) **TaxSCQ**: single-choice questions with four candidate options; the output is restricted to a single letter in {A, B, C, D}. (3) **TaxMCQ**: multiple-choice questions with five candidate options; the output is a set of letters drawn from {A, B, C, D, E} (at least two options, reported in alphabetical order). (4) TaxQA: tax consultation question–answer samples, requiring the model to generate standardized responses. (5) **TaxBoard**: government message-board Q&A; the model must provide an answer and state the corresponding policy basis. (6) **TaxCrime**: a Criminal Law article classification task grounded in tax-related criminal fact descriptions. The label space contains 13 categories: A. Article 201 of the Criminal Law; B. Article 203 of the Criminal Law; C. Article 204 of the Criminal Law; D. Article 205 of the Criminal Law; E. Article 206 of the Criminal Law; F. Article 207 of the Criminal Law; G. Article 208 of the Criminal Law; H. Article 209 of the Criminal Law; I. Article 210 of the Criminal Law; J. Article 211 of the Criminal Law; K. Article 212 of the Criminal Law; L. Article 227 of the Criminal Law; M. Article 264 of the Criminal Law. (7) **TaxOpinion**: includes tax-related public opinion texts and reference summaries; the model generates a summary of approximately 300 words. (8) **TaxRisk**: risk management and control texts with structured extraction targets, namely "Risks Encountered" and "Corresponding Mitigation Measures". (9) **TaxInspect**: an information extraction task over tax inspection cases, requiring the model to output "Criminal Act", "Charged Offense", "Penalty Outcome", the Charged Offense is restricted to one of seven categories: Crime of holding forged invoices; Crime of illegally selling special VAT invoices; Crime of illegally purchasing special VAT invoices; Crime of purchasing forged special VAT invoices; Crime of fraudulently obtaining export tax refunds; Crime of tax evasion; Crime of falsely issuing invoices. (10) **TaxPlan**: A structured tax-planning generation task where only the core rationale is textual and all other key fields are numeric, enabling automatic evaluation.

All datasets are converted into instruction sets in JSONL format, with the following structure:

Metrics	Description
Accuracy (ACC)	Measure if the response meets the target answer provided by the query, avoiding errors and bias.
Naturalness (NAT)	Assesses whether the response description is fluent, natural, and consistent with human expression.
Informativeness (INF)	Evaluate if the response description includes enough key information and fully answers the question.

Table C18: Briefly description of the three manual metrics (ACC,NAT,INF) for prompt quality evaluation.

Task	Dataset	Prompt Consistency Analysis			
		Total	Retained	Fleiss's κ	Krippendorff's α
KM	TaxRecite	20	10	0.842	0.867
KU	TaxSum	20	12	0.824	0.856
	TaxSum	20	13	0.833	0.872
	TaxTopic	20	11	0.843	0.891
	TaxRead	20	16	0.837	0.861
KA	TaxCalc	20	13	0.856	0.872
	TaxSCQ	20	15	0.828	0.865
	TaxMCQ	20	16	0.853	0.857
	TaxQA	20	13	0.819	0.842
	TaxBoard	20	15	0.832	0.873
	TaxCrime	20	14	0.851	0.863
	TaxOpinion	20	16	0.876	0.881
	TaxRisk	20	18	0.846	0.858
	TaxInspect	20	15	0.862	0.871
	TaxPlan	20	14	0.846	0.893

Table C22: Prompt evaluation consistency analysis: average scores for Fleiss' κ and Krippendorff's α.

```
{
 id:[integer] unique sample ID
 query:[string] input question & prompt
 text:[string] input question content
 answer:[string] expected response
}
```

In addition, for all other classification datasets (including TaxSCQ, TaxTopic, and TaxCrime), the instructions contain two extra fields.

```
{
 choices:[list] a list of responses
 gold:[integer] the ideal target response
}
```

C.5 Prompt Annotation

To guide LLMs in producing outputs in a specified format for evaluation, a rigorous process of prompt selection, adaptation, and design is required. Given the specialized knowledge and complexity involved in designing prompts for the tax domain, expertise in the field is essential, and it is recognized that different LLMs exhibit significant variation in their adaptability to prompts. Therefore, a customized set of prompt samples is developed for each dataset to ensure the generalizability and comparability of prompt performance across different models. This procedure consists of two phases: annotation and evaluation. (1) **Annotation Phase**: We assemble a team of five instructors from the faculties of finance and economics, who specialize in auditing, tax administration, tax planning, tax agency, tax inspection, and public administration. For each dataset, at least two instructors collaborate—one is responsible for drafting diverse prompts, and the other reviews them to ensure the professionalism of the tax scenario descriptions and the accuracy of terminology. The team provides 20 prompt examples for each dataset. To ensure consistency in the quality of prompt annotation, we address the issue of variability in LLM adaptability caused by differences in how experts understand the data and their individual annotation styles. (2) **Evaluation Phase**: We form a professional annotation and evaluation team consisting of eight undergraduate students and two graduate students. These students have backgrounds in economics and taxation and have demonstrated strong academic performance, ensuring both expertise and accountability. Each participant evaluates 5 prompts using a questionnaire that assesses 3 metrics: Accuracy (ACC), Naturalness (NAT), and Informativeness (INF), as shown in Table C18. The evaluation options are divided into four levels with corresponding scores: "Strongly Disagree (0)", "Disagree (1)", "Agree (2)", and "Strongly Agree (3)". Examples of these manual scores are shown in Tables C19-C21. All written prompts are tested via the ChatALL [3] platform using APIs of randomly selected LLMs, like OpenAI's ChatGPT, and Alibaba's Qwen.

To ensure the reliability and quality of our evaluation annotation process, we calculate two key inter-annotator agreement scores in each evaluation round: Fleiss' Kappa (κ) and Krippendorff's Alpha (α) (Yu et al., 2025; Peng et al., 2025). Fleiss' Kappa measures the consistency among multiple raters, while Krippendorff's Alpha accounts for imbalances in category distribution. We individually assess the Kappa and Alpha scores for 3 manual metrics (ACC, NAT, and INF), averaging these scores to derive the final inter-annotator agreement score. Regular training sessions and discussions on scoring guidelines help minimize subjective bias. As summarized in Table C22, after 5 rounds of cross-evaluation, the average inter-annotator agree-

[3] https://github.com/ai-shifu/ChatALL

ment scores for each prompt set across different raters range from 0.8 to 1, demonstrating the robustness and quality of the manual assessments. To ensure consistency in prompt evaluation, we retain only those prompts with average scores exceeding 2 (AVG>2) as high-quality paired prompts. Examples of human-annotated prompts and their corresponding English translations for all datasets are presented in Table C23.

To enhance the adaptability of LLMs to prompts, each data sample is paired with a randomly selected high-quality prompt during evaluation.

D Scalable Promotion

Our TaxPraBen proposed the structured evaluation approach can be naturally extended to specialized fields like law, finance, and medicine. (1) **Legal**: For tasks like verdict summarization and sentencing prediction, our framework pairs BERTScore to validate legal reasoning soundness with accuracy metrics to verify numerical outcomes including fines or sentence lengths, enabling reliable, controllable legal AI evaluation. (2) **Finance**: For audit reports and risk analysis workflows, it validates the rationality of qualitative risk assessments while cross-checking the accuracy of transaction amounts and discrepancy ratios, standardizing performance comparisons across models for high-stakes financial automation. (3) **Medicine**: It supports granular assessment of hybrid clinical outputs, evaluating diagnostic narrative quality via semantic metrics while validating the precision of vitals, lab values and other quantitative biomedical measurements, ideal for discharge summary automation and diagnostic support tools. This structured paradigm establishes TaxPraBen as scalable evaluation framework for compliance-critical domains, enabling holistic assessment of both interpretive reasoning quality and numerical precision for LLM hybrid outputs. Table D presents adapted practice case studies for these extended domains.

E Benchmark Leaderboard

To enable transparent and standardized comparison of tax-oriented LLMs, we build the TaxPraBen open leaderboard, which supports model submission, automatic evaluation, and result visualization. The TaxPraBen open leaderboard [4] and its representative interface are shown in Figure E.

(1) Platform design. We adopt a modular architecture to continuously add tasks and metrics while keeping results comparable across versions, and provide an end-to-end pipeline from submission to display. **(2) Submission protocol.** Submissions should include necessary metadata and reproducible settings, and comply with privacy and governance requirements, avoiding sensitive data and improper behaviors. **(3) Leaderboard Evaluation.** The leaderboard covers the main tasks and metrics of TaxPraBen, reporting overall and per-task scores with filtering and sorting for quick diagnosis of strengths and weaknesses. **(4) Community Collaboration.** The platform accepts new models and benchmark extensions, and we maintain versioned tasks and rules to ensure long-term traceability and comparability.

Overall, TaxPraBen offers a scalable, open, and transparent platform for evaluating LLMs in Chinese tax practice. It enables model submissions, interactive performance comparisons, community contributions, and iterative improvements, providing a reliable reference for researchers and practitioners while bridging the gap between academic research and real-world applications in tax domain.

F Evaluation Details

F.1 Structured Alignment

To evaluate the three typical practical cases (TaxRisk, TaxInSpect, and TaxPlan), we require LLMs to produce JSON responses that strictly follow a predefined structure with specified fields. However, even with carefully designed prompts that show high consistency in human evaluations, it remains challenging to guide all LLMs to produce fully compliant outputs. Thus, we further employ ChatGPT, combined with the prompts shown in Table 3, to process the non-compliant outputs: removing redundant text and extracting all key fields. Examples of raw outputs and their standardized structured versions are provided in Table F.

F.2 Models Introduction

We evaluate 19 LLMs, organized into three groups: **(1) 7 Multilingual General LLMs**, namely ChatGPT-3.5-Turbo-1106&GPT-4o (Brown et al., 2020) (OpenAI), Mistral-v0.3 (Jiang et al., 2023) (Mistral AI), Gemma (Team et al., 2024) (Google), BayLing2 (Zhang et al., 2024) (Chinese Academy of Sciences), LLaMA3 (Dubey et al., 2024) (Meta), and Grok-3 [5] (xAI), covering both open-source and commercial models; **(2) 11 Chinese-oriented**

[4] huggingface.co/spaces/hzkkk/TaxPraBen

[5] https://grok.com/

LLMs, including Baichuan2 (Yang et al., 2023) (Baichuan AI), DeepSeek llm&R1 (Guo et al., 2025a) (DeepSeek), Qwen2.5 (Yang et al., 2024) (Alibaba), ChatGLM3&GLM4 (GLM et al., 2024) (Zhipu AI), ChineseLLaMA3 (Cui et al., 2023) (Community), YI (Young et al., 2024) (01 AI), ERNIE-3.5-8K (Baidu), Atom (Zhao et al., 2024) (AtomEcho), and InternLM2.5 (Cai et al., 2024b) (Shanghai AI Lab), pretrained on Chinese text and generally surpass multilingual LLMs in Chinese tasks; **(3) 1 Tax-related LLM**, YaYi2 (Luo et al., 2023) (Zhongke Wenge), fine-tuned with tax-related corpora to enhance its understanding of Chi-nese tax knowledge. Their respective descriptions are as follows. **ChatGPT-3.5**: Developed by OpenAI with the Turbo-1106 version, this LLM features 175B parameters and excels in conversational skills, widely used in NLP tasks. **GPT-4o**: A multimodal LLM by OpenAI, capable of text, image, and voice processing, known for fast responses and strong conversational abilities, accessible via API. **Mistral-V0.3**: Developed by Mistral AI, this 7B parameter LLM emphasizes efficient reasoning and is suitable for local deployment. **Gemma**: A 7B parameter model by Google, designed for multilingual processing in resource-constrained environments, lightweight and practical. **LLaMA3**: Developed by Meta, this LLM features 8B parameters and excels in multilingual generation for complex tasks. **Bayling2**: Based on LLAMA2 architecture, this 7B parameter LLM from the Chinese Academy of Sciences focuses on multilingual alignment, particularly for low-resource languages. **Grok3**: Developed by xAI, this model boasts 1200B parameters and is API-accessible, capable of handling complex tasks. **DeepSeek-llm**: A 7B parameter model by DeepSeek, optimized for Chinese processing and suitable for localized applications. **Baichuan2**: Developed by Baichuan AI, this 7B parameter LLM specializes in Chinese content generation with stable performance and open weight access. **Atom**: Fine-tuned from LLAMA2-7B, this 7B parameter LLM is optimized for Chinese tasks by a Chinese research team. **Qwen2.5**: Developed by Alibaba Cloud, this 7B parameter LLM supports Chinese and multilingual processing with excellent performance and weight access. **ChineseLLaMA3 (ChnLLaMA3)**: A community-developed variant fine-tuned on Llama-3-8B, featuring 8B parameters and multilingual capabilities. **ERNIE-3.5-8K**: Developed by Baidu, this model excels in Chinese language understanding with about 1,000B parameters, ideal for search and Q&A tasks, accessible via API. **ChatGLM3**: Optimized for Chinese conversation, this 6B parameter model by Zhipu AI offers weight access for chat applications. **YI**: A 6B parameter model from 01.AI, excelling in Chinese processing and code generation, suitable for development scenarios with weight access. **GLM4**: Developed by Zhipu AI, this 9B parameter model emphasizes high-precision performance in Chinese tasks for professional applications. **DeepSeek-R1**: A 7B parameter LLM by DeepSeek, known for powerful reasoning capabilities, addressing complex problem-solving needs. **InternLM2.5**: Developed by the Shanghai Artificial Intelligence Laboratory, this 7B parameter model excels in both Chinese and multilingual tasks, applicable across various scenarios. **YaYi2**: A 30B parameter LLM by Zhongke WenGe, offering weight access, designed as a general-purpose LLM with some tax-related fine-tuning for professional applications.

F.3 Metric Computation

Below is the detailed explanations of our used evaluation metrics.

Accuracy (ACC) (Goutte and Gaussier, 2005) : It measures text classification performance by the proportion of correctly predicted samples. Defined as:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (True Positive) is the count of correctly identified positive samples, TN (True Negative) is the count of correctly identified negatives, FP(False Positive) is the count of negatives misclassified as positives, and FN (False Negative) is the count of positives misclassified as negatives.

F1 Score (Goutte and Gaussier, 2005): It is the harmonic mean of Precision and Recall. It is useful for evaluating model performance in sentiment analysis, especially with class imbalance. Defined as:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

where Precision and Recall are defined as:

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}$$

Macro F1: It is the average of the F1 scores across classes, reflecting performance consistency

$$\mathrm{MCC}=\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}C_{ij}\left(\delta_{ij}-\frac{C_{i+}C_{+j}}{N^2}\right)}{\sqrt{\left(\sum_{i=1}^{K}C_{i+}\right)\left(\sum_{j=1}^{K}C_{+j}\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{C_{i+}C_{+j}}{N^2}\right)}}$$

in multi-class label analysis. Defined as:

$$\text{Macro F1}=\frac{1}{N}\sum_{i=1}^{N}\text{F1}_i$$

where N is the number of classes and F1_i is the F1 Score for class i.

Matthews Correlation Coefficient (MCC) (Matthews, 1975): It is a comprehensive classification metric for imbalanced data in sentiment analysis, as defined in the formula, where K is the total number of classes ($K = 2$ for binary classification and $K > 2$ for multi-class classification); C_{ij} is the entry in the confusion matrix at row i and column j (the number of samples whose true class is i and predicted class is j); C_{i+} is the sum of row i (the number of samples with true class i); C_{+j} is the sum of column j (the number of samples predicted as class j); and N is the total number of samples, i.e., $N=\sum_{i=1}^{K}C_{i+}=\sum_{j=1}^{K}C_{+j}$.

Exact Match (EM) Accuracy (Rajpurkar et al., 2016): It requires exact matches between predicted and true labels, defined as follows:

$$\text{EM Accuracy}=\frac{\text{Number of matches}}{\text{Total samples}}$$

BERTScore (Zhang et al., 2019): It is the harmonic mean of Precision and Recall, computed as follows:

$$\text{BERTScore}=\frac{2\cdot(\text{Precision}\cdot\text{Recall})}{\text{Precision}+\text{Recall}}$$

BARTScore (Yuan et al., 2021): It evaluates generated text quality by calculating reconstruction error with the BART model, defined as follows:

$$\text{BARTScore}=\text{BART}(Ref, Gen)$$

where Ref and Gen denote reference and generated texts, respectively. BART(Ref, Gen) is the reconstruction error calculated by the BART model.

RiskScore: It averages BERTScore for semantic matching between "risks encountered" and "matching solutions", detailed as follows:

$$\text{RiskScore}=\frac{1}{2}\sum_{i=1}^{2}\text{BERTScore}(Ref_i, Gen_i)$$

where Ref_i and Gen_i denote reference and generated texts.

InspectScore: It combines numerical matching for "offense charge" and semantic matching for "criminal conduct" and "penalty result", defined as:

$$\text{InspectScore}=\frac{1}{2}(\text{EM Accuracy}(Ref_o, Gen_o)+\frac{1}{2}\sum_{i=1}^{2}\text{BERTScore}(Ref_i, Gen_i))$$

where Ref_i and Gen_i denote the reference and prediction for two semantic matching fields, while Ref_o and Gen_o denote those for precise numerical (exact) matching.

PlanScore: It combines semantic matching for "core idea" and precise numerical matching for multiple fields (as each data entry varies, e.g., "pre-tax deductible amount of original plan", "taxable amount of original plan", "pre-tax deductible amount after planning", and "tax savings amount"), defined as follows:

$$\text{PlanScore}=\frac{1}{2}(\text{BERTScore}(Ref_c, Gen_c)+\frac{1}{p}\sum_{i=1}^{p}\text{EM Accuracy}(Ref_i, Gen_i))$$

where Ref_c and Gen_c represent the reference and prediction for semantic matching, while Ref_i and Gen_i denote those for exact matching in p fields.

F.4 Overall Metrics

Besides the TaxRisk, TaxInspection, and TaxPlan datasets, our custom-built RiskScore, InspectScore, and PlanScore can be applied as integrated evaluation metrics. For the other datasets, we assess the model's overall performance across datasets using the weighted average of all specialized metrics. Specifically, in the TaxSCQ, TaxTopic, and TaxCrime datasets, the final overall score is calculated as the average of ACC, F1, and Macro F1 metrics, defined as follows :

$$\text{S}_{\text{Overall}}=\frac{\text{ACC}+\text{F1}+\text{Macro F1}}{3}.$$

For the TaxKQA, TaxBoard, TaxRecite, TaxSum and TaxOpinion datasets, the overall score is calculated by transforming BARTScore to be greater than 0 with a weight of 10, and then averaging it with BERTScore, ensuring normalization to the range [0, 1], defined as follows :

$$\mathrm{S}_{\mathrm{Overall}} = \frac{\mathrm{BERTScore} + \frac{\mathrm{BARTScore}+10}{10}}{2}.$$

G Detailed Results

The detailed results of these models are presented in Table G. Unless otherwise specified, the issues discussed are relevant to both zero-shot and one-shot scenarios. We conduct a thorough analysis of certain notable commonalities or differences observed in each task.

TaxRecite. The TaxRecite dataset requires models to faithfully reproduce the original wording of given tax provisions, which is highly challenging for models that have not explicitly encountered the corresponding regulations. Closed-source large models (e.g., ERNIE-3.5) perform best on this task, often generating article texts that closely match the reference answers. In contrast, most open-source models, lacking explicit memorization of tax statutes, tend to produce incomplete or paraphrased outputs. Typical errors include omitting critical legal phrasing, summarizing provisions in free-form language instead of verbatim recitation, and adding unnecessary explanations. We also observe that, while some models can recover partial snippets in the zero-shot setting, providing a demonstration may interfere with recall for certain models: they rigidly imitate the example format and deviate from the exact wording, leading to performance degradation under the one-shot setting.

TaxSum. In TaxSum, models are required to read long policy news articles and generate concise summaries. This task evaluates both language understanding and summarization, as well as the ability to capture tax-related key information. Results show that closed-source models generally exhibit clear advantages in identifying salient points and maintaining factual consistency, whereas many smaller open-source models produce overly generic summaries, miss critical details (e.g., dates, numbers, or policy names), or even introduce content inconsistent with the source. A plausible explanation is that tax policy news contains abundant domain-specific terminology and complex sentence structures, making comprehensive understanding difficult for weaker models.

TaxTopic. TaxTopic is a topic classification task where models must determine the tax theme of a news article, testing keypoint extraction and topic recognition. Results indicate that open-source models can sometimes capture key terms and contextual cues to identify the relevant tax category. Under the zero-shot setting, GLM4 outperforms closed-source models; however, beyond this case, many open-source models achieve near-random performance and struggle on certain categories. This discrepancy likely stems from the fact that tax news often involves specialized terminology and implicit domain context, requiring substantial tax knowledge to classify accurately.

TaxRead. TaxRead provides relevant laws or official notices as context and asks models to extract answers accordingly, resembling an open-book question answering scenario. Most models achieve strong performance, likely because answers are explicitly present in the provided materials and can be retrieved via pattern matching, reducing reliance on external knowledge. Even smaller models can locate correct spans through surface matching. Moreover, providing demonstrations helps models better follow the expected answer format and reduce verbosity, leading to improved accuracy for most models in the one-shot setting.

TaxCalc. All models perform poorly on TaxCalc, making it one of the most challenging datasets. This task requires computing tax liabilities or tax differences given a scenario, involving numerical operations, logical reasoning, and tax-rule application. Such requirements exceed the strengths of current LLMs: even top models such as GPT-4o and ERNIE-3.5 frequently fail on complex tax planning computations, with accuracy close to zero. While a few large models occasionally produce near-correct answers for simple cases, almost no model can solve more complex instances (e.g., multi-step progressive calculations or comparative scenario reasoning).

TaxSCQ. TaxSCQ is a single-choice multiple-choice task where models must select exactly one correct option. It evaluates detailed tax knowledge and basic reasoning in an exam-like format. Closed-source models perform relatively well. Among open-source models, InternLM2.5 and GLM4 achieve competitive performance with closed-source models under the zero-shot setting. However, most open-source models remain only slightly above random guessing (25%). We also observe that some models simply restate the prompt or generate responses unrelated to the options, revealing failures in aligning with the multiple-choice answering format.

TaxMCQ. TaxMCQ is a more difficult multi-choice format where models must select all correct options. Compared to single-choice questions, this task demands more comprehensive judgment: models must identify all correct points while avoiding distractors. Results show that most models cannot correctly select all correct options simultaneously. The best-performing models are Yi, Qwen2.5, and InternLM2.5, all of which surpass closed-source models under the zero-shot setting. A typical error pattern is a combination of omissions and false positives: models often identify only the most salient correct option while missing other correct ones, and they also guess incorrect options due to uncertainty.

TaxQA. TaxQA requires models to generate free-form answers to tax questions, assessing domain knowledge application and explanation quality. Overall, model differences are not pronounced and performance is generally moderate. Among open-source models, Qwen2.5 and GLM4 perform relatively well in the zero-shot setting. We also find that knowledge-limited models may fabricate non-existent tax rules or provide incorrect tax rates, producing factually unreliable answers.

TaxBoard. TaxBoard requires models to produce official-style responses to taxpayers' inquiries, typically demanding concise answers supported by legal bases. Results confirm that models with strong instruction-following capabilities perform well on this dataset. ERNIE-3.5 can emulate the style of official tax authority replies: it tends to answer directly, respond clearly, and cite relevant statutes or regulatory references to enhance credibility. In contrast, some open-source models produce poorly structured answers lacking legal support, and weaker models may deviate from the question altogether.

TaxCrime. TaxCrime asks models to determine which tax law or regulation is violated given a case description, requiring both statute-level knowledge and case-to-law mapping ability. Results show that most models struggle to identify the correct legal provisions. ERNIE-3.5 consistently outperforms other models; this advantage may be associated with its knowledge-enhanced paradigm and broader Chinese knowledge coverage. In comparison, open-source models generally lack statute-level Chinese tax knowledge injection and therefore fail to map case details to specific provisions.

TaxOpinion. TaxOpinion requires models to summarize informal online public opinion texts, which are often colloquial and noisy (e.g., comments on tax evasion cases). Results exhibit substantial divergence across models. Some open-source models can extract main viewpoints from lengthy posts and generate coherent summaries aligned with the original intent. By contrast, weaker models frequently miss the main point: they either paraphrase fragmented sentences without distilling the central idea, or ignore the stance and sentiment expressed in the post, resulting in unfocused summaries.

TaxRisk. TaxRisk asks models to read tax risk-control articles and extract risk points along with corresponding mitigation measures, outputting them in a predefined JSON structure. This task combines information extraction and structured generation, representing a typical knowledge-application scenario. Results vary markedly across models. In addition to closed-source models, GLM4 and Qwen2.5 perform well under the zero-shot setting. Inspection of the outputs shows that models may capture only major risks while missing secondary ones, or provide overly generic mitigation strategies.

TaxInspect. TaxInspect requires models to read tax inspection cases and extract key elements in a structured format, emphasizing both fine-grained factual extraction and strict adherence to output schemas. Results show that only closed-source models and a few open-source models (e.g., GLM4, DeepSeek-R1) perform well. For most open-source models, outputs either largely copy the original case text without extracting fields, or contain severe misalignment where background information is incorrectly filled into target fields.

TaxPlan. TaxPlan is among the most demanding datasets, requiring models to propose tax planning solutions for enterprise scenarios, compute tax liabilities and savings under different strategies, and output results in a structured format. This task integrates scenario understanding, quantitative reasoning, and plan generation, making it extremely challenging. Results show that ERNIE-3.5 performs slightly worse, while GLM4 is the best among open-source models. Overall performance remains unsatisfactory across models: few can fully cover all elements in the reference answers, and accurate numerical computation is particularly rare. Qualitatively, models may propose broadly reasonable planning ideas, but they frequently fail when precise numerical calculations are required.

Task	Dataset	Definition	Collection Details
KM	*TaxRecite*	**Tax Law Recitation:** Provide the section number of the tax law, and require the model to accurately recite the original description.	Download legal and regulatory documents from the official website of the State Taxation Administration (国家税务总局), extract a portion of the original text of the provisions, and process them into a question-and-answer format.
KU	*TaxSum*	**Tax News Summarization:** Provide the news report and require the model to extract key information and generate a summary.	Select tax-related news articles from the Tax House website, retain the manually annotated summaries from the site, and compress the original text while preserving its main points to optimize the input length.
	TaxTopic	**Tax Topic Classification:** Provide the news title and require the model to determine its corresponding thematic category.	Select tax-related articles from the Tax House website, extract their titles, and record the categories they belong to on the website as labels for thematic identification.
	TaxRead	**Tax Reading Comprehension:** Provide the article along with relevant questions, and require the model to extract relevant content for answer.	Select tax-related articles from the Tax House (税屋) website, compress the original text using ChatGPT while retaining its main points to optimize input length, and utilize ChatGPT to extract questions and corresponding answers from the compressed articles.
KA	*TaxCalc*	**Tax Payment Calculation:** Provide a corporate tax case and require the model to calculate and provide the accurate amount.	Carefully select calculation problems from the exercise book for the Certified Tax Agent exam, ensuring the questions cover multiple tax types to comprehensively address and encompass the tax domain.
	TaxSCQ	**Tax Single-Choice Exam:** Provide the question along with four options, requiring the model to identify one correct answer.	Carefully select single-choice questions from the exercise book for the Certified Tax Agent Exam (注册税务师考试). To ensure broad coverage, we extract relevant exam questions from both real exam papers and mock exam papers, encompassing various knowledge areas and key exam topics.
	TaxMCQ	**Tax Multiple-Choice Exam:** Provide the question along with five options, requiring the model to identify all correct answers.	Carefully select multiple-choice questions from the exercise book for the Certified Tax Agent exam. To ensure broad coverage, we extract relevant exam questions from both real exam papers and mock exam papers, encompassing various knowledge areas and key exam topics.
	TaxQA	**Tax Knowledge Q&A:** Provide a question that asks the model to give semantically similar answers with limited length.	Download legal regulations, tax service guides, and tax incentive documents from the official website of the State Taxation Administration, as well as tax knowledge documents from Baidu Wenku (百度文库). Then, extract tax-related questions and their answers from these files.
	TaxBoard	**Tax Board Q&A:** Provide an official user inquiry question, requiring the model to give a response with limited length.	Collect real user questions and official responses from the message board of the State Taxation Administration's website, simulating authentic user consultation scenarios to provide users with more comprehensive references and guidance.
	TaxCrime	**Tax Law Identification:** Provide tax-related crime facts and ask the model to identify the violated law article from the options.	Download tax-related criminal case judgments from China Judgments Online, extract criminal behaviors and violated law articles, and use ChatGPT to label the articles for a classification task that identifies the violated laws based on the behaviors.
	TaxOpinion	**Tax Opinion Summarization:** Provide tax-related public opinion articles and require the model to generate corresponding summaries.	Retrieve public opinions on tax evasion from the StoneDT (思通数科) Opinion system, use ChatGPT to assess relevance to taxation, clean and compress the tax-related content by removing emoticons and ensuring fluent, logical sentences, and finally generate summaries with ChatGPT as dataset answers.
	TaxRisk	**Tax Risk Prevention:** Provide articles on fiscal and tax risk control and ask the model to extract the mentioned risks and their solutions.	Download articles on fiscal and tax risk control from Tax House (税屋) website, compress them using ChatGPT while retaining key information as input text, then extract the faced risks and corresponding solutions information using ChatGPT as answers.
	TaxInspect	**Tax Inspection Analysis:** Provide tax inspection cases and extract mentioned criminal behaviors, charges, and penalties.	Download tax inspection case articles from Tax House website, compress them using ChatGPT while retaining key information as input text, then extract "criminal behaviors", "crime charges", and "punishment results" information using ChatGPT as answers.
	TaxPlan	**Tax Strategy Planning:** Provide specific tax scenarios and require the model to perform tax planning to achieve tax-saving effects.	Extract tax planning cases from tax planning case books, including scenario descriptions and tax planning solutions, use ChatGPT to extract structured information as answers, replacing the original descriptive tax planning solutions with structured answers to form the dataset.

Table C1: Description of the data definitions and collection details for the three-level tax tasks in TaxPraBen.

Note: *Our annotation process on the EasyDataset platform goes beyond simple question(问题)& answer(回答) interface-based matching. We treat each question as an "annotation requirement + original document" serving as a cross-reference annotation guideline, with each answer, including the Q&A pairs generated by ChatGPT, requiring proofreading by domain experts.*

Figure C1: An example screenshot demonstrating manual adjustment of the TaxRead and TaxQA datasets.

Prompt: 根据问题需求，请提供法条的原文部分，不必对其含义做出解释，也不需要指明是第几条。(*According to the question requirements, please provide the verbatim excerpt of the relevant legal provisions. Do not explain their meaning, and do not indicate article numbers.*)

Zero-Shot

Query: {Prompt}

Text: 2007年12月6日通过的《中华人民共和国企业所得税法实施条例》中第一百一十一条的内容是什么？ (*What is the content of Article 111 of the Implementation Regulations of the Enterprise Income Tax Law of the People's Republic of China, adopted on December 6, 2007?*)

Answer: 企业所得税法第四十一条所称合理方法，包括：(一)可比非受控价格法，是指按照没有关联关系的交易各方进行相同或者类似业务往来的价格进行定价的方法；(二)再销售价格法，是指按照从关联方购进商品再销售给没有关联关系的交易方的价格，减除相同或者类似业务的销售毛利进行定价的方法；(三)成本加成法，是指按照成本加合理的费用和利润进行定价的方法；(四)交易净利润法，是指按照没有关联关系的交易各方进行相同或者类似业务往来取得的净利润水平确定利润的方法；(五)利润分割法，是指将企业与其关联方的合并利润或者亏损在各方之间采用合理标准进行分配的方法；(六)其他符合独立交易原则的方法。(*The "reasonable methods" referred to in Article 41 of the Enterprise Income Tax Law include: (1) the Comparable Uncontrolled Price (CUP) method, i.e., determining the price by reference to the prices applied in identical or similar transactions between parties without an associated relationship; (2) the Resale Price method, i.e., determining the price by reference to the price at which goods purchased from an associated party are resold to an unrelated party, minus the gross margin for identical or similar transactions;(3) the Cost Plus method, i.e., determining the price based on cost plus reasonable expenses and profit; (4) the Transactional Net Margin method (TNMM), i.e., determining profit based on the net profit level realized in identical or similar transactions between unrelated parties; (5) the Profit Split method, i.e., allocating the combined profit or loss between an enterprise and its associated parties according to reasonable allocation keys; and (6) other methods consistent with the arm's-length principle.*)

One-Shot

Query: {Prompt}

Example-Text: 2010年2月10日通过的《税务行政复议规则》中第二十五条的内容是什么？ (*What is the content of Article 25 of the Rules on Tax Administrative Reconsideration, adopted on February 10, 2010?*)

Example-Answer: 同一行政复议案件申请人超过5人的，应当推选1至5名代表参加行政复议。(*In an administrative reconsideration case where more than five applicants are involved, the applicants shall designate one to five representatives to participate in the administrative reconsideration proceedings.*)

Text: 2007年12月6日通过的《中华人民共和国企业所得税法实施条例》中第一百一十一条的内容是什么？ (*What is the content of Article 111 of the Implementation Regulations of the Enterprise Income Tax Law of the People's Republic of China, adopted on December 6, 2007?*)

Answer: 企业所得税法第四十一条所称合理方法，包括：(一)可比非受控价格法，是指按照没有关联关系的交易各方进行相同或者类似业务往来的价格进行定价的方法；(二)再销售价格法，是指按照从关联方购进商品再销售给没有关联关系的交易方的价格，减除相同或者类似业务的销售毛利进行定价的方法；(三)成本加成法，是指按照成本加合理的费用和利润进行定价的方法；(四)交易净利润法，是指按照没有关联关系的交易各方进行相同或者类似业务往来取得的净利润水平确定利润的方法；(五)利润分割法，是指将企业与其关联方的合并利润或者亏损在各方之间采用合理标准进行分配的方法；(六)其他符合独立交易原则的方法。(*The "reasonable methods" referred to in Article 41 of the Enterprise Income Tax Law include: (1) the Comparable Uncontrolled Price (CUP) method, i.e., determining the price by reference to the prices applied in identical or similar transactions between parties without an associated relationship; (2) the Resale Price method, i.e., determining the price by reference to the price at which goods purchased from an associated party are resold to an unrelated party, minus the gross margin for identical or similar transactions;(3) the Cost Plus method, i.e., determining the price based on cost plus reasonable expenses and profit; (4) the Transactional Net Margin method (TNMM), i.e., determining profit based on the net profit level realized in identical or similar transactions between unrelated parties; (5) the Profit Split method, i.e., allocating the combined profit or loss between an enterprise and its associated parties according to reasonable allocation keys; and (6) other methods consistent with the arm's-length principle.*)

Table C4: An instance format with paired prompts for the TaxRecite dataset in the KM task.

Prompt: 请根据以下文本编写一个简短的概要，字数不超过300字，重点在于提炼核心内容。(*Please write a concise summary of the following text in no more than 300 words, focusing on extracting the key points.*)

Zero-Shot

Query: {Prompt}

Text: 为配合个人所得税专项附加扣除工作，依据相关法律法规，三部门联合发布通知，明确个人所得税住房贷款利息专项附加扣除相关信息归集事项。信息归集范围为1989年1月1日（含）后发放的商业性个人住房贷款，不包括个人商用房贷款。数据项涵盖借款人姓名、证件号码、贷款银行、合同编号、是否为首套住房贷款等11项内容。住房公积金贷款由主管部门负责归集。各商业银行于2019年3月1日前向征信系统报送信息。....。(*To facilitate the work related to the special additional deductions for individual income tax, the three departments have jointly issued a notice clarifying the information collection matters related to the special additional deduction for housing loan interest. The scope of information collection includes commercial personal housing loans issued after January 1, 1989 (inclusive), excluding personal commercial property loans. The data items cover 11 contents, including the borrower's name, identification number, loan bank, contract number, and whether it is a first home loan, among others. Housing provident fund loans are to be collected by the relevant management department. All commercial banks must report this information to the credit reporting system by March 1, 2019. ... .*)

Answer: 信息归集工作时间紧、任务重、要求高，各商业银行要高度重视，充分认识贯彻施行《个人所得税专项附加扣除暂行办法》的重要性和紧迫性，切实加强组织领导，务必在规定时间内做好各项工作。(*Given the tight timelines, heavy workload, and stringent requirements associated with information consolidation, all commercial banks should attach great importance to this effort. They should fully recognize the significance and urgency of implementing the* Interim Measures for Special Additional Deductions of Individual Income Tax*, strengthen organizational leadership, and ensure that all related tasks are completed within the prescribed time frame.*)

One-Shot

Query: {Prompt}

Example-Text: 国家税务总局厦门市税务局通知，自即日起，通过12366纳税服务平台受理网上咨询以提升服务品质。纳税人、缴费人可点击网站"我要咨询"，使用智能咨询、网上留言、在线咨询和众包互助服务。原有智能咨询服务平台继续用于社保自助咨询；历史留言可在"办税服务-历史咨询"查看。智能咨询页面功能如下：系统根据IP地址默认显示地区，用户可修改。1. 智能咨询：录入问题自助查询。2. 在线咨询：工作时间可与客服对话。3. 网上留言：新用户需注册登录后提交留言。4. 众包互助：登录后可查看或帮助回复问题。感谢纳税人对12366纳税咨询服务工作的支持。(*The Xiamen Municipal Tax Service of the State Taxation Administration (STA) announces that, effective immediately, online inquiries will be handled through the 12366 Tax Service Platform to enhance service quality. Taxpayers and contribution payers may click "I Want to Consult" on the platform website and use the Intelligent Consultation, Online Message, Live Chat, and Crowdsourced Mutual-Help services. The original Intelligent Consultation platform will remain available for self-service inquiries related to social insurance, and historical messages can be accessed via "Tax Service"–"Inquiry History". The Intelligent Consultation page provides the following functions: the system automatically displays the user's region based on the IP address, which can be modified manually. (1) Intelligent Consultation: users enter questions for self-service retrieval; (2) Live Chat: during working hours, users may communicate with customer service staff in real time; (3) Online Message: new users must register and log in before submitting messages; and (4) Crowd sourced Mutual Help: after logging in, users may browse and assist in answering questions. The Xiamen Municipal Tax Service of the STA appreciates taxpayers' continued support for the 12366 tax consultation services.*)

Example-Answer: 为更好地服务纳税人和缴费人，提升网上咨询服务品质，我局网站自即日起，通过国家税务总局12366纳税服务平台受理网上咨询。点击首页"我要咨询"，即可使用智能咨询、网上留言、在线咨询和众包互助等咨询服务。(*To better serve taxpayers and fee payers and to enhance the quality of online consultation services, our bureau will, effective immediately, accept online inquiries via the State Taxation Administration's 12366 Tax Service Platform. Users may click "I want to consult" on the homepage to access consultation services including intelligent consultation, online messaging, live online consultation, and crowdsourced mutual assistance.*)

Text: 为配合个人所得税专项附加扣除工作，依据相关法律法规，三部门联合发布通知，明确个人所得税住房贷款利息专项附加扣除相关信息归集事项。信息归集范围为1989年1月1日（含）后发放的商业性个人住房贷款，不包括个人商用房贷款。数据项涵盖借款人姓名、证件号码、贷款银行、合同编号、是否为首套住房贷款等11项内容。住房公积金贷款由主管部门负责归集。各商业银行于2019年3月1日前向征信系统报送信息。....。(*To facilitate the work related to the special additional deductions for individual income tax, the three departments have jointly issued a notice clarifying the information collection matters related to the special additional deduction for housing loan interest. The scope of information collection includes commercial personal housing loans issued after January 1, 1989 (inclusive), excluding personal commercial property loans. The data items cover 11 contents, including the borrower's name, identification number, loan bank, contract number, and whether it is a first home loan, among others. Housing provident fund loans are to be collected by the relevant management department. All commercial banks must report this information to the credit reporting system by March 1, 2019. ... ..*)

Answer: 信息归集工作时间紧、任务重、要求高，各商业银行要高度重视，充分认识贯彻施行《个人所得税专项附加扣除暂行办法》的重要性和紧迫性，切实加强组织领导，务必在规定时间内做好各项工作。(*Given the tight timelines, heavy workload, and stringent requirements associated with information consolidation, all commercial banks should attach great importance to this effort. They should fully recognize the significance and urgency of implementing the* Interim Measures for Special Additional Deductions of Individual Income Tax*, strengthen organizational leadership, and ensure that all related tasks are completed within the prescribed time frame.*)

Table C5: An instance format with paired prompts for the TaxSum dataset in the KU task.

Prompt: 阅读标题以了解文章主题，并在提供的选项中选择最适合表达这一主题的一个，选项有：A.财商资讯 B.财税表单 C.地方法规 D.法规解读 E.纳税调整 F.纳税评估 G.税收优惠 H.税收筹划 I.税务问答 J.中央法规。(*Read the title to identify the main topic of the article, and select the single option that best matches this topic from the following candidates: A. Financial literacy news; B. Tax and finance forms; C. Local regulations; D. Regulation interpretation; E. Taxable income adjustments; F. Tax assessment; G. Tax incentives; H. Tax planning; I. Tax Q&A; J. National regulations.*)

Zero-Shot

Query: {Prompt}
Text: 浙江省国家税务局公告2014年第10号 浙江省国家税务局关于企业资产损失税前扣除申报有关问题的公告 (*Announcement No. 10 (2014) of the Zhejiang Provincial Office of the State Administration of Taxation: On Issues Concerning the Pre-tax Deduction Declaration for Enterprise Asset Losses.*)

Answer: C

Choices: ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J"]

Gold: 2

One-Shot

Query: {Prompt}
Example-Text: 税务总局解答热点财税问题 (*Hotspot Q&A on Tax and Fiscal Issues Issued by China's State Taxation Administration*)
Example-Answer: I
Text: 浙江省国家税务局公告2014年第10号 浙江省国家税务局关于企业资产损失税前扣除申报有关问题的公告 (*Announcement No. 10 (2014) of the Zhejiang Provincial Office of the State Administration of Taxation: On Issues Concerning the Pre-tax Deduction Declaration for Enterprise Asset Losses.*)

Answer: C

Choices: ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J"]

Gold: 2

Table C6: An instance format with paired prompts for the TaxTopic dataset in the KU task.

Prompt: 关于下面提出的问题，请仔细考虑四个可能的答案，并选出你认为正确的那个，仅需指出"A"、"B"、"C"或"D"，不需要解释。(*Please carefully consider the four candidate answers to the question below and select the one you deem correct. Return only the letter "A", "B", "C", or "D" without providing any explanation.*)

Zero-Shot

Query: {Prompt}
Text: 关于退还关税，下列说法正确的是()。A.纳税人自缴纳税款之日起3年内，可以申请退还关税 B.海关发现实际征收税款多于应收税款，可以不用通知纳税义务人办理退还手续 C.已征出口关税的货物，因故未装运出口申请退关的，可以申请退还关税 D.海关应当自受理退税申请之日起60日内查实并通知纳税义务人办理退还手续 (*Regarding the refund of customs duties, which of the following statements is correct()? A.Within three years from the date on which the taxpayer pays the duty, the taxpayer may apply for a refund of customs duties. B.If the Customs authorities discover that the duties actually collected exceed the duties payable, they may refund the overpaid amount without notifying the taxpayer or tax obligor to complete the refund procedures. C.For goods on which export duties have already been levied, if the goods are not shipped for certain reasons and a duty refund is requested, an application may be filed for a refund of the customs duties. D.The Customs authorities shall, within 60 days from the date of accepting the refund application, verify the relevant facts and notify the taxpayer or tax obligor to complete the refund procedures.*)

Answer: C

Choices: ["A", "B", "C", "D"]

Gold: 2

One-Shot

Query: {Prompt}
Example-Text: 下列各课税要素，与纳税期限的选择密切相关的是（）。A.计税依据 B.纳税环节 C.课税对象的性质 D.纳税人 (*Which of the following tax elements is most closely associated with the selection of the tax payment period ( )? A.Tax base B.Taxable stage C.Nature of the taxable object D.Taxpayer*)
Example-Answer: C
Text: 关于退还关税，下列说法正确的是()。A.纳税人自缴纳税款之日起3年内，可以申请退还关税 B.海关发现实际征收税款多于应收税款，可以不用通知纳税义务人办理退还手续 C.已征出口关税的货物，因故未装运出口申请退关的，可以申请退还关税 D.海关应当自受理退税申请之日起60日内查实并通知纳税义务人办理退还手续 (*Regarding the refund of customs duties, which of the following statements is correct()? A.Within three years from the date on which the taxpayer pays the duty, the taxpayer may apply for a refund of customs duties. B.If the Customs authorities discover that the duties actually collected exceed the duties payable, they may refund the overpaid amount without notifying the taxpayer or tax obligor to complete the refund procedures. C.For goods on which export duties have already been levied, if the goods are not shipped for certain reasons and a duty refund is requested, an application may be filed for a refund of the customs duties. D.The Customs authorities shall, within 60 days from the date of accepting the refund application, verify the relevant facts and notify the taxpayer or tax obligor to complete the refund procedures.*)

Answer: C

Choices: ["A", "B", "C", "D"]

Gold: 2

Table C7: An instance format with paired prompts for the TaxSCQ dataset in the KA task.

Prompt: 在阅读完下面的文本之后，请根据其内容回答问题，确保回复是从文中摘录且言简意赅，不要重复问题，答案中不要出现句号。(*After reading the following text, please answer the question based on its content. Ensure that your response is a concise excerpt from the text, without repeating the question, and do not include periods in your answer.*)
Zero-Shot
Query: {Prompt} **Text:** 为综合反映与科学评价资产评估机构整体实力和专业胜任能力，推动资产评估行业信用体系建设，提升行业公信力和影响力，促进高质量发展，中国资产评估协会修订并发布了《资产评估机构综合评价办法》。该办法旨在通过系统化的评估体系，全面衡量资产评估机构的业务水平、服务质量、风险管理等多方面的能力，从而引导行业健康发展。新办法已经第六届常务理事会第三次会议审议通过，于2025年2月18日正式发布。问题：《资产评估机构综合评价办法》何时正式发布？ (*To comprehensively reflect and rigorously assess the overall strength and professional competence of asset appraisal firms, promote the development of an industry-wide credit system, enhance credibility and influence, and foster high-quality growth, the China Appraisal Society (CAS) revised and released the Measures for the Comprehensive Evaluation of Asset Appraisal Firms. The revised measures aim to establish a systematic evaluation framework that holistically assesses firms' capabilities across multiple dimensions, including operational performance, service quality, and risk management, thereby guiding the healthy development of the industry. The new measures were reviewed and approved at the Third Meeting of the Third Executive Council of the Sixth Standing Council and were formally released on February 18, 2025.* ***Question:*** *When was the Measures for the Comprehensive Evaluation of Asset Appraisal Firms formally released?*)
Answer: 2025年2月18日 (*February 18, 2025*)
One-Shot
Query: {Prompt} **Example-Text:** 国家税务总局平潭综合实验区税务局通知，已委托中国邮政集团公司福建省平潭县分公司为符合条件的自然人代开增值税普通发票。为便于工作开展，自即日起启用"国家税务总局平潭综合实验区税务局委托代开发票专用章"印章。附件为该印章印模。通知时间为2019年8月21日。问题：国家税务总局平潭综合实验区税务局委托哪家公司为符合条件的自然人代开增值税普通发票？ (*The Pingtan Comprehensive Experimental Zone Taxation Bureau of the State Taxation Administration (STA) issued a notice stating that it has entrusted the Pingtan County Branch of China Post Group Co., Ltd. (Fujian Province) to issue VAT ordinary invoices on behalf of eligible natural persons. To facilitate implementation, the bureau has, effective immediately, put into use the seal entitled "Special Seal for Entrusted Invoice Issuance of the Pingtan Comprehensive Experimental Zone Taxation Bureau of the State Taxation Administration." An impression specimen of the seal is provided in the appendix. The notice was issued on August 21, 2019.* ***Question:*** *Which company was entrusted by the Pingtan Comprehensive Experimental Zone Taxation Bureau of the STA to issue VAT ordinary invoices on behalf of eligible natural persons?*) **Example-Answer:** 中国邮政集团公司福建省平潭县分公司 (*Pingtan County Branch, Fujian Province, China Post Group Corporation.*) **Text:** 为综合反映与科学评价资产评估机构整体实力和专业胜任能力，推动资产评估行业信用体系建设，提升行业公信力和影响力，促进高质量发展，中国资产评估协会修订并发布了《资产评估机构综合评价办法》。该办法旨在通过系统化的评估体系，全面衡量资产评估机构的业务水平、服务质量、风险管理等多方面的能力，从而引导行业健康发展。新办法已经第六届常务理事会第三次会议审议通过，于2025年2月18日正式发布。问题：《资产评估机构综合评价办法》何时正式发布？ (*To comprehensively reflect and rigorously assess the overall strength and professional competence of asset appraisal firms, promote the development of an industry-wide credit system, enhance credibility and influence, and foster high-quality growth, the China Appraisal Society (CAS) revised and released the Measures for the Comprehensive Evaluation of Asset Appraisal Firms. The revised measures aim to establish a systematic evaluation framework that holistically assesses firms' capabilities across multiple dimensions, including operational performance, service quality, and risk management, thereby guiding the healthy development of the industry. The new measures were reviewed and approved at the Third Meeting of the Third Executive Council of the Sixth Standing Council and were formally released on February 18, 2025.* ***Question:*** *When was the Measures for the Comprehensive Evaluation of Asset Appraisal Firms formally released?*)
Answer: 2025年2月18日 (*February 18, 2025*)

Table C8: An instance format with paired prompts for the TaxRead dataset in the KA task.

Prompt: 针对题目，请直接给出xx处的答案，并将答案带入到问题中复述一遍，例如："业务（1）应缴纳消费税为xx万元"，保证输出字符在30个字符以内。(*For the question, please directly provide the answer for* xx*, and restate it in the question, such as: 'For transaction (1), the consumption tax payable is* xx *yuan,' ensuring the output is within 30 characters.*)

Zero-Shot

Query: {Prompt}

Text: 位于某市区国家重点扶持的高新技术企业为增值税一般纳税人（制造业企业），2023年销售产品取得不含税收入6500万元，另外取得投资收益320万元，全年发生产品销售成本和相关费用共计5300万元。缴纳的税金及附加339万元，发生的营业外支出420万元，12月末企业自行计算的全年会计利润总额761万元，预缴企业所得税96万元。2024年1月经聘请的税务师事务所审核，发现以下问题：（1）8月中旬以预收款方式销售一批产品，预收含税价款234万元，并收存银行。12月下旬将该批产品发出，但未将预收款转作收入。（不考虑相关成本）... 问题为：新产品研发费应调整的应纳税所得额为xx万元？ (*A nationally supported key high-tech enterprise located in the urban district of a certain city is a general VAT taxpayer (manufacturing enterprise). In 2023, it generated RMB 65.00 million in product sales revenue (VAT-exclusive) and additionally earned investment income of RMB 3.20 million. For the year, the total cost of goods sold and related expenses amounted to RMB 53.00 million. Taxes and surcharges paid were RMB 3.39 million, and non-operating expenses were RMB 4.20 million. As of the end of December, the enterprise calculated an annual accounting profit before tax of RMB 7.61 million, and it had prepaid corporate income tax (CIT) of RMB 0.96 million. In January 2024, after an audit by a hired tax agent firm, the following issues were identified:(1) In mid-August, the enterprise sold a batch of products under an advance-receipt arrangement, receiving an advance payment of RMB 2.34 million (VAT-inclusive), which was deposited in the bank. In late December, the batch of products was delivered, but the advance receipt was not reclassified as revenue. (Relevant costs are ignored.) ... **Question:** The adjustment to taxable income attributable to the new product R&D expenses should be* xx *10,000 RMB.*)

Answer: -72.75

One-Shot

Query: {Prompt}

Example-Text: 我国某居民企业在甲国设立一家分公司，在乙国设立一家持股80%的子公司，2023年该企业计算的境内利润总额4000万元。相关涉税资料如下：（1）甲国分公司按我国税法确认的销售收入300万元，销售成本500万元。（2）收到乙国子公司投资收益1900万元，子公司已在乙国缴纳企业所得税1000万元，子公司当年税后利润全部按持股比例分配，乙国预提所得税税率5%。（其他相关资料：该居民企业适用25%的企业所得税税率，无纳税调整金额，境外已纳税额选择"分国不分项"抵免方式）问题为：2023年该居民企业来源于子公司投资收益的境外实际缴纳税额是多少万元？ (*A Chinese resident enterprise establishes a branch in Country A and an 80% owned subsidiary in Country B. In 2023, the enterprise reports a domestic total profit of RMB 40 million. The relevant tax information is as follows: (1) The branch in Country A has sales revenue of RMB 3 million and sales cost of RMB 5 million as recognized under Chinese tax law. (2) The enterprise receives RMB 19 million in investment income (dividends) from the subsidiary in Country B. The subsidiary has paid RMB 10 million in corporate income tax in Country B. The subsidiary distributes all of its after-tax profit for the year in proportion to shareholding, and the withholding income tax rate in Country B is 5%. (Other information: the resident enterprise is subject to a 25% corporate income tax rate in China, with no tax adjustments, and adopts the "per-country, no-itemization" foreign tax credit method.) **Question:** In 2023, what is the amount of foreign taxes actually paid attributable to the investment income from the subsidiary (in RMB million)?*)

Example-Answer: 2023年该居民企业来源于子公司投资收益的境外实际缴纳税额是900.00万元 (*In 2023, the resident enterprise incurred foreign income taxes actually paid amounting to RMB 9.00 million on investment income received from its subsidiaries.*)

Text: 位于某市区国家重点扶持的高新技术企业为增值税一般纳税人（制造业企业），2023年销售产品取得不含税收入6500万元，另外取得投资收益320万元，全年发生产品销售成本和相关费用共计5300万元。缴纳的税金及附加339万元，发生的营业外支出420万元，12月末企业自行计算的全年会计利润总额761万元，预缴企业所得税96万元。2024年1月经聘请的税务师事务所审核，发现以下问题：（1）8月中旬以预收款方式销售一批产品，预收含税价款234万元，并收存银行。12月下旬将该批产品发出，但未将预收款转作收入。（不考虑相关成本）...问题为：新产品研发费应调整的应纳税所得额为xx万元？ (*A nationally supported key high-tech enterprise located in the urban district of a certain city is a general VAT taxpayer (manufacturing enterprise). In 2023, it generated RMB 65.00 million in product sales revenue (VAT-exclusive) and additionally earned investment income of RMB 3.20 million. For the year, the total cost of goods sold and related expenses amounted to RMB 53.00 million. Taxes and surcharges paid were RMB 3.39 million, and non-operating expenses were RMB 4.20 million. As of the end of December, the enterprise calculated an annual accounting profit before tax of RMB 7.61 million, and it had prepaid corporate income tax (CIT) of RMB 0.96 million. In January 2024, after an audit by a hired tax agent firm, the following issues were identified:(1) In mid-August, the enterprise sold a batch of products under an advance-receipt arrangement, receiving an advance payment of RMB 2.34 million (VAT-inclusive), which was deposited in the bank. In late December, the batch of products was delivered, but the advance receipt was not reclassified as revenue. (Relevant costs are ignored.)...**Question:** The adjustment to taxable income attributable to the new product R&D expenses should be* xx *10,000 RMB.*)

Answer: -72.75

Table C9: An instance format with paired prompts for the TaxCalc dataset in the KA task.

Prompt: 针对以下问题，请分析五个选项，并选择最适合的两个或更多，然后按字母顺序提供"A"、"B"、"C"、"D"或"E"作为你的答案，输出格式为字母组合，字母之间不要有任何其他字符，且必须是顺序的，例如："AB"或"ABC"。(*For the following question, analyze the five options and select the two or more most appropriate ones. Then provide your answer as a sequence of letters chosen from "A", "B", "C", "D", and "E", listed in alphabetical order (e.g., "AB", or "ABC").*)

Zero-Shot

Query: {Prompt}

Text: 关于不动产租赁服务的增值税处理，正确的有()。A.个体工商户异地出租不动产，在不动产所在地预缴的增值税款，可在当期增值税税款中抵减 B.以经营租赁方式将土地出租给他人使用，按不动产经营租赁服务缴纳增值税 C.一般纳税人出租其2016年5月1日前取得的不动产可选择简易计税方法计税 D.其他个人异地出租不动产，向不动产所在地预缴税款，向居住所在地申报纳税 E.纳税人向其他个人出租不动产，可以开具增值税专用发票 (*Regarding the VAT treatment of real estate leasing services, which of the following statements are correct? A. Individual businesses leasing real estate in a different location can offset the prepaid VAT at the property's location against their current VAT payable. B. Leasing land to others under an operating lease is subject to VAT as real estate leasing services. C. General taxpayers renting out real estate acquired before May 1, 2016, can choose the simplified tax calculation method. D. Individuals leasing real estate in a different location should prepay VAT at the property's location and declare taxes at their place of residence. E. Taxpayers leasing real estate to individuals can issue a VAT special invoice.*)

Answer: ABC

One-Shot

Query: {Prompt}

Example-Text: 下列业务产生的收入中，可以作为计算业务招待费、广告费和业务宣传费税前扣除限额基数的有（）。A.让渡无形资产使用权 B.让渡固定资产所有权 C.从事债权投资 D.接受捐赠 E.将资产转移至境外分支机构用于销售 (*Among the following types of income, which can be included in the base used to compute the pre-tax deduction limits for business entertainment expenses, advertising expenses, and business promotion expenses? A.Granting the right to use intangible assets B.Transferring ownership of fixed assets C.Engaging in debt investment D.Receiving donations E.Transferring assets to an overseas branch for sale*)

Example-Answer: AE

Text: 关于不动产租赁服务的增值税处理，正确的有()。A.个体工商户异地出租不动产，在不动产所在地预缴的增值税款，可在当期增值税税款中抵减 B.以经营租赁方式将土地出租给他人使用，按不动产经营租赁服务缴纳增值税 C.一般纳税人出租其2016年5月1日前取得的不动产可选择简易计税方法计税 D.其他个人异地出租不动产，向不动产所在地预缴税款，向居住所在地申报纳税 E.纳税人向其他个人出租不动产，可以开具增值税专用发票 (*Regarding the VAT treatment of real estate leasing services, which of the following statements are correct? A. Individual businesses leasing real estate in a different location can offset the prepaid VAT at the property's location against their current VAT payable. B. Leasing land to others under an operating lease is subject to VAT as real estate leasing services. C. General taxpayers renting out real estate acquired before May 1, 2016, can choose the simplified tax calculation method. D. Individuals leasing real estate in a different location should prepay VAT at the property's location and declare taxes at their place of residence. E. Taxpayers leasing real estate to individuals can issue a VAT special invoice.*)

Answer: ABC

Table C10: An instance format with paired prompts for the TaxMCQ dataset in the KA task.

Prompt: 对于下面提出的税务问题，请准备简要的答案，注意每项答复应保持在300字之内。(*For the tax-related question below, please provide concise answers. Each response should be limited to no more than 300 words.*)

Zero-Shot

Query: {Prompt}

Text: 转让债权如何确定其应予减除的财产原值和合理费用？ (*How should the original value of the property to be deducted and the reasonable expenses be determined when transferring a debt?*)

Answer: 采用"加权平均法"确定其应予减除的财产原值和合理费用。即以纳税人购进的同一种类债券买入价和买进过程中缴纳的税费总和，除以纳税人购进的该种类债券数量之和，乘以纳税人卖出的该种类债券数量，再加上卖出的该种类债券过程中缴纳的税费。(*he "weighted average method" is used to determine the deductible original value and reasonable expenses of the asset. Specifically, the total purchase cost and related taxes and fees for the same type of bonds are divided by the total quantity purchased to obtain the average cost per bond. This average is then multiplied by the quantity sold, and any taxes and fees paid during the sale are added.*)

One-Shot

Query: {Prompt}

Example-Text: 人均耕地超过二亩但不超过三亩的地区的耕地占用税每平方米是多少？ (*What is the cultivated land use tax (per square meter) in regions where the per-capita cultivated land area exceeds 2 mu but does not exceed 3 mu?*)

Example-Answer: 六元至三十元。(*from six to thirty yuan.*)

Text: 转让债权如何确定其应予减除的财产原值和合理费用？ (*How should the original value of the property to be deducted and the reasonable expenses be determined when transferring a debt?*)

Answer: 采用"加权平均法"确定其应予减除的财产原值和合理费用。即以纳税人购进的同一种类债券买入价和买进过程中缴纳的税费总和，除以纳税人购进的该种类债券数量之和，乘以纳税人卖出的该种类债券数量，再加上卖出的该种类债券过程中缴纳的税费。(*The "weighted average method" is used to determine the deductible original value and reasonable expenses of the asset. Specifically, the total purchase cost and related taxes and fees for the same type of bonds are divided by the total quantity purchased to obtain the average cost per bond. This average is then multiplied by the quantity sold, and any taxes and fees paid during the sale are added.*)

Table C11: An instance format with paired prompts for the TaxQA dataset in the KA task.

Prompt: 对于由群众提出的税务问题，请提供确切的解答，并指出相关的法律依据，确保每个回答都在300字以内。(*For tax-related inquiries raised by the general public, please provide precise answers and cite the relevant legal basis. Ensure that each response is no more than 300 Chinese characters in length.*)

Zero-Shot

Query: {Prompt}

Text: 标题：文化事业建设费今年还有减半优惠吗？ 内容：文化事业建设费今年还有减半优惠吗？ (*Title: Is there still a 50% reduction policy for the Cultural Undertaking Construction Fee this year? Content: Is there still a 50% reduction policy for the Cultural Undertaking Construction Fee this year?*)

Answer: 根据《财政部关于延续实施文化事业建设费优惠政策的通知》(财税〔2025〕7号)规定，自2025年1月1日至2027年12月31日，对归属中央收入的文化事业建设费，将按应缴费额的50%减征；对归属地方收入的文化事业建设费，各省可以结合当地经济发展和文化事业需求，在50%的幅度内减征。若已征收的费用符合减免条件，可予以退还或抵减以后应缴费用。(*According to the* Notice of the Ministry of Finance on Extending Preferential Policies for the Cultural Undertakings Construction Fee *(Caishui [2025] No. 7), from January 1, 2025 to December 31, 2027, the cultural undertakings construction fee attributable to central government revenue shall be reduced by 50% of the payable amount. For the portion attributable to local government revenue, each province may, in light of local economic development and cultural-sector needs, grant a reduction of up to 50%. If fees already collected meet the eligibility criteria for such reductions or exemptions, they may be refunded or credited against future payable amounts.*)

One-Shot

Query: {Prompt}

Example-Text: 标题：企业录用脱贫人口可以享受什么优惠？ 内容：企业录用脱贫人口可以享受什么优惠？ (*Title: What preferential tax policies are available to enterprises that employ individuals lifted out of poverty? Content: What preferential tax policies are available to enterprises that employ individuals lifted out of poverty?*)

Example-Answer: 根据《关于进一步支持重点群体创业就业有关税收政策的公告》(苏财税〔2023〕23号)第二条的规定，自2023年1月1日至2027年12月31日，企业招用脱贫人口，以及在人力资源社会保障部门公共就业服务机构登记失业半年以上且持《就业创业证》或《就业失业登记证》的人员，与其签订1年以上期限劳动合同并依法缴纳社会保险费的，自签订劳动合同并缴纳社会保险当月起，在3年内按实际招用人数，每人每年可享受7800元的定额扣减，适用于增值税、城市维护建设税、教育费附加、地方教育附加和企业所得税。(*According to Article 2 of the* Announcement on Further Supporting Tax Policies for Entrepreneurship and Employment of Key Groups *(Su Cai Shui [2023] No. 23), from January 1, 2023 to December 31, 2027, enterprises that hire (i) individuals from alleviated-poverty households or (ii) persons who have been registered as unemployed for more than six months with public employment service agencies of the human resources and social security authorities and who hold an* Employment and Entrepreneurship Certificate *or an* Employment and Unemployment Registration Certificate, *and that sign labor contracts with a term of at least one year and pay social insurance premiums in accordance with the law, are eligible, starting from the month in which the labor contract is executed and the social insurance premiums are paid, to claim a fixed-amount tax credit of RMB 7,800 per eligible employee per year, based on the actual number of hires, for up to three years. This credit may be used to offset value-added tax, urban maintenance and construction tax, education surcharge, local education surcharge, and corporate income tax.*)

Text: 标题：文化事业建设费今年还有减半优惠吗？ 内容：文化事业建设费今年还有减半优惠吗？ (*Title: Is there still a 50% reduction policy for the Cultural Undertaking Construction Fee this year? Content: Is there still a 50% reduction policy for the Cultural Undertaking Construction Fee this year?*)

Answer: 根据《财政部关于延续实施文化事业建设费优惠政策的通知》(财税〔2025〕7号)规定，自2025年1月1日至2027年12月31日，对归属中央收入的文化事业建设费，将按应缴费额的50%减征；对归属地方收入的文化事业建设费，各省可以结合当地经济发展和文化事业需求，在50%的幅度内减征。若已征收的费用符合减免条件，可予以退还或抵减以后应缴费用。(*According to the* Notice of the Ministry of Finance on Extending Preferential Policies for the Cultural Undertakings Construction Fee *(Caishui [2025] No. 7), from January 1, 2025 to December 31, 2027, the cultural undertakings construction fee attributable to central government revenue shall be reduced by 50% of the payable amount. For the portion attributable to local government revenue, each province may, in light of local economic development and cultural-sector needs, grant a reduction of up to 50%. If fees already collected meet the eligibility criteria for such reductions or exemptions, they may be refunded or credited against future payable amounts.*)

Table C12: An instance format with paired prompts for the TaxBoard dataset in the KA task.

Prompt: 阅读下面的犯罪事实，请判断其主要违反了以下哪一法律条款：A. 刑法第二百零一条 B. 刑法第二百零三条 C. 刑法第二百零四条 D. 刑法第二百零五条 E. 刑法第二百零六条 F. 刑法第二百零七条 G. 刑法第二百零八条 H. 刑法第二百零九条 I. 刑法第二百一十条 J. 刑法第二百一十一条 K. 刑法第二百一十二条 L. 刑法第二百二十七条 M. 刑法第二百六十四条。只需输出对应字母，不需要解释原因。(*Read the following statement of criminal facts and determine which of the following legal provisions is primarily violated: A. Article 201 of the Criminal Law; B. Article 203 of the Criminal Law; C. Article 204 of the Criminal Law; D. Article 205 of the Criminal Law; E. Article 206 of the Criminal Law; F. Article 207 of the Criminal Law; G. Article 208 of the Criminal Law; H. Article 209 of the Criminal Law; I. Article 210 of the Criminal Law; J. Article 211 of the Criminal Law; K. Article 212 of the Criminal Law; L. Article 227 of the Criminal Law; M. Article 264 of the Criminal Law. Output only the corresponding letter; no explanation is required.*)

Zero-Shot

Query: {Prompt}

Text: 被告人叶俊在2017年4月11日10时许，在北京市丰台区七里庄"速八酒店"门前向樊某出售北京市增值税普通发票，现场被北京市公安局丰台分局东铁匠营派出所民警当场查获，起获增值税普通发票19张，票面金额共计人民币1875921.23元，且经鉴定均为真发票。(*The defendant, Ye Jun, was apprehended at approximately 10:00 on April 11, 2017, while selling Beijing value-added tax ordinary invoices to an individual surnamed Fan in front of the Super 8 Hotel in Qilizhuang, Fengtai District, Beijing. Officers from the Dongtiejianying Police Station of the Fengtai Branch of the Beijing Municipal Public Security Bureau seized 19 VAT ordinary invoices with a total face value of RMB 1,875,921.23; an appraisal confirmed that all of the invoices were genuine.*)

Answer: H

Choices: ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L", "M"]

Gold: 7

One-Shot

Query: {Prompt}

Example-Text: 2015年1月至2016年7月，被告人金斌锋在没有实际业务往来的情况下，以收取一定比例的费用差价为条件，介绍上海欣原纸塑包装材料有限公司虚开增值税专用发票。共虚开27份，销货单位为上海江岚电子科技有限公司和上海泉易电子科技有限公司，价税合计人民币240.1万余元，税额人民币34.8万余元。上述发票已由欣原公司入账并向税务机关申报抵扣税款。金斌锋在案发后如实供述了犯罪事实。(*From January 2015 to July 2016, the defendant, Jin Binfeng, without any genuine underlying business transactions, arranged for Shanghai Xinyuan Paper & Plastic Packaging Materials Co., Ltd. to issue value-added tax (VAT) special invoices by charging a service fee at a certain percentage of the invoice amount. In total, 27 invoices were falsely issued, with Shanghai Jianglan Electronic Technology Co., Ltd. and Shanghai Quanyi Electronic Technology Co., Ltd. listed as the sellers. The total amount (including VAT) was approximately RMB 2.401 million, including VAT of approximately RMB 0.348 million. These invoices were recorded in Xinyuan's accounts and used to declare input VAT credits to the tax authorities. After the case was uncovered, Jin truthfully confessed to the criminal facts.*)

Example-Answer: D

Text: 被告人叶俊在2017年4月11日10时许，在北京市丰台区七里庄"速八酒店"门前向樊某出售北京市增值税普通发票，现场被北京市公安局丰台分局东铁匠营派出所民警当场查获，起获增值税普通发票19张，票面金额共计人民币1875921.23元，且经鉴定均为真发票。(*The defendant, Ye Jun, was apprehended at approximately 10:00 on April 11, 2017, while selling Beijing value-added tax ordinary invoices to an individual surnamed Fan in front of the Super 8 Hotel in Qilizhuang, Fengtai District, Beijing. Officers from the Dongtiejianying Police Station of the Fengtai Branch of the Beijing Municipal Public Security Bureau seized 19 VAT ordinary invoices with a total face value of RMB 1,875,921.23; an appraisal confirmed that all of the invoices were genuine.*)

Answer: H

Choices: ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L", "M"]

Gold: 7

Table C13: An instance format with paired prompts for the TaxCrime dataset in the KA task.

Prompt: 请将下述税务舆情概括为一段约300字的摘要，要求突出主要事件、观点或问题核心。(*Please summarize the following tax-related public opinion into an approximately 300-word abstract, highlighting the main events and the core viewpoints or issues.*)

Zero-Shot

Query: {Prompt}

Text: 近日，网红二驴涉嫌商业欺诈及逃税事件引发舆论关注。多名受害人和合作方纷纷揭露其不当行为，特别是一位女老板称自己被二驴骗走3000万元。...。更有消息称，二驴的主持人华少也曾为其花费40万元包机，整场活动的费用累计超过3000万元。在这一事件的背后，二驴的税务问题也受到关注。有报道称，二驴通过其父亲并玉坤的账户进行大额收款，意图规避税务责任。这一做法引发了外界对其是否存在逃税行为的质疑。值得注意的是，二驴的妻子曾因涉嫌偷税漏税而被处罚，因此人们对二驴的税务问题更加关注。此举是否涉及逃税，以及这些收入是否依法缴纳税款，目前尚无官方回应。...。(*Recently, the influencer Er Lu has come under public scrutiny due to allegations of commercial fraud and tax evasion. Numerous victims and partners have come forward to expose his misconduct, particularly a female business owner who claims to have been defrauded of 30 million yuan. ... Furthermore, reports indicate that Er Lu's host, Hua Shao, spent 400,000 yuan on a charter flight for him, with the total expenses for the event exceeding 30 million yuan. Behind this incident, Er Lu's tax issues have also attracted attention. It has been reported that he received large sums through his father Bing Yukun's account in an apparent attempt to evade tax responsibilities. This practice has raised questions about whether he is engaging in tax evasion. Notably, Er Lu's wife has previously been penalized for tax evasion, which has drawn even more scrutiny to his tax matters. It remains unclear whether these actions involve tax evasion or if the income has been properly taxed, as there has been no official response yet. ...*)

Answer: 近日，网红二驴因涉嫌商业欺诈及逃税引发广泛关注。一名女老板曝光称，与二驴合作举办演唱会时投入了3000万元，但二驴未完成合同约定的销售目标，且未投入任何资金。女老板还提供了聊天记录和转账截图，揭露二驴虚假承诺，甚至收取了其他商家的抗位费。...。(*Recently, Chinese livestream celebrity Erlv has attracted widespread attention following allegations of commercial fraud and tax evasion. A female business owner claimed that she invested RMB 30 million to co-organize a concert with Erlv; however, Erlv allegedly failed to meet the contractual sales targets and did not contribute any capital. The business owner further provided purported chat logs and bank-transfer screenshots, alleging that Erlv made false promises and even collected so-called booth (slotting) fees from other merchants. ....*)

One-Shot

Query: {Prompt}

Example-Text: 记者从市市场监管委获悉，为期一年的加油机作弊专项治理行动已接近尾声。截至目前，我市已完成对全市加油站的全面检查，立案查处9起案件，规范了成品油零售市场秩序，有力遏制了加油机作弊、计量作弊和偷逃税等违法行为。...。下一步，市市场监管委将继续加大执法力度，将专项治理与常态化执法检查相结合，推动建立健全加油机计量监管的长效机制，进一步规范成品油零售市场秩序。(*The reporter learned from the Municipal Market Supervision Administration that the year-long special campaign against fuel dispenser fraud is nearing its conclusion. So far, comprehensive inspections of all gas stations in the city have been completed, resulting in the filing and investigation of nine cases, thereby standardizing the retail market order for refined oil and effectively curbing illegal activities such as fuel dispenser fraud, measurement cheating, and tax evasion. ... Going forward, the market regulation authority will continue to intensify enforcement, integrate the special campaign with routine inspections, and promote the establishment of a long-term mechanism for metrological supervision of fuel dispensers, thereby further standardizing the refined-oil retail market.*)

Example-Answer: 自去年8月起，市市场监管委联合多部门开展为期一年的加油机作弊专项治理行动，....。接下来，市市场监管委将加强执法力度，推动建立长期有效的加油机计量监管机制，进一步规范市场秩序。(*Since August of last year, the Municipal Market Supervision Administration, in collaboration with multiple departments, has launched a year-long special campaign against fuel dispenser fraud. ... Going forward, the Administration will further strengthen enforcement and promote the establishment of a long-term, effective regulatory mechanism for fuel-dispenser metrology to better standardize market order.*)

Text: 近日，网红二驴涉嫌商业欺诈及逃税事件引发舆论关注。多名受害人和合作方纷纷揭露其不当行为，特别是一位女老板称自己被二驴骗走3000万元。...。更有消息称，二驴的主持人华少也曾为其花费40万元包机，整场活动的费用累计超过3000万元。在这一事件的背后，二驴的税务问题也受到关注。有报道称，二驴通过其父亲并玉坤的账户进行大额收款，意图规避税务责任。这一做法引发了外界对其是否存在逃税行为的质疑。值得注意的是，二驴的妻子曾因涉嫌偷税漏税而被处罚，因此人们对二驴的税务问题更加关注。此举是否涉及逃税，以及这些收入是否依法缴纳税款，目前尚无官方回应。.... (*Recently, the influencer Er Lu has come under public scrutiny due to allegations of commercial fraud and tax evasion. Numerous victims and partners have come forward to expose his misconduct, particularly a female business owner who claims to have been defrauded of 30 million yuan. ... Furthermore, reports indicate that Er Lu's host, Hua Shao, spent 400,000 yuan on a charter flight for him, with the total expenses for the event exceeding 30 million yuan. Behind this incident, Er Lu's tax issues have also attracted attention. It has been reported that he received large sums through his father Bing Yukun's account in an apparent attempt to evade tax responsibilities. This practice has raised questions about whether he is engaging in tax evasion. Notably, Er Lu's wife has previously been penalized for tax evasion, which has drawn even more scrutiny to his tax matters. It remains unclear whether these actions involve tax evasion or if the income has been properly taxed, as there has been no official response yet. ...*)

Answer: 近日，网红二驴因涉嫌商业欺诈及逃税引发广泛关注。一名女老板曝光称，与二驴合作举办演唱会时投入了3000万元，但二驴未完成合同约定的销售目标，且未投入任何资金。女老板还提供了聊天记录和转账截图，揭露二驴虚假承诺，甚至收取了其他商家的抗位费。...。(*Recently, Chinese livestream celebrity Erlv has attracted widespread attention following allegations of commercial fraud and tax evasion. A female business owner claimed that she invested RMB 30 million to co-organize a concert with Erlv; however, Erlv allegedly failed to meet the contractual sales targets and did not contribute any capital. The business owner further provided purported chat logs and bank-transfer screenshots, alleging that Erlv made false promises and even collected so-called booth (slotting) fees from other merchants. ....*)

Table C14: An instance format with paired prompts for the TaxOpinion dataset in the KA task.

Prompt: 阅读下列财税风险控制相关内容，提炼出“面临的风险”及“风险对应的解决方案”，只需直接输出如下格式内容：{"面临的风险": , "风险对应的解决方案": } (*Read the following content related to fiscal and tax risk control, and extract (i) the “Risks Encountered” and (ii) the “Corresponding Mitigation Measures.” Output* only *the following JSON format: {"Risks Encountered": , "Corresponding Mitigation Measures": }*)
Zero-Shot
Query: {Prompt} **Text:** 建筑企业存在较大的税收风险，主要涉及错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。针对这些风险，建议企业要加强对税务风险的防控。对于总承包业务未分别核算的风险，建议按照不同税率核算销售额，避免高税率缴纳；对于简易计税进项税额未转出的风险，要注意用于简易计税项目的进项税额不可抵扣；对于合同金额与结算金额不一致的风险，应根据实际情况及时进行印花税补缴。企业在实际操作中需谨慎遵守政策规定，避免风险的发生。(*Construction enterprises face substantial tax risks, which mainly stem from (i) misapplication of taxable items and applicable tax rates, (ii) failure to reverse input VAT credits when adopting the simplified taxation method, and (iii) inconsistencies between contract amounts and settlement amounts. To mitigate these risks, enterprises should strengthen tax risk management and internal controls. Regarding the risk that general contracting activities are not accounted for separately, firms are advised to segregate sales by the corresponding tax rates to avoid being taxed at a higher rate. With respect to the failure to reverse input VAT credits under the simplified method, firms should note that input VAT attributable to simplified-taxation projects is not creditable. For discrepancies between contract and settlement amounts, enterprises should, where appropriate, promptly pay any additional stamp duty based on the actual circumstances. In practice, enterprises should adhere strictly to relevant tax policies and regulations to minimize the likelihood of non-compliance.*)
Answer: {"面临的风险": "错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。", "对应的解决措施": "加强税务风险防控、按照不同税率核算销售额、注意简易计税项目进项税额不可抵扣、及时进行印花税补缴。"} (*{"Risks Encountered": "Misapplication of tax items and rates; failure to reverse input VAT under the simplified taxation regime; inconsistencies between the contract amount and the settlement amount.", "Corresponding Mitigation Measures": "Strengthen tax risk control; account for sales revenue by applicable tax rates; note that input VAT for simplified taxation items is non-creditable; promptly settle any outstanding stamp duty through supplementary payment."}*.)
One-Shot
Query: {Prompt} **Example-Text:** 国家金融监管局数据显示，截至2023年末，我国普惠型小微企业贷款余额29.06万亿元，同比增长23.27%。普惠型涉农贷款余额12.59万亿元，同比增长20.34%。根据税法，金融企业需确保涉农和中小企业贷款损失的税前扣除符合条件。部分金融企业未正确提供税前扣除所需的完整证据，例如司法追索记录，导致税务风险。税务检查发现，甲银行未能提供完整的追索和资产清偿证明，致使部分贷款损失未获得税前扣除。政策规定要求金融企业对不同金额的贷款损失提供具体的追索记录和资产证明。建议金融机构遵守相关法规，确保贷款损失税前扣除的合规性和透明度。(*According to data released by the National Financial Regulatory Administration, by the end of 2023 China’s inclusive micro- and small-enterprise loan balance reached 29.06 trillion RMB, representing a year-on-year increase of 23.27%. Inclusive agriculture-related loans totaled 12.59 trillion RMB, up 20.34% year on year. Under the tax law, financial institutions must ensure that the pre-tax deduction of credit losses on agriculture-related and small- and medium-sized enterprise (SME) loans satisfies the applicable eligibility requirements. However, some institutions fail to furnish complete supporting documentation required for such deductions (e.g., records of judicial recovery efforts), thereby increasing tax compliance risk. In tax audits, Bank A was unable to provide comprehensive evidence of recovery actions and asset liquidation, resulting in the denial of pre-tax deductions for a portion of its loan losses. Relevant policies require financial institutions to submit specific recovery records and asset documentation for loan losses at different loss amounts. We therefore recommend that financial institutions strictly comply with the relevant regulations to improve the compliance and transparency of pre-tax deductions for loan losses.*) **Example-Answer:** {"面临的风险": "部分金融企业未正确提供税前扣除所需的完整证据，导致税务风险", "对应的解决措施": "建议金融机构遵守相关法规，确保贷款损失税前扣除的合规性和透明度。"} (*{"Risks Encountered": "Some financial institutions may fail to provide complete and appropriate documentation required for pre-tax deductions, thereby increasing tax compliance risks.", "Corresponding Mitigation Measures": "It is recommended that financial institutions comply with relevant regulations and ensure the compliance and transparency of pre-tax deductions for loan losses."}*) **Text:** 建筑企业存在较大的税收风险，主要涉及错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。针对这些风险，建议企业要加强对税务风险的防控。对于总承包业务未分别核算的风险，建议按照不同税率核算销售额，避免高税率缴纳；对于简易计税进项税额未转出的风险，要注意用于简易计税项目的进项税额不可抵扣；对于合同金额与结算金额不一致的风险，应根据实际情况及时进行印花税补缴。企业在实际操作中需谨慎遵守政策规定，避免风险的发生。(*Construction enterprises face substantial tax risks, which mainly stem from (i) misapplication of taxable items and applicable tax rates, (ii) failure to reverse input VAT credits when adopting the simplified taxation method, and (iii) inconsistencies between contract amounts and settlement amounts. To mitigate these risks, enterprises should strengthen tax risk management and internal controls. Regarding the risk that general contracting activities are not accounted for separately, firms are advised to segregate sales by the corresponding tax rates to avoid being taxed at a higher rate. With respect to the failure to reverse input VAT credits under the simplified method, firms should note that input VAT attributable to simplified-taxation projects is not creditable. For discrepancies between contract and settlement amounts, enterprises should, where appropriate, promptly pay any additional stamp duty based on the actual circumstances. In practice, enterprises should adhere strictly to relevant tax policies and regulations to minimize the likelihood of non-compliance.*)
Answer: {"面临的风险": "错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。", "对应的解决措施": "加强税务风险防控、按照不同税率核算销售额、注意简易计税项目进项税额不可抵扣、及时进行印花税补缴。"} (*{"Risks Encountered": "Misapplication of tax items and rates; failure to reverse input VAT under the simplified taxation regime; inconsistencies between the contract amount and the settlement amount.", "Corresponding Mitigation Measures": "Strengthen tax risk control; account for sales revenue by applicable tax rates; note that input VAT for simplified taxation items is non-creditable; promptly settle any outstanding stamp duty through supplementary payment."}*.)

Table C15: An instance format with paired prompts for the TaxRisk dataset in the KA task.

Prompt: 请提取以下税务稽查文章中的关键信息，内容包括：犯罪行为、所犯罪名与处罚结果。所犯罪名仅可在以下罪名中选择其一：持有伪造的发票罪，非法出售增值税专用发票罪，非法购买增值税专用发票罪，购买伪造的增值税专用发票罪，骗取出口退税罪，逃税罪，虚开发票罪。直接按照以下格式作答，无需添加解释说明：{"犯罪行为": , "所犯罪名": , "处罚结果": } (*Please extract the key information from the following tax audit case report, including: criminal act, the charged offense, and the penalty outcome. The charged offense must be selected from exactly one of the following options: Crime of holding forged invoices, Crime of illegally selling special VAT invoices, Crime of illegally purchasing special VAT invoices, Crime of purchasing forged special VAT invoices, Crime of fraudulently obtaining export tax refunds, Crime of tax evasion, Crime of falsely issuing invoices. Provide the answer strictly in the following format without any additional explanation: {"Criminal Act": , "Charged Offense": , "Penalty Outcome": }*)

Zero-Shot

Query: {Prompt}

Text: 税务局检查发现，A石油化工有限公司在2017年至2019年期间存在多起偷税行为。具体违法事实包括将购进的非应征消费税产品变为应征消费税产品进行销售，涉及煤油和普柴的数量分别为2017年590.7吨和2018年1927.6吨，以及2019年将有机化工原料变为汽油的6700.842吨。根据相关法律规定，该行为构成偷税行为，导致公司未缴纳的消费税总额达到1769.1万元。对此，税务局依法对A公司分别处以总计8845.5万元的罚款。(*During a tax inspection, the tax authority found that Company A (Petrochemical) committed multiple acts of tax evasion between 2017 and 2019. The specific violations included converting purchased non–excise-taxable products into excise-taxable products for sale. The quantities involved were 590.7 tons of kerosene and 1,927.6 tons of diesel in 2017 and 2018, respectively; additionally, in 2019, 6,700.842 tons of organic chemical feedstock were converted into gasoline. Under the relevant statutory provisions, these actions constitute tax evasion and resulted in unpaid excise tax totaling RMB 17.691 million. Accordingly, the tax authority imposed a total fine of RMB 88.455 million on Company A in accordance with law.*)

Answer: {"犯罪行为": "A石油化工有限公司将购进的非应征消费税产品变为应征消费税产品进行销售，涉及煤油、普柴及有机化工原料，未缴纳消费税总额达到1769.1万元。", "所犯罪名": "逃税罪", "处罚结果": "对A公司分别处以总计8845.5万元的罚款。"} (*{"Criminal Act": "Company A (a petroleum and petrochemical enterprise) sold products initially purchased as non-excise-taxable goods after converting them into excise-taxable products, involving kerosene, ordinary diesel, and organic chemical raw materials, and evaded a total of RMB 17.691 million in excise tax.", "Charged Offense": "Tax Evasion", "Penalty Outcome": "Company A was fined a total of RMB 88.455 million."}.*)

One-Shot

Query: {Prompt}

Example-Text: 2021年6月，曾某因涉嫌虚开增值税专用发票罪被捕，并于2021年12月接受了东莞市第三人民法院的公开审理。检方指控曾某利用瑞航塑胶公司名义，通过虚构交易方式获取5224059.35元的税款，并获得32份虚假发票。法院认定曾某系公司实际经营者，直接负责且主观恶性较轻。尽管曾某在庭审中未否认指控，但其辩护人提出其初犯、自愿认罪等情节。最终，法院依据《中华人民共和国刑法》第二百零五条，判处曾某有期徒刑11年。(*In June 2021, Mr. Zeng was arrested on suspicion of the crime of falsely issuing special value-added tax (VAT) invoices, and his case was publicly tried by the Third People's Court of Dongguan in December 2021. The prosecution alleged that Mr. Zeng, acting in the name of Ruihang Plastics Co., Ltd., obtained RMB 5,224,059.35 in tax proceeds by fabricating transactions and acquired 32 fraudulent invoices. The court found that Mr. Zeng was the de facto operator of the company, bore direct responsibility, and exhibited a relatively low degree of subjective maliciousness. Although Mr. Zeng did not deny the allegations during the trial, his defense counsel argued mitigating circumstances, including that he was a first-time offender and voluntarily pleaded guilty. Ultimately, pursuant to Article 205 of the Criminal Law of the People's Republic of China, the court sentenced Mr. Zeng to 11 years of fixed-term imprisonment.*)

Example-Answer: {"犯罪行为": "利用瑞航塑胶公司名义，通过虚构交易方式获取税款，并获得虚假发票。", "所犯罪名": "虚开发票罪", "处罚结果": "判处有期徒刑11年。"} (*{"Criminal Act": "Using the name of Ruihang Plastics Co., Ltd., the defendant obtained tax refunds through fictitious transactions and procured fraudulent invoices.", "Charged Offense": "Crime of issuing false invoices.", "Penalty Outcome": "Sentenced to 11 years' imprisonment."}*)

Text: 税务局检查发现，A石油化工有限公司在2017年至2019年期间存在多起偷税行为。具体违法事实包括将购进的非应征消费税产品变为应征消费税产品进行销售，涉及煤油和普柴的数量分别为2017年590.7吨和2018年1927.6吨，以及2019年将有机化工原料变为汽油的6700.842吨。根据相关法律规定，该行为构成偷税行为，导致公司未缴纳的消费税总额达到1769.1万元。对此，税务局依法对A公司分别处以总计8845.5万元的罚款。(*During a tax inspection, the tax authority found that Company A (Petrochemical) committed multiple acts of tax evasion between 2017 and 2019. The specific violations included converting purchased non–excise-taxable products into excise-taxable products for sale. The quantities involved were 590.7 tons of kerosene and 1,927.6 tons of diesel in 2017 and 2018, respectively; additionally, in 2019, 6,700.842 tons of organic chemical feedstock were converted into gasoline. Under the relevant statutory provisions, these actions constitute tax evasion and resulted in unpaid excise tax totaling RMB 17.691 million. Accordingly, the tax authority imposed a total fine of RMB 88.455 million on Company A in accordance with law.*)

Answer: {"犯罪行为": "A石油化工有限公司将购进的非应征消费税产品变为应征消费税产品进行销售，涉及煤油、普柴及有机化工原料，未缴纳消费税总额达到1769.1万元。", "所犯罪名": "逃税罪", "处罚结果": "对A公司分别处以总计8845.5万元的罚款。"} (*{"Criminal Act": "Company A (a petroleum and petrochemical enterprise) sold products initially purchased as non-excise-taxable goods after converting them into excise-taxable products, involving kerosene, ordinary diesel, and organic chemical raw materials, and evaded a total of RMB 17.691 million in excise tax.", "Charged Offense": "Tax Evasion", "Penalty Outcome": "Company A was fined a total of RMB 88.455 million."}.*)

Table C16: An instance format with paired prompts for the TaxInspect dataset in the KA task.

Prompt: 请补全下方空项，完成后请按题目原格式返回填写结果。核心思路为文字回答，其余字段统一填写数字，若为百分比项，请勿添加百分号，仅填数字。{"核心思路": , "原年收入": , "筹划后年收入": , "节增收入": , "原应纳增值税": , "筹划后应纳增值税": } (*Please complete the missing fields below. After completion, return the filled results in the original format. The "Core Rationale" should be answered in text, and all other fields must be filled with numbers only. For percentage fields, do not include the percent sign; provide only the numeric value.{"Core Rationale": , "Original Annual Income": , "Post-Planning Annual Income": , "Incremental Income": , "Original VAT Payable": , "Post-Planning VAT Payable": }*)
Zero-Shot
Query: {Prompt} **Text:** 某个体工商户销售水果、杂货，每月含税销售额为20600元左右，当地财政厅和国家税务局规定的增值税起征点为20000元。请计算该个体工商户全年应纳增值税额，并提出纳税筹划方案。（不考虑小微企业免征增值税优惠）(*A sole proprietorship engages in the retail sale of fruits and general groceries, with an average monthly tax-inclusive sales revenue of approximately RMB 20,600. The VAT threshold (minimum taxable turnover) stipulated by the local Department of Finance and the State Taxation Administration is RMB 20,000. Please calculate the annual value-added tax (VAT) payable by this taxpayer and propose an appropriate tax planning strategy. (Assume that the VAT exemption for small and micro enterprises does not apply.)*)
Answer: {"核心思路": "通过将每月含税销售额降低到略低于起征点，个体工商户可以避免增值税的缴纳，从而实现全年收入的增加。", "原年收入": 24, "筹划后年收入": 24.6, "节增收入": 0.6, "原应纳增值税": 0.72, "筹划后应纳增值税": 0} (*{"Core Rationale": "By reducing the monthly VAT-inclusive sales revenue to just below the VAT exemption threshold, a self-employed household business can avoid paying VAT, thereby increasing annual income.", "Original Annual Income": 24, "Post-Planning Annual Income": 24.6, "Incremental Income": 0.6, "Original VAT Payable": 0.72, "Post-Planning VAT Payable": 0}*.)
One-Shot
Query: {Prompt} **Example-Text:** 某企业集团下属甲、乙两个企业。其中，甲企业适用25%的企业所得税税率，乙企业属于需要国家扶持的高新技术企业，适用15%的企业所得税税率。2019纳税年度，甲企业的应纳税所得额为8000万元，乙企业的应纳税所得额为9000万元。请计算甲乙两个企业以及该企业集团在2019纳税年度分别应当缴纳的企业所得税税款，并提出纳税筹划方案。(*A corporate group comprises two subsidiaries, Firm A and Firm B. Firm A is subject to the standard corporate income tax (CIT) rate of 25%, whereas Firm B qualifies as a state-supported high-tech enterprise and is subject to a preferential CIT rate of 15%. In the 2019 tax year, Firm A reports taxable income of RMB 80 million, and Firm B reports taxable income of RMB 90 million. Please compute the CIT liabilities payable by Firm A and Firm B, as well as the total CIT liability of the corporate group for the 2019 tax year, and propose an appropriate tax-planning strategy.*) **Example-Answer:** {"核心思路": "通过将高税率企业的部分应纳税所得转移至低税率企业，可以在集团内实现整体税负优化，降低企业集团应缴税款。", "甲企业税前所得（原始）": 8000, "乙企业税前所得（原始）": 9000, "甲企业所得税（原始）": 2000, "乙企业所得税（原始）": 1350, "集团总所得税（原始）": 3350, "甲企业税前所得（筹划后）": 7000, "乙企业税前所得（筹划后）": 10000, "甲企业所得税（筹划后）": 1750, "乙企业所得税（筹划后）": 1500, "集团总所得税（筹划后）": 3250, "税负减少": 100} (*{"Core Rationale": "By reallocating a portion of taxable income from the high-tax-rate entity to the low-tax-rate entity, the group can achieve overall tax burden optimization and reduce the total corporate income tax payable.", "Taxable Income of Company A (Pre-tax, Original)": 8000, "Taxable Income of Company B (Pre-tax, Original)": 9000, "Corporate Income Tax of Company A (Original)": 2000, "Corporate Income Tax of Company B (Original)": 1350, "Total Group Corporate Income Tax (Original)": 3350, "Taxable Income of Company A (Pre-tax, After Planning)": 7000, "Taxable Income of Company B (Pre-tax, After Planning)": 10000, "Corporate Income Tax of Company A (After Planning)": 1750, "Corporate Income Tax of Company B (After Planning)": 1500, "Total Group Corporate Income Tax (After Planning)": 3250, "Tax Burden Reduction": 100}*) **Text:** 某个体工商户销售水果、杂货，每月含税销售额为20600元左右，当地财政厅和国家税务局规定的增值税起征点为20000元。请计算该个体工商户全年应纳增值税额，并提出纳税筹划方案。（不考虑小微企业免征增值税优惠）(*A sole proprietorship engages in the retail sale of fruits and general groceries, with an average monthly tax-inclusive sales revenue of approximately RMB 20,600. The VAT threshold (minimum taxable turnover) stipulated by the local Department of Finance and the State Taxation Administration is RMB 20,000. Please calculate the annual value-added tax (VAT) payable by this taxpayer and propose an appropriate tax planning strategy. (Assume that the VAT exemption for small and micro enterprises does not apply.)*)
Answer: {"核心思路": "通过将每月含税销售额降低到略低于起征点，个体工商户可以避免增值税的缴纳，从而实现全年收入的增加。", "原年收入": 24, "筹划后年收入": 24.6, "节增收入": 0.6, "原应纳增值税": 0.72, "筹划后应纳增值税": 0} (*{"Core Rationale": "By reducing the monthly VAT-inclusive sales revenue to just below the VAT exemption threshold, a self-employed household business can avoid paying VAT, thereby increasing annual income.", "Original Annual Income": 24, "Post-Planning Annual Income": 24.6, "Incremental Income": 0.6, "Original VAT Payable": 0.72, "Post-Planning VAT Payable": 0}*.)

Table C17: An instance format with paired prompts for the TaxPlan dataset in the KA task.

Data	Prompt	Input	Target	Output	Metric AC NA IN AVG			
TaxRecite	根据问题需求，请提供法条的原文部分，不必对其含义做出解释，也不需要指明是第几条。	2007年12月6日 通 过的《中华人民共和国企业所得税法实施条例》中第一百一十一条的内容是什么？	企业所得税法第四十一条所称合理方法，包括：(一)可比非受控价格法；(二)再销售...(六)其他符合独立交易原则的方法。	企业所得税法第四十一条所称合理方法，包括：(一)可比非受控价格法；(二)再销售...(六)其他符合独立交易原则的方法。	3	3	3	3
TaxSum	根请根据以下文本编写一个简短的概要，字数不超过300字，重点在于提炼核心内容。	为配合个人所得税专项附加扣除工作，依据相关法律法规，三部门联合发布通知，明确个人所得税住房贷款利息专项附加扣除相关信息归集事项。信息归集范围为1989年1月1日（含）后发放的商业性个人住房贷款，不包括个人商用房贷款。...	信息归集工作时间紧、任务重、要求高，各商业银行要高度重视，充分认识贯彻施行《个人所得税专项附加扣除暂行办法》的重要性和紧迫性，切实加强组织领导，务必在规定时间内做好各项工作。	本次信息归集工作期限紧、工作量大、标准要求严格。各商业银行应提高站位，充分把握《个人所得税专项附加扣除暂行办法》贯彻实施的重要意义与紧迫要求，强化统筹协调和责任落实，完善工作机制与保障措施，确保在规定时限内高质量完成各项工作任务。	3	2	2	2.3
TaxTopic	阅读标题以了解文章主题，并在提供的选项中选择最适合表达这一主题的一个，选项有：A.财商资讯 B.财税表单 C.地方法规 D.法规解读 E.纳税调整 F.纳税评估 G.税收优惠 H.税收筹划 I.税务问答 J.中央法规。	浙江省国家税务局公告2014年第10号 浙江省国家税务局关于企业资产损失税前扣除申报有关问题的公告	C	C	3	3	3	3
TaxRead	通过阅读以下文本，然后回答以下问题。	为综合反映与科学评价资产评估机构整体实力和专业胜任能力，...问题：《资产评估机构综合评价办法》何时正式发布？	2025年2月18日	2024年7月6日。《资产评估机构综合评价办法》的发布时间为2024年7月6日。	1	1	3	1.7
TaxCalc	针对题目，请直接给出xx处的答案，并将答案带入到问题中复述一遍，例如："业务（1）应缴纳消费税为xx万元"，保证输出字符在30个字符以内。	位于某市区国家重点扶持的高新技术企业为增值税一般纳税人（制造业企业），2023年销售产品取得不含税收入6500万元... 问题为：新产品研发费应调整的应纳税所得额为xx万元？	-72.75	-72.75	3	3	3	3
TaxSCQ	关于下面提出的问题，请仔细考虑四个可能的答案，并选出你认为正确的那个，仅需指出A、B、C或D，不需要解释。	关于退还关税，下列说法正确的是()。A.纳税人自缴纳税款之日起3年内...D.海关应当自受理退税申请之日起60日内查实并通知纳税义务人办理退还手续	C	C	3	3	3	3

Table C19: Examples of human evaluation with metrics like accuracy(AC), naturalness(NA), informativeness(IN) and average(AVG) for human written prompts in various tax datasets. We retain the prompts with AVG above 2.

Data	Prompt	Input	Target	Output	Metric AC NA IN AVG
TaxMCQ	针对以下问题，请分析五个选项，并选择最适合的两个或更多，然后按字母顺序提供A、B、C、D或E作为你的答案，输出格式为字母组合，字母之间不要有任何其他字符，且必须是顺序的。	关于不动产租赁服务的增值税处理，正确的有()。A.个体工商户异地出租不动产，在不动产所在地预缴的增值税款，可在当期增值税税款中抵减 B.以经营租赁方式将土地出租给他人使用... E.纳税人向其他个人出租不动产，可以开具增值税专用发票	ABC	A	1 3 1 1.7
TaxQA	对于下面提出的税务问题，请准备简要的答案，注意每项答复应保持在300字之内。	转让债权如何确定其应予减除的财产原值和合理费用?	采用“加权平均法”确定其应予减除的财产原值和合理费用。即以纳税人购进的同一种类债券买入价和买进过程中缴纳的税费总和，除以纳税人购进的该种类债券数量之和，乘以纳税人卖出的该种类债券数量，再加上卖出的该种类债券过程中缴纳的税费。	转让债权时，一般可采用加权平均法来确定应予减除的财产原值和合理费用：将纳税人购入同一类别债券的累计买入金额汇总后，除以累计购入数量，计算出单位成本；再用该单位成本乘以本次卖出的数量，作为可减除的财产原值，相关合理费用按实际发生情况一并计入扣除。	2 3 2 2.3
TaxBoard	对于由群众提出的税务问题，请提供确切的解答，并指出相关的法律依据，确保每个回答都在300字以内。	标题：文化事业建设费今年还有减半优惠吗？ 内容：文化事业建设费今年还有减半优惠吗?	根据《财政部关于延续实施文化事业建设费优惠政策的通知》(财税 〔2025〕7号)规定，自2025年1月1日至2027年12月31日，对归属中央收入的文化事业建设费，将按应缴费额的50%减征；对归属地方收入...。	有。依据《财政部关于延续实施文化事业建设费优惠政策的通知》（财税〔2025〕7号）：（1）归属中央收入...。	2 2 3 2.3
TaxCrime	阅读下面的犯罪事实，请判断其主要违反了以下哪一法律条款：A. 刑法第二百零一条 B. 刑法第二百零三条 C. 刑法第二百零四条 D. 刑法第二百零五条 E. 刑法第二百零六条 F. 刑法第二百零七条 G. 刑法第二百零八条 H. 刑法第二百零九条 I. 刑法第二百一十条 J. 刑法第二百一十一条 K. 刑法第二百一十二条 L. 刑法第二百二十七条 M. 刑法第二百六十四条。只需输出对应字母，不需要解释原因。	被告人叶俊在2017年4月11日10时许，在北京市丰台区七里庄“速八酒店”门前向樊某出售北京市增值税普通发票，现场被北京市公安局丰台分局东铁匠营派出所民警当场查获，起获增值税普通发票19张，票面金额共计人民币1875921.23元，且经鉴定均为真发票。	H	H	3 3 3 3

Table C20: Examples of human evaluation with metrics like accuracy(AC), naturalness(NA), informativeness(IN) and average(AVG) for human written prompts in various tax datasets. We retain the prompts with AVG above 2.

Data	Prompt	Input	Target	Output	Metric AC	NA	IN	AVG
TaxOpinion	请将下述税务舆情概括为一段约300字的摘要，要求突出主要事件、观点或问题核心。	记者从市市场监管委获悉，为期一年的加油机作弊专项治理行动已接近尾声。...。下一步，市市场监管委将继续加大执法力度，将专项治理与常态化执法检查相结合，推动建立健全加油机计量监管的长效机制，进一步规范成品油零售市场秩序。	自去年8月起，市市场监管委联合多部门开展为期一年的加油机作弊专项治理行动，已基本完成。...接下来，市市场监管委将加强执法力度，推动建立长期有效的加油机计量监管机制，进一步规范市场秩序。	市市场监管委通报，自去年8月起，会同公安、商务、税务等部门在全市开展为期一年的加油机作弊专项治理行动，目前已接近尾声并完成对加油站的全面检查。...下一步将把专项治理与常态化检查结合，健全长效监管机制。	3	1	3	2.3
TaxRisk	阅读下列财税风险控制相关内容，提炼出“面临的风险”及“风险对应的解决方案”，只需直接输出如下格式内容：{"面临的风险": ,"风险对应的解决方案": }	建筑企业存在较大的税收风险，主要涉及错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。针对这些风险，建议企业要加强对税务风险的防控。...企业在实际操作中需谨慎遵守政策规定，避免风险的发生。	{"面临的风险": "错误适用税目税率、简易计税进项税额未转出、合同金额与结算金额不一致。", "对应的解决措施": "加强税务风险防控、按照不同税率核算销售额、注意简易计税项目进项税额不可抵扣、及时进行印花税补缴。"}	{"面临的风险":"税目税率适用错误、简易计税进项未转出、合同金额与结算金额不一致。","风险对应的解决方案":"加强税务风控；按不同税率分别核算；简易计税进项不得抵扣并转出；按实际补缴印花税。"}	2	3	3	2.7
TaxInspect	请提取以下税务稽查文章中的关键信息，内容包括：犯罪行为、所犯罪名与处罚结果。所犯罪名仅可在以下罪名中选择其一：持有伪造的发票罪，非法出售增值税专用发票罪，非法购买增值税专用发票罪，购买伪造的增值税专用发票罪，骗取出口退税罪，逃税罪，虚开发票罪。直接按照以下格式作答，无需添加解释说明：{"犯罪行为": ,"所犯罪名": ,"处罚结果": }	税务局检查发现，A石油化工有限公司在2017年至2019年期间存在多起偷税行为。...根据相关法律规定，该行为构成偷税行为，导致公司未缴纳的消费税总额达到1769.1万元。对此，税务局依法对A公司分别处以总计8845.5万元的罚款。	{"犯罪行为": "A石油化工有限公司将购进的非应征消费税产品变为应征消费税产品进行销售，涉及煤油、普柴及有机化工原料，未缴纳消费税总额达到1769.1万元。", "所犯罪名": "逃税罪", "处罚结果": "对A公司分别处以总计8845.5万元的罚款。"}	{"犯罪行为":"A石油化工有限公司将购进的非应征消费税产品转换为应征消费税产品后销售，涉及煤油、普柴及有机化工原料，少缴消费税1769.1万元。","所犯罪名":"逃税罪","处罚结果":"被税务机关处罚款8845.5万元。"}	3	3	2	2.7
TaxPlan	请补全下方空项，完成后请按题目原格式返回填写结果。核心思路为文字回答，其余字段统一填写数字，若为百分比项，请勿添加百分号，仅填数字。{"核心思路": ,"原年收入": ,"筹划后年收入": ,"节增收入": ,"原应纳增值税": ,"筹划后应纳增值税": }	某个体工商户销售水果、杂货，每月含税销售额为20600元左右，当地财政厅和国家税务局规定的增值税起征点为20000元。请计算该个体工商户全年应纳增值税额，并提出纳税筹划方案。（不考虑小微企业免征增值税优惠）	{"核心思路": "通过将每月含税销售额降低到略低于起征点，个体工商户可以避免增值税的缴纳，从而实现全年收入的增加。", "原年收入": 24, "筹划后年收入": 24.6, "节增收入": 0.6, "原应纳增值税": 0.72, "筹划后应纳增值税": 0}	{"核心思路": "个体工商户如果把每月含税销售额控制在起征点以下，就可以免缴增值税，从而提高全年到手收入。", "原年收入": 24, "筹划后年收入": 27, "节增收入": 3, "原应纳增值税": 0.72, "筹划后应纳增值税": 0}	1	3	3	2.3

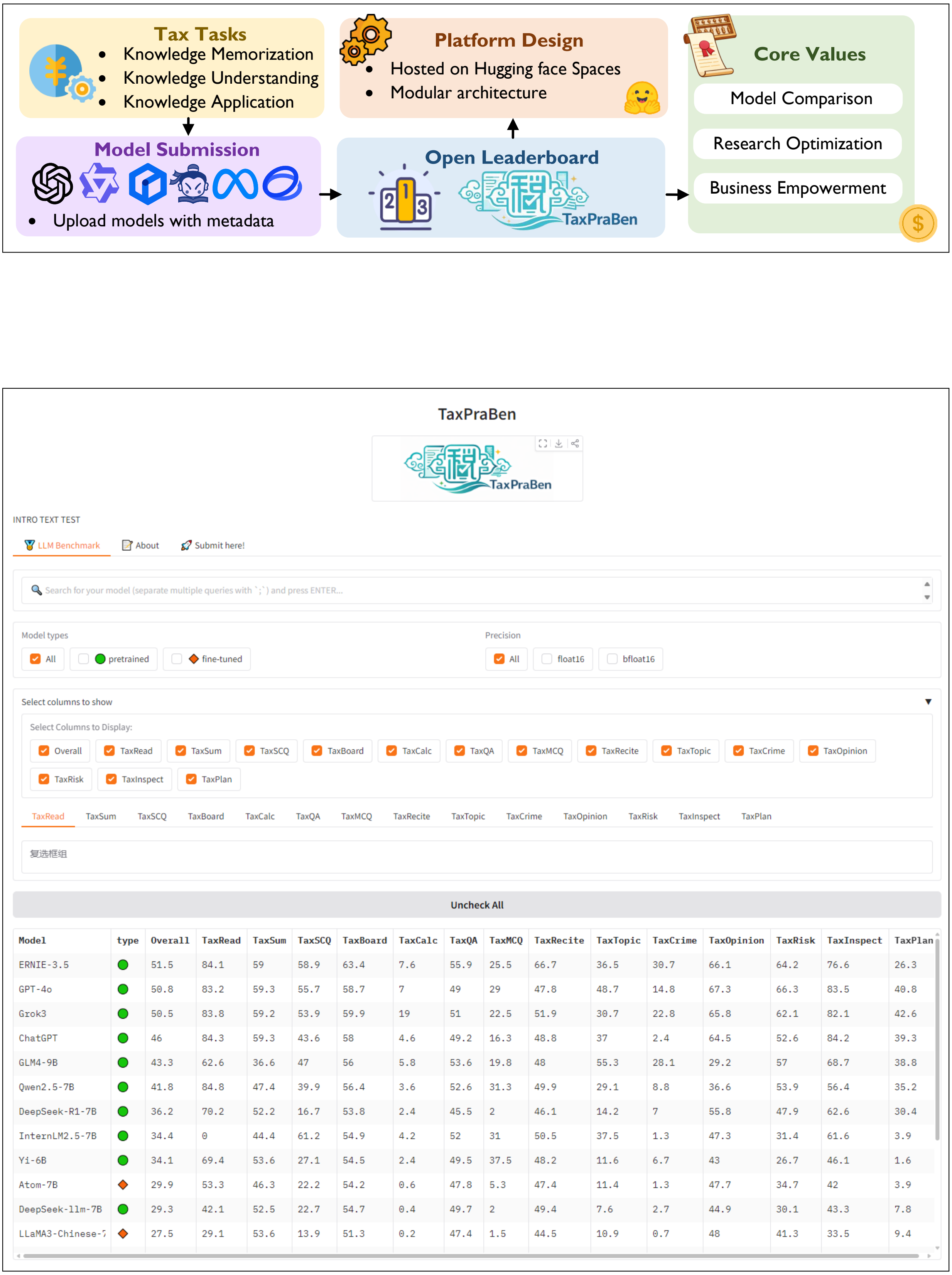
Table C21: Examples of human evaluation with metrics like accuracy(AC), naturalness(NA), informativeness(IN) and average(AVG) for human written prompts in various tax datasets. We retain the prompts with AVG above 2.

Task	Dataset	Prompt	English Translation
KM	TaxRecite	根据问题需求，请提供法条的原文部分，不必对其含义做出解释，也不需要指明是第几条。	Please provide the original text of the legal provision based on the question requirements, without explaining its meaning or indicating which article it is from.
KU	TaxSum	请根据以下文本编写一个简短的概要，字数不超过300字，重点在于提炼核心内容。	Please provide a brief summary based on the following text, keeping it under 300 words and focusing on extracting the core content.
	TaxTopic	阅读标题以了解文章主题，并在提供的选项中选择最适合表达的一个选项，选项有 A.财商资讯 B.财税表单 C.地方法规 D.法规解读 E.纳税调整 F.纳税评估 G.税收优惠 H.税收筹划 I.税务问答 J.中央法规	Read the title to grasp the article's theme and select the option that best reflects it. The options are: A. Financial Information B. Tax Forms C. Local Regulations D. Regulatory Interpretation E. Tax Adjustments F. Tax Assessment G. Tax Incentives H. Tax Planning I. Tax Q&A J. Central Regulations.
	TaxRead	在阅读完下面的文本后，请根据其内容回答问题，确保回复是从文中摘录且言简意赅，不要重复问题，答案中不要出现句号。	After reading the text, please provide concise excerpts for the questions without repeating them or using periods in your answers.
KA	TaxCalc	针对题目，请直接给出xx处的答案，并将答案带入到问题中复述一遍，例如："业务（1）应缴纳消费税为xx万元"，保证输出字符在30个字符以内。	For the question, please provide the answer from the designated section, and incorporate it into the question. For example: "The consumption tax payable by the business (1) is xx million yuan." Ensure the output is within 30 characters.
	TaxSCQ	关于下面提出的问题，请仔细考虑四个可能的答案，并选出你认为正确的那个，仅需指出“A”、“B”、“C”或“D”，不需要解释。	Regarding the question posed below, please carefully consider the four possible answers and select the one you believe is correct. Simply indicate "A", "B", "C", or "D" without any explanation.
	TaxMCQ	针对以下问题，请分析五个选项，并选择最适合的两个或更多，然后按字母顺序提供“A”、“B”、“C”、“D”或“E”作为你的答案，输出格式为字母组合，字母之间不要有任何其他字符，且必须按顺序的排列的。	For the following question, please analyze the five options and select the two or more that are most suitable. Provide your answer in the format of a letter combination, with letters in order and without any other characters between them: "A", "B", "C", "D" or "E."
	TaxQA	对于下面提出的税务问题，请准备简要的答案，注意每项答复应保持在300字之内。	For the tax-related questions posed below, please prepare concise answers, ensuring that each response remains within 300 words.
	TaxBoard	对于由群众提出的税务问题，请提供确切的解答，并指出相关的法律依据，确保每个回答都在300字以内。	For the tax questions raised by the public, please provide precise answers and indicate the relevant legal basis, ensuring that each response is within 300 words.
	TaxCrime	请判断其最主要违反哪个条款：A.刑法第二百零一条 B.刑法第二百零三条 C.刑法第二百零四条 D.刑法第二百零五条 E.刑法第二百零六条 F.刑法第二百零七条 G.刑法第二百零八条 H.刑法第二百零九条 I.刑法第二百一十条 J.刑法第二百一十一条 K.刑法第二百一十二条 L.刑法第二百二十七条 M.刑法第二百六十四条。请只返回选项字母。	Please determine which article is primarily violated: A. No. 201 of the Criminal Law B. No. 203 of the Criminal Law C. No. 204 of the Criminal Law D. No. 205 of the Criminal Law E. No. 206 of the Criminal Law F. No. 207 of the Criminal Law G. No. 208 of the Criminal Law H. No. 209 of the Criminal Law I. No. 210 of the Criminal Law J. No. 211 of the Criminal Law K. No. 212 of the Criminal Law L. No. 227 of the Criminal Law M. No. 264 of the Criminal Law. Please return only the letter of the option.
	TaxOpinion	根据所提供的税务舆情内容，请提炼出主要信息，形成一段300字左右的摘要，确保表述准确、重点突出。	Based on the provided tax-related public opinion content, please extract the main information and create a summary of approximately 300 words, ensuring that the expression is accurate and the key points are highlighted.
	TaxRisk	请根据下文内容提取文章所描述的税务风险及其应对措施，按以下结构输出即可：\n 面临的风险：xxx \n 风险对应的解决方案：xxx	Please extract the tax risks and their corresponding countermeasures described in the following content, and output them in the specified structure: \n Facing risks: xxx \n Corresponding solutions: xxx
	TaxInspect	阅读下文税务稽查材料，请识别并输出文章中涉及的三项内容：犯罪行为、所犯罪名及处罚结果。所犯罪名应从以下列表中选取：持有伪造的发票罪、非法出售增值税专用发票罪、非法购买增值税专用发票罪、购买伪造的增值税专用发票罪、骗取出口退税罪、逃税罪、虚开发票罪。只需使用下述格式作答，无需其他说明：\n 犯罪行为：xxx \n 所犯罪名：xxx \n 处罚结果：xxx	Please identify the three items: criminal behavior, crime committed, and penalty result. Select the crime from the following list: possession of forged invoices, illegal sale of special VAT invoices, illegal purchase of special VAT invoices, purchase of forged special VAT invoices, fraudulently obtaining export tax refunds, tax evasion, and issuing false invoices. Use the format below without additional explanation: \n Criminal behavior: xxx \n Crime committed: xxx \n Penalty result: xxx: .
	TaxPlan	请填写以下空项内容，并按格式输出。除“核心思路”需写完整句子外，其它所有空项均应只写数字，不含单位，百分数项也不要带“%”。{"核心思路": ,"原始应调增所得额": ,"原始企业所得税增加额": , "原始代扣代缴个税": , "筹划后企业所得税减少": ,"筹划后个税减少": }	Please fill in the fields and output as specified. "Core Idea" should be a complete sentence; all other fields should contain only numbers without units, and percentage fields should not include "%". "Core Idea": "xxx", "Original Adjusted Income Amount": xxx, "Original Corporate Income Tax Increase": xxx, "Original Withheld Individual Income Tax": xxx, "Post-Planning Corporate Income Tax Reduction": xxx, "Post-Planning Individual Income Tax Reduction": xxx.

Table C23: Examples of human-annotated prompts and their corresponding English translations in TaxPraBen.

Domain	Case Description	Raw Text	Structured JSON	Evaluation Metrics
Legal	刑事案件裁判要旨与量刑结果生成：给定案情与法条要点，让模型输出“裁判理由＋刑期/罚金等可核对数值”。(*Case Ruling Summary and Sentencing Generation: Given the circumstances of the case and legal provisions, the model is asked to output "ruling reasons + verifiable numerical values such as sentence length/fines"*)	被告人张某入户盗窃财物价值约人民币50,000元，到案后如实供述并退赔部分损失，依法可从轻处罚。综合犯罪情节与认罪认罚情况，建议判处有期徒刑一年六个月，缓刑二年，并处罚金人民币10,000元。(*The defendant, Zhang Mou, stole property worth approximately 50,000 yuan by entering a house, and truthfully confessed to the crime and partially compensated for the losses after being apprehended. He can be given a lighter punishment according to the law. Considering the criminal circumstances and his guilty plea, it is recommended to sentence him to one year and six months in prison, with a two-year probation period, and a fine of 10,000 yuan.*)	{"核心思路":"基于入户盗窃的法定构成与认罪认罚、退赔等从轻情节，依法确定从轻量刑并适用缓刑。", "刑期_月":18, "缓刑_月":24, "罚金_元":10000} *{"Core Idea": "Based on the legal composition of housebreaking and mitigating circumstances such as a guilty plea and compensation, a lighter sentence is determined and probation is applied.", "Sentence_Months": 18, "Probation_Months": 24, "Fine_Yuan": 10000}*	语义维度: BERTScore(核心思路) 数值维度: EM Accuracy(刑期_月, 缓刑_月, 罚金_元) *Semantic Dimension: BERTScore (Core_Reasoning)* *Numerical Dimension: EM Accuracy (Sentence_Months, Probation_Months, Fine_Yuan)*
Finance	审计风险说明＋差异率核对：给定对账摘要与抽样结果，让模型输出“风险结论与原因＋交易金额/差异率等关键数字”。(*Audit Risk explanation + discrepancy rate verification: given reconciliation summaries and sampling results, the model outputs "risk conclusion with reasons + key figures such as transaction amount/discrepancy rate".*)	本期收入确认存在较高错报风险：部分合同在未完成履约义务前提前确认收入，且回款与发货节点不一致。抽样核对显示异常交易金额约1,250,000元，占抽样总金额12,500,000元的10.0%。建议对相关收入进行重分类或调减，并将该类交易纳入重点复核。(*There is a high risk of misstatement in revenue recognition: Some contracts recognized revenue prematurely before fulfilling obligations, and payment/receipt nodes are inconsistent with delivery. Sampling verification shows abnormal transaction amount of approximately RMB 1,250,000, accounting for 10.0% of the total sampled amount (RMB 12,500,000). It is recommended to reclassify or reduce relevant revenue and include such transactions in key review.*)	{"风险结论":"收入确认与履约/回款节点不一致导致潜在提前确认，形成重大错报风险，需调整并重点复核。", "异常交易金额":1250000, "抽样总金额":12500000, "差异率":0.10} *{"Risk_Conclusion": "Inconsistency between revenue recognition and performance/payment nodes leads to potential premature recognition, creating a material misstatement risk requiring adjustment and key review"., "Abnormal_Transaction_Amount": 1250000, "Total_Sampled_Amount": 12500000, "Discrepancy_Rate": 0.10 }*	语义维度: BERTScore(风险结论) 数值维度: EM Accuracy(异常交易金额, 抽样总金额, 差异率) *Semantic Dimension: BERTScore (Risk_Conclusion)* *Numerical Dimension: EM Accuracy (Abnormal_Transaction_Amount, Total_Sampled_Amount, Discrepancy_Rate)*
Medicine	出院小结自动生成：给定病程与化验单，模型输出“诊断与处置叙述＋生命体征/化验值等可核对数值”。(*Discharge summary auto-generation: given medical records and test reports, the model outputs "diagnosis and treatment narrative + verifiable values such as vital signs/laboratory results".*)	患者以发热、咳嗽咳痰入院，影像提示右下肺感染，诊断为肺炎；合并2型糖尿病血糖控制欠佳。入院体温 38.5°C，WBC 12.3 × 10^9/L，空腹血糖 11.2 mmol/L。予以头孢类抗感染、雾化祛痰及胰岛素强化控制血糖后症状改善，体温恢复正常，建议规律随访复查血常规与血糖。(*The patient was admitted with fever and cough; imaging indicated right lower lung infection, diagnosed as pneumonia; complicated with poorly controlled type 2 diabetes. Admission Temperature: 38.5°C, WBC 12.3×10^9/L, Fasting Blood Glucose: 11.2mmol/L. After cephalosporin anti-infection, aerosol expectoration, and insulin-intensive glucose control, symptoms improved with normalized temperature. Regular follow-up for blood routine and glucose recheck is recommended.*)	{"诊断叙述":"依据症状与影像诊断肺炎，结合白细胞升高支持感染；同时强化控糖并随访复查关键指标。", "体温(°C)":38.5, "WBC(10^9/L)":12.3, "空腹血糖(mmol/L)":11.2} *{"Diagnosis_Narrative": "Pneumonia diagnosed based on symptoms and imaging, supported by elevated white blood cells; concurrent intensive glucose control with follow-up for key indicators"., "Temperature(°C)": 38.5, "WBC(10^9/L)": 12.3, "Fasting_Glucose(mmol/L)": 11.2 }*	语义维度: BERTScore(诊断叙述) 数值维度: EM Accuracy(体温_°C, WBC_10^9/L, 空腹血糖_mmol/L) *Semantic Dimension: BERTScore (Diagnosis_Narrative)* *Numerical Dimension: EM Accuracy (Temperature_°C, WBC_10^9/L, Fasting_Glucose(mmol/L))*

Table D: Examples of our structured evaluation approach applied in legal, financial, and medical domains.



Model	type	Overall	TaxRead	TaxSum	TaxSCQ	TaxBoard	TaxCalc	TaxQA	TaxMCQ	TaxRecite	TaxTopic	TaxCrime	TaxOpinion	TaxRisk	TaxInspect	TaxPlan
ERNIE-3.5	●	51.5	84.1	59	58.9	63.4	7.6	55.9	25.5	66.7	36.5	30.7	66.1	64.2	76.6	26.3
GPT-4o	●	50.8	83.2	59.3	55.7	58.7	7	49	29	47.8	48.7	14.8	67.3	66.3	83.5	40.8
Grok3	●	50.5	83.8	59.2	53.9	59.9	19	51	22.5	51.9	30.7	22.8	65.8	62.1	82.1	42.6
ChatGPT	●	46	84.3	59.3	43.6	58	4.6	49.2	16.3	48.8	37	2.4	64.5	52.6	84.2	39.3
GLM4-9B	●	43.3	62.6	36.6	47	56	5.8	53.6	19.8	48	55.3	28.1	29.2	57	68.7	38.8
Qwen2.5-7B	●	41.8	84.8	47.4	39.9	56.4	3.6	52.6	31.3	49.9	29.1	8.8	36.6	53.9	56.4	35.2
DeepSeek-R1-7B	●	36.2	70.2	52.2	16.7	53.8	2.4	45.5	2	46.1	14.2	7	55.8	47.9	62.6	30.4
InternLM2.5-7B	●	34.4	0	44.4	61.2	54.9	4.2	52	31	50.5	37.5	1.3	47.3	31.4	61.6	3.9
Yi-6B	●	34.1	69.4	53.6	27.1	54.5	2.4	49.5	37.5	48.2	11.6	6.7	43	26.7	46.1	1.6
Atom-7B	◆	29.9	53.3	46.3	22.2	54.2	0.6	47.8	5.3	47.4	11.4	1.3	47.7	34.7	42	3.9
DeepSeek-llm-7B	●	29.3	42.1	52.5	22.7	54.7	0.4	49.7	2	49.4	7.6	2.7	44.9	30.1	43.3	7.8
LLaMA3-Chinese-7	◆	27.5	29.1	53.6	13.9	51.3	0.2	47.4	1.5	44.5	10.9	0.7	48	41.3	33.5	9.4

Figure E: Overview of the TaxPraBen open leaderboard and its representative interface screenshot.

Dataset	LLM Raw Response	Aligned Structured Output
TaxRisk	{"风险": "企业存在“不应享而享”的税收优惠问题，可能被追缴税款并取消高新技术企业资格。","措施": "企业需重视税务风险，确保合规经营，精准把握资格认定的关键要素，提升知识产权质量，加强研发费用核算，确保申报材料的真实性、准确性和完整性，严格执行后续管理措施。"} *{"Risk": "The enterprise may have issues of improperly claiming tax incentives to which it is not entitled, and may be required to repay taxes and have its high and new technology enterprise qualification revoked.", "Measures": "The enterprise should attach importance to tax-related risks and ensure compliant operations; accurately grasp the key elements of qualification accreditation; improve the quality of intellectual property; strengthen the accounting of R&D expenses; ensure the authenticity, accuracy, and completeness of the application materials; and strictly implement subsequent management measures."}*	{"面临的风险": "企业存在“不应享而享”的税收优惠问题，可能被追缴税款并取消高新技术企业资格。", "对应的解决措施": "企业需重视税务风险，确保合规经营，精准把握资格认定的关键要素，提升知识产权质量，加强研发费用核算，确保申报材料的真实性、准确性和完整性，严格执行后续管理措施。"} *{"Risks Encountered": "The enterprise may have issues of improperly claiming tax incentives to which it is not entitled, and may be required to repay taxes and have its high and new technology enterprise qualification revoked.", "Corresponding Mitigation Measures": "The enterprise should attach importance to tax-related risks and ensure compliant operations; accurately grasp the key elements of qualification accreditation; improve the quality of intellectual property; strengthen the accounting of R&D expenses; ensure the authenticity, accuracy, and completeness of the application materials; and strictly implement subsequent management measures."}*
TaxInspect	从案情看，公司让财务人员黄某用其私人账户代收货款，导致账外收入无法完整入账，并在申报时通过虚假申报的方式少缴税款，属于典型的隐匿收入逃避纳税义务行为。按照刑法相关规定，可能构成逃税罪。处罚一般包括补缴税款、滞纳金及罚款，情节严重还可能追究刑责。答案：犯罪行为：通过财务人员黄某的私人账户收取货款并虚假申报，隐匿销售收入逃避税款。所犯罪名：逃税罪，处罚结果：补缴税款、加收滞纳金，并处罚款共计1411.65万元。 *From the case, the company instructed its finance employee, Mr.Huang, to collect sales proceeds through his personal account. As a result, off-book income could not be fully recorded, and the company underpaid taxes by making false declarations when filing. This constitutes a typical act of concealing income to evade tax obligations. Under the relevant provisions of the Criminal Law, it may constitute the crime of tax evasion. Penalties generally include paying back the underpaid taxes, late payment surcharges, and fines; in serious cases, criminal liability may also be pursued. Answer: Criminal Act: Collecting sales payments through the personal account of the finance employee, Mr. Huang, and making false tax filings to conceal sales revenue and evade taxes. Charged Offense: Tax evasion. Penalty Outcome: Payment of back taxes, imposition of late payment surcharges, and fines totaling RMB 14.1165 million.*	{"犯罪行为": "通过财务人员黄某的私人账户收取货款并虚假申报，隐匿销售收入逃避税款。","所犯罪名": "逃税罪","处罚结果": "补缴税款、加收滞纳金，并处罚款共计1411.65万元。"} *{"Criminal Act": "Collecting sales payments through the personal account of the finance employee, Mr. Huang, and making false tax filings to conceal sales revenue and evade taxes.","Charged Offense": "Tax evasion", "Penalty Outcome": "Payment of back taxes, imposition of late payment surcharges, and fines totaling RMB 14.1165 million."}*
TaxPlan	其中包含的计算过程：0.548-0.524 = 0.024(万元/人/年)，若10000人则合计240万元。{"核心思路": "通过为全体员工统一购买符合税法规定的商业健康保险，员工缴费可税前扣除，从而减少个人所得税负担。","人均年应纳个人所得税_购买健康保险": 0.524, "人均年应纳个人所得税_未购买健康保险": 0.548, "甲公司全体员工年节税": 240, "备注": "按题设人数推算"} *The calculation involved is as follows: 0.548-0.524=0.024 (ten-thousand yuan per person per year). For 10,000 people, the total is 2.4 million RMB. {"Core Rationale": "By purchasing commercial health insurance for all employees in compliance with tax law, employees' premiums can be deducted pre-tax, thereby reducing their individual income tax burden.","Per-capita annual individual income tax payable_with health insurance": 0.524, "Per-capita annual individual income tax payable_without health insurance": 0.548, "Annual tax savings for all employees of Company A": 240, "Note": "Estimated based on the headcount given in the problem."}*	{"核心思路": "通过为全体员工统一购买符合税法规定的商业健康保险，员工缴费可税前扣除，从而减少个人所得税负担。", "人均年应纳个人所得税_购买健康保险": 0.524, "人均年应纳个人所得税_未购买健康保险": 0.548, "人均节税": 0.024, "甲公司全体员工年节税": 240} *{"Core Rationale": "By purchasing commercial health insurance for all employees in compliance with tax law, employees' premiums can be deducted pre-tax, thereby reducing their individual income tax burden.", "Per-capita annual individual income tax payable_with health insurance": 0.524, "Per-capita annual individual income tax payable_without health insurance": 0.548, "Per-capita tax savings": 0.024, "Annual tax savings for all employees of Company A": 240}*

Table F: LLM raw responses with redundant descriptions are converted into aligned structured outputs by ChatGPT.

Tax Task	Tax Dataset	Metrics	ChatGPT	GPT-4o	Mistral-V0.3	Gemma	LLaMA3	Bayling2	Grok3	DeepSeek$_{llm}$	Baichuan2	Atom	Qwen2.5	ChnLLaMA3	ERNIE-3.5	ChatGLM3	Yi	GLM4	DeepSeek$_{R1}$	InternLM2.5	YaYi2
Zero-Shot																					
KM	TaxRecite	BERTScore	0.493	0.475	0.390	0.224	0.390	0.465	0.537	0.504	0.470	0.479	0.502	0.428	**0.721**	0.493	0.489	0.481	0.454	0.518	0.491
		BARTScore	-5.166	-5.195	-5.903	-7.357	-5.657	-5.331	-4.991	-5.162	-5.306	-5.304	-5.044	-5.367	**-3.869**	-5.189	-5.245	-5.221	-5.327	-5.073	-5.220
KU	TaxSum	BERTScore	**0.624**	0.621	0.043	0.234	0.360	0.524	0.620	0.541	0.413	0.464	0.471	0.552	0.618	0.336	0.549	0.331	0.536	0.432	0.331
		BARTScore	-4.378	**-4.362**	-6.826	-7.288	-5.934	-5.003	-4.364	-4.907	-5.523	-5.380	-5.234	-4.794	-4.389	-5.965	-4.775	-5.994	-4.932	-5.444	-6.058
	TaxTopic	Accuracy	0.442	0.556	0.152	0.000	0.222	0.054	0.346	0.128	0.146	0.211	0.340	0.193	0.424	0.116	0.210	**0.630**	0.226	0.434	0.004
		F1	0.402	0.559	0.092	0.000	0.094	0.084	0.368	0.071	0.083	0.093	0.356	0.095	0.416	0.147	0.097	**0.618**	0.143	0.437	0.007
		Macro F1	0.267	0.346	0.039	0.000	0.039	0.035	0.207	0.030	0.034	0.038	0.177	0.039	0.255	0.054	0.041	**0.411**	0.057	0.254	0.005
	TaxRead	Accuracy	0.843	0.832	0.558	0.001	0.207	0.658	0.838	0.421	0.210	0.533	**0.848**	0.291	0.841	0.806	0.694	0.626	0.702	0.000	0.554
KA	TaxCalc	Accuracy	0.046	0.070	0.000	0.000	0.002	0.004	**0.190**	0.004	0.000	0.006	0.036	0.002	0.076	0.018	0.024	0.058	0.024	0.042	0.002
	TaxSCQ	Accuracy	0.436	0.556	0.079	0.129	0.231	0.147	0.541	0.259	0.323	0.264	0.407	0.226	0.589	0.283	0.307	0.460	0.230	**0.607**	0.346
		F1	0.436	0.557	0.067	0.166	0.128	0.145	0.538	0.209	0.319	0.199	0.394	0.092	0.589	0.279	0.251	0.475	0.133	**0.614**	0.345
		Macro F1	0.437	0.557	0.072	0.167	0.134	0.148	0.538	0.213	0.320	0.202	0.395	0.098	0.589	0.281	0.255	0.474	0.138	**0.614**	0.344
	TaxMCQ	Accuracy	0.163	0.290	0.003	0.000	0.000	0.013	0.225	0.020	0.060	0.053	0.313	0.015	0.255	0.030	**0.375**	0.198	0.020	0.310	0.005
	TaxQA	BERTScore	0.477	0.470	0.373	0.243	0.407	0.477	0.487	0.480	0.458	0.460	0.508	0.457	0.531	0.491	0.479	**0.538**	0.434	0.513	0.508
		BARTScore	-4.922	-4.895	-5.651	-7.287	-5.446	-5.091	-4.675	-4.860	-4.979	-5.051	-4.555	-5.098	**-4.126**	-4.926	-4.893	-4.649	-5.243	-4.721	-4.999
	TaxBoard	BERTScore	0.636	0.640	0.210	0.234	0.506	0.572	0.653	0.583	0.551	0.576	0.604	0.540	**0.688**	0.611	0.578	0.598	0.574	0.584	0.532
		BARTScore	-4.771	-4.648	-6.452	-7.289	-5.330	-4.991	-4.541	-4.882	-5.007	-4.918	-4.766	-5.135	**-4.200**	-4.862	-4.888	-4.788	-4.981	-4.854	-5.212
	TaxCrime	Accuracy	0.025	0.135	0.000	0.035	0.010	0.010	0.290	0.030	0.025	0.030	0.100	0.015	**0.385**	0.000	0.080	0.315	0.080	0.030	0.125
		F1	0.012	0.135	0.000	0.001	0.001	0.001	0.168	0.001	0.002	0.001	0.123	0.001	**0.349**	0.000	0.079	0.347	0.096	0.002	0.060
		Macro F1	0.034	0.175	0.000	0.005	0.003	0.004	**0.225**	0.050	0.005	0.007	0.041	0.005	0.188	0.000	0.042	0.181	0.035	0.007	0.032
	TaxOpinion	BERTScore	0.764	**0.807**	0.104	0.179	0.237	0.311	0.786	0.499	0.203	0.544	0.384	0.532	0.784	0.078	0.468	0.285	0.650	0.527	0.374
		BARTScore	-4.729	**-4.614**	-7.407	-7.531	-7.238	-6.911	-4.705	-6.007	-7.353	-5.900	-6.531	-5.712	-4.632	-8.019	-6.073	-7.008	-5.351	-5.813	-6.485
	TaxRisk	RiskScore	0.526	**0.663**	0.063	0.096	0.145	0.157	0.621	0.301	0.097	0.347	0.539	0.413	0.642	0.055	0.267	0.570	0.479	0.314	0.186
	TaxInspect	InspectScore	**0.842**	0.835	0.061	0.230	0.295	0.288	0.821	0.433	0.085	0.420	0.564	0.335	0.766	0.186	0.461	0.687	0.626	0.616	0.318
	TaxPlan	PlanScore	0.393	0.408	0.075	0.049	0.048	0.078	**0.426**	0.078	0.066	0.039	0.352	0.094	0.263	0.294	0.016	0.388	0.304	0.039	0.088
One-Shot																					
KM	TaxRecite	BERTScore	0.510	0.511	0.398	0.218	0.374	0.482	0.547	0.470	0.484	0.467	0.495	0.401	**0.666**	0.342	0.474	0.152	0.454	0.074	0.318
		BARTScore	-5.161	-5.133	-5.716	-7.399	-5.593	-5.308	-4.940	-5.208	-5.225	-5.325	-5.075	-5.495	**-4.281**	-5.986	-5.214	-7.014	-5.326	-7.448	-6.191
KU	TaxSum	BERTScore	0.624	**0.629**	0.339	0.549	0.426	0.471	0.623	0.576	0.562	0.573	0.264	0.587	0.617	0.200	0.589	0.254	0.534	0.206	0.066
		BARTScore	-4.385	**-4.300**	-5.579	-4.800	-5.519	-5.308	-4.348	-4.630	-4.638	-4.670	-6.406	-4.566	-4.418	-6.739	-4.512	-6.412	-4.936	-6.707	-7.483
	TaxTopic	Accuracy	0.472	**0.594**	0.168	0.004	0.124	0.008	0.394	0.208	0.149	0.185	0.280	0.111	0.551	0.182	0.176	0.370	0.322	0.017	0.001
		F1	0.456	**0.603**	0.125	0.008	0.107	0.108	0.421	0.113	0.120	0.097	0.323	0.097	0.545	0.237	0.100	0.464	0.323	0.029	0.002
		Macro F1	0.273	**0.379**	0.052	0.003	0.039	0.052	0.228	0.047	0.060	0.040	0.199	0.037	0.305	0.085	0.044	0.283	0.117	0.012	0.002
	TaxRead	Accuracy	0.840	0.833	0.833	0.639	0.018	0.525	0.831	0.785	0.815	0.656	0.780	0.732	**0.844**	0.592	0.781	0.093	0.785	0.758	0.611
KA	TaxCalc	Accuracy	0.002	0.002	0.000	0.000	0.000	0.000	0.002	0.002	0.000	0.000	0.002	0.002	**0.006**	0.000	0.000	0.000	0.000	0.000	0.000
	TaxSCQ	Accuracy	0.444	0.551	0.241	0.234	0.224	0.271	0.569	0.241	0.279	0.221	0.490	0.230	**0.590**	0.201	0.303	0.333	0.246	0.301	0.206
		F1	0.441	0.551	0.141	0.129	0.087	0.279	0.566	0.111	0.238	0.093	0.492	0.088	**0.591**	0.236	0.251	0.451	0.204	0.313	0.249
		Macro F1	0.442	0.551	0.146	0.134	0.094	0.277	0.566	0.116	0.240	0.099	0.492	0.095	**0.590**	0.236	0.254	0.452	0.207	0.315	0.249
	TaxMCQ	Accuracy	0.058	0.035	0.018	0.005	0.018	0.030	0.035	0.055	0.050	0.043	**0.060**	0.020	0.035	0.010	0.035	0.025	0.010	0.030	0.000
	TaxQA	BERTScore	0.500	0.490	0.428	0.232	0.393	0.513	0.513	0.467	0.469	0.441	**0.542**	0.426	0.526	0.508	0.465	0.492	0.450	0.538	0.454
		BARTScore	-4.778	-4.700	-5.412	-7.484	-5.501	-5.059	-4.498	-4.934	-4.992	-5.271	-4.400	-5.283	**-4.227**	-4.948	-4.924	-4.820	-5.103	-4.810	-5.563
	TaxBoard	BERTScore	0.640	0.642	0.482	0.232	0.496	0.588	0.660	0.576	0.566	0.543	0.596	0.547	**0.676**	0.606	0.578	0.470	0.583	0.517	0.498
		BARTScore	-4.682	-4.594	-5.393	-7.343	-5.334	-4.902	-4.460	-4.892	-4.966	-5.099	-4.779	-5.061	**-4.282**	-4.803	-4.840	-5.474	-4.927	-5.246	-5.418
	TaxCrime	Accuracy	0.150	0.320	0.080	0.025	0.070	0.060	0.425	0.130	0.090	0.035	0.250	0.030	**0.535**	0.085	0.195	0.180	0.140	0.035	0.105
		F1	0.180	0.373	0.097	0.002	0.080	0.089	0.408	0.153	0.101	0.002	0.275	0.021	**0.505**	0.089	0.146	0.221	0.156	0.011	0.082
		Macro F1	0.099	0.281	0.031	0.006	0.052	0.040	**0.335**	0.088	0.032	0.009	0.120	0.010	0.275	0.032	0.177	0.096	0.081	0.020	0.042
	TaxOpinion	BERTScore	0.769	**0.808**	0.363	0.299	0.221	0.183	0.792	0.690	0.678	0.643	0.506	0.674	0.793	0.063	0.513	0.088	0.670	0.101	0.403
		BARTScore	-4.710	-4.663	-6.365	-6.956	-7.328	-7.522	-4.669	-5.136	-5.224	-5.327	-6.030	-5.098	**-4.616**	-8.098	-5.863	-7.971	-5.278	-7.888	-6.462
	TaxRisk	RiskScore	0.593	**0.667**	0.184	0.280	0.215	0.358	0.621	0.477	0.424	0.460	0.543	0.516	0.658	0.041	0.412	0.255	0.529	0.081	0.226
	TaxInspect	InspectScore	**0.845**	0.843	0.379	0.530	0.345	0.477	0.793	0.487	0.398	0.451	0.577	0.420	0.796	0.140	0.438	0.372	0.637	0.068	0.136
	TaxPlan	PlanScore	0.418	**0.443**	0.068	0.097	0.045	0.234	0.471	0.052	0.065	0.059	0.409	0.141	0.408	0.176	0.106	0.310	0.298	0.031	0.049

***Note**: To ensure fair assessment across tasks and data types, we compute the overall average of all evaluation metrics from 5 runs as the results.*

Table G: The detailed zero-shot and one-shot results for 19 popular LLMs evaluated on the TaxPraBen benchmark.