SmartBench: Is Your LLM Truly a Good Chinese Smartphone Assistant?

Xudong Lu[*1,2], **Haohao Gao**[*1], **Renshou Wu**[*†1], **Shuai Ren**[1], **Xiaoxin Chen**[1], **Hongsheng Li**[2✉], **Fangyuan Li**[1✉]
[1]vivo AI Lab [2]CUHK MMLab
[*]Equal contribution [✉]Corresponding author [†]Project lead
{luxudong@link,hsli@ee}.cuhk.edu.hk {gaohaohao,wurenshou,lifangyuan}@vivo.com

Abstract

Large Language Models (LLMs) have become integral to daily life, especially advancing as intelligent assistants through on-device deployment on smartphones. However, existing LLM evaluation benchmarks predominantly focus on objective tasks like mathematics and coding in English, which do not necessarily reflect the practical use cases of on-device LLMs in real-world mobile scenarios, especially for Chinese users. To address these gaps, we introduce **SmartBench**, the first benchmark designed to evaluate the capabilities of on-device LLMs in Chinese mobile contexts. We analyze functionalities provided by representative smartphone manufacturers and divide them into five categories: text summarization, text Q&A, information extraction, content creation, and notification management, further detailed into 20 specific tasks. For each task, we construct high-quality datasets comprising 50 to 200 question-answer pairs that reflect everyday mobile interactions, and we develop automated evaluation criteria tailored for these tasks. We conduct comprehensive evaluations of on-device LLMs and MLLMs using SmartBench and also assess their performance after quantized deployment on real smartphone NPUs. Our contributions provide a standardized framework for evaluating on-device LLMs in Chinese, promoting further development and optimization in this critical area. Code and data will be available at https://github.com/vivo-ai-lab/SmartBench.

1 Introduction

Large Language Models (LLMs) have significantly transformed everyday life in recent years by serving as intelligent, context-aware assistants (OpenAI., 2024; Team et al., 2024; Anthropic, 2023; Anil et al., 2023; Lu et al., 2024a; Jiang et al., 2024; Abdin et al., 2024; Guo et al., 2025). To further enhance the capabilities of LLMs in serving human needs, various academic research and engineering efforts have focused on deploying smaller LLMs on edge devices, such as smartphones (Xue et al., 2024; Yao et al., 2024; Chu et al., 2023, 2024; Lu et al., 2024b). As companions in our daily lives, smartphones serve as crucial platforms for people to experience the capabilities of on-device LLMs. The local deployment of LLMs on end-side smartphones eliminates the need for a network connection, which not only broadens the scope of possible application scenarios but also enhances user privacy by keeping sensitive data processing on the device (Qu et al., 2024; Ding et al., 2024).

The current trend in smartphone technology shows that major manufacturers are increasingly adopting on-device LLMs (Ashkboos et al., 2024), integrating advanced AI capabilities into their devices. Industry leaders such as Apple with OpenELM (Mehta et al., 2024), HUAWEI's Pangu E (Zeng et al., 2021), Xiaomi's MiLM (XiaomiTime, 2024), and vivo's BlueLM-3B (Lu et al., 2024b) have demonstrated significant progress in this domain. These on-device LLMs support various real-time tasks (Wu et al., 2024), offering users seamless and responsive AI-powered mobile interactions (Xu et al., 2024).

However, we find that there are still notable gaps in the comprehensive evaluations for assessing the capabilities of on-device LLMs deployed on smartphones. Traditional LLM evaluations are typically categorized into two dimensions, i.e., ***objective*** tasks and ***subjective*** tasks. Objective tasks primarily focus on the assessment of knowledge, encompassing areas such as mathematical proficiency with benchmarks like GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), coding competence evaluated through HumanEval (Chen et al., 2021), and multitask accuracy measured by MMLU (Hendrycks et al., 2020). Subjective tasks typically evaluate the model's ability to generate coherent, contextually appropriate, and human-like responses. These tasks often consider the model's

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 3907–3932
November 4-9, 2025

Text Summarization				Text Q&A			Information Extraction		
Document Summ 文档摘要	Call Summ 通话摘要	Recording Summ 录音摘要	Meeting Summ 会议摘要	Document Q&A 文档问答	Retrieval Q&A 检索问答	Personal Q&A 个人问答	Entity Extraction 实体抽取	Relation Extraction 关系抽取	Event Extraction 事件抽取
Content Creation								**Notification Management**	
Text Polishing 文本润色	Text Continuation 文本续写	Text Abbreviation 文本缩写	Text Expansion 文本扩写	Text Creation 文本创作	Text Formatting 文本排版	Instant Reply 即时回复	Text Correction 文本纠错	Notification Sorting 通知排序	Message Summ 消息总结

Table 1: We analyze the on-device LLM features currently released on mobile phones by major manufacturers, dividing them into 5 major categories with 20 tasks. Based on this, we propose SmartBench, the first (Chinese) benchmark for assessing the capabilities of on-device LLMs in mobile scenarios.

creativity, fluency, adaptability to nuanced instructions, and alignment with user intent. Subjective evaluation datasets are often derived from user-constructed scenarios (Liu et al., 2023), curated human-chatbot conversations (Lin et al., 2024), and filtered interactions from platforms like Chatbot Arena (Li et al., 2024a,b). For on-device smartphone applications, the evaluation predominantly emphasizes subjective capabilities. Through our investigation, we identify the following critical gaps in existing subjective evaluation benchmarks:

1) The scenario gap: Current benchmarks emphasize tasks like mathematics and coding (Li et al., 2023; Lin et al., 2024; Li et al., 2024b), which are rarely handled by on-device LLMs in practical applications. Instead, on-device LLMs place greater emphasis on lightweight tasks such as text refinement, and notification processing.

2) The language gap: Mobile users who speak different languages often have varying living environments and language habits. Currently, most evaluation protocols for subjective tasks are all in English. As a market with over 1 billion smartphone users (Statista, 2025), it is crucial to have an evaluation benchmark for LLMs deployed on Chinese-oriented smartphones.

To tackle these gaps, in this paper, starting from a functional investigation of on-device LLMs, we construct SmartBench, the first (Chinese) benchmark for evaluating the capabilities of on-device LLMs in mobile scenarios. Specifically, we analyze the on-device LLM functionalities provided by Apple, HUAWEI, OPPO, vivo, Xiaomi, and HONOR (up to December 2024), dividing them into five categories: text summarization, text Q&A, information extraction, content creation, and notification management. Building on these functionalities, we further refine the 5 categories into 20 tasks, as outlined in Tab. 1. To evaluate each task, we construct 50 to 200 question-answer (QA) pairs per task that reflect everyday life scenarios by screening open-source datasets and generating additional pairs using manual collection or LLMs, resulting in a total of 2973 QA pairs. Evaluations of subjective tasks are commonly conducted using the LLM-as-a-Judge paradigm (Zheng et al., 2023). In SmartBench, we develop detailed automated evaluation criteria for each category/task. We further conduct comprehensive evaluations of multiple on-device LLMs and MLLMs on SmartBench and assess their performance after quantized deployment on the NPU of real smartphones.

Our contributions are summarized as follows:

1) We investigate the on-device LLM features offered by representative smartphone manufacturers, organizing them into 5 categories comprising 20 tasks. We then introduce SmartBench, the first Chinese benchmark designed to evaluate the capabilities of on-device LLMs in mobile scenarios, featuring 2973 QA pairs.

2) For each task, we construct high-quality text QA pairs tailored to mobile usage scenarios by screening open-source datasets, manually collecting data, and synthesizing data using LLMs. Additionally, we develop high-quality automated evaluation methods for each category/task.

3) We evaluate the performance of representative end-side LLMs/MLLMs using SmartBench. Additionally, we assess the accuracy of quantized models running on real smartphone NPUs, which offers greater practical value.

2 Related Works

2.1 Large Language Models on Edge Devices

The deployment of LLMs on edge devices has garnered significant attention in recent years. In the academic community, there are currently numerous open-source LLMs and MLLMs, such as Qwen2.5 3B (Yang et al., 2024b), InternVL 2.5 4B (Chen et al., 2024), and MiniCPM 3.0 4B (Hu et al., 2024a). Most of these models have between

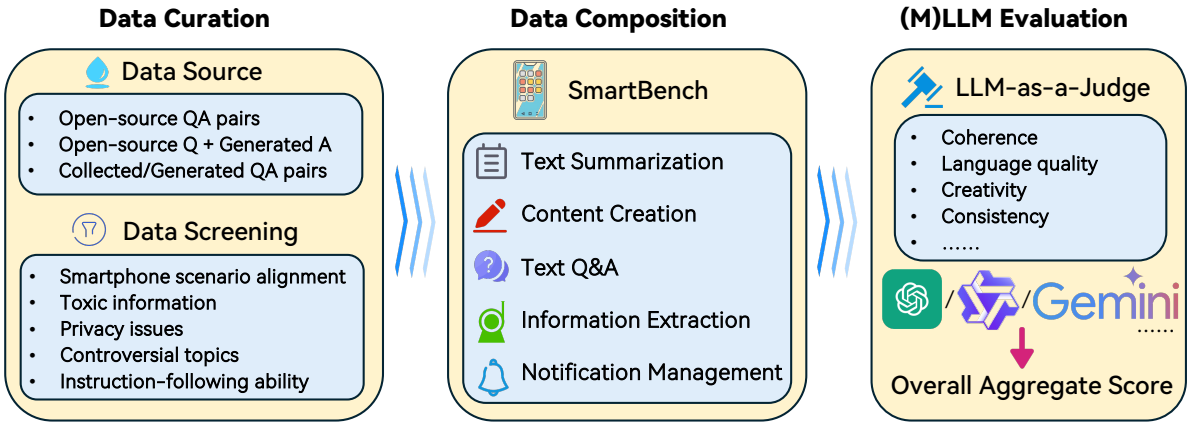


Figure 1: Overview of SmartBench, including data curation, data composition, and LLM-as-a-Judge evaluation.

3B and 4B parameters, making them well-suited for deployment on edge devices with limited computational capabilities. Besides, major smartphone manufacturers have also introduced their own LLMs, including Gemini Nano by Google, BlueLM by vivo, Magic LM by HONOR, OpenELM by Apple, and MiLM by Xiaomi (Wu et al., 2024). These advancements pave the way for more efficient and powerful AI applications on edge devices.

2.2 Benchmarks for Realworld Assistance

How to comprehensively evaluate LLMs has long been a widely researched topic (Chang et al., 2024). The vast majority of benchmarks are designed to assess the knowledge capabilities of these models, including general knowledge (Hendrycks et al., 2020; Wang et al., 2024c; Clark et al., 2018), mathematics and science knowledge (Cobbe et al., 2021; Hendrycks et al., 2021; Rein et al., 2023), and coding ability (Austin et al., 2021; Chen et al., 2021), etc. Recently, there have been new datasets introduced to test the ability of models to handle real users' questions in the wild (Liu et al., 2023; Lin et al., 2024). These datasets often consist of subjective questions that focus on the creativity and ability of models to follow instructions in real-world usage scenarios (Li et al., 2024b,a, 2023), providing a more direct reflection of user comfort and satisfaction during real-world usage. SmartBench is the first benchmark designed to evaluate the practical functionalities of LLMs deployed on smartphones.

2.3 Chinese LLM Benchmarks

With the rapid development of Chinese LLMs (Sun et al., 2021; Team, 2023; Guo et al., 2025), specialized evaluation benchmarks have been established to assess their performance in understanding and generating content within a Chinese context. Prominent Chinese LLM benchmarks include CMRC (Cui et al., 2019), CLUE (Xu et al., 2020), SuperCLUE (Xu et al., 2023), and C-Eval (Huang et al., 2023), etc. Additionally, there are datasets like AlignBench (Liu et al., 2023) designed for evaluating subjective tasks in Chinese. However, SmartBench distinguishes itself by focusing specifically on everyday mobile scenarios, offering a unique perspective on the practical functionalities of on-device LLMs in real-life smartphone usage.

2.4 LLM Agent on Smartphones

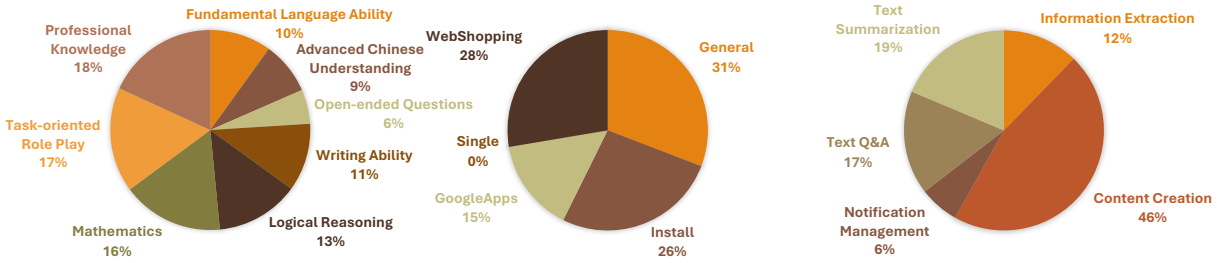
There is another type of task on mobile phones that helps solve real-world tasks, called mobile agents (Wang et al., 2024a; Zhang et al., 2023a; Chai et al., 2024; Rawles et al., 2024). These tasks often involve executing multi-step commands on the phone based on user instructions (Zhang et al., 2024). In contrast, Smartbench focuses on the functionality of on-device LLMs for handling common daily tasks in a single step, without planning action trajectories or calling external APIs.

3 SmartBench

In this section, we present a detailed description of the proposed SmartBench benchmark, specifically focusing on the scenario of smartphone deployment. We cover the data composition (Sec. 3.1), data sources (Sec. 3.2), filtering criteria (Sec. 3.3), and evaluation protocol (Sec. 3.4) used in the construction of the benchmark. The overview of SmartBench is illustrated in Fig. 1.

3.1 Data Composition

We divide the on-device LLM features released by representative smartphone manufacturers into 5 categories, encompassing a total of 20 tasks.



(A) AlignBench (B) AitZ (test) (C) SmartBench

Figure 2: Data composition comparison between AlignBench (Liu et al., 2023), AitZ (Zhang et al., 2024) and SmartBench. AlignBench (zh) is a general benchmark designed for Chinese scenarios, and AitZ (en) is a mobile agent benchmark. SmartBench (zh) is specifically designed for evaluating end-side LLM functionality on smartphones.

1) Text Summarization: This category is focused on providing a concise summary of the text in one sentence and listing key information in bullet points. The benefit of this function is that it allows users to quickly grasp the main ideas and essential details without needing to read through lengthy content. We categorize the content into four scenarios. Document summarization primarily targets emails, scientific knowledge, and news reports. Call summarization focuses on conversations between two people. Recording summarization focuses on recordings that have significant background noise. Meeting summarization specifically refers to the summarization of meetings.

2) Content Creation: This category highlights the functionality of creating content on mobile devices, enabling users to effortlessly share their creations on social media platforms such as Weibo, WeChat Moments, and RedNote (Xiaohongshu). With the widespread use of smartphones, mobile content creation has become increasingly accessible and convenient. We focus on the commonly utilized functions for content creation, i.e., polishing, continuation, abbreviation, expansion, (automatic) creation, formatting, and correction. Additionally, on-device LLMs are employed to refine users' message replies; therefore, we also incorporate tests for the instant reply functionality.

3) Text Q&A: This feature allows users to quickly obtain information or answer questions through simple text inputs. We categorize it into three scenarios: Document Q&A, where a specific document is provided and questions are answered based on it; Retrieval Q&A, where answers are summarized based on multiple relevant retrieval contents and questions; and Personal Q&A, where information from synthesized personal records (such as memos or personal notes) is used.

4) Information Extraction: This category involves automatically identifying and extracting specific data from text inputs, such as names, dates, addresses, or other relevant information. The information extraction functionality on mobile phones is primarily divided into three aspects: Entity Extraction, which involves identifying and extracting specific entities from text, such as names, locations, dates, etc.; Relation Extraction, which analyzes and extracts relationships between entities, such as "someone works at a certain company"; and Event Extraction, which identifies specific events and their related elements from text, such as time, location, and participants. These functionalities collectively contribute to intelligent applications, such as automatic summarization, smart search, and personalized recommendations.

5) Notification Management: Effective notification management on smartphones is essential to

Category	Count	Input	Target
Text Summarization	550	1890	244
Content Creation	1377	210	143
Text Q&A	495	930	115
Information Extraction	362	682	74
Notification Management	189	376	101

Table 2: We present the number of QA pairs for each category in SmartBench. For each category, we also provide the average input (query) length and the average target (reference answer) length of all QA pairs.

Message Summarization Query
月亮上的海：我家布偶超级爱掉毛，尤其是换季的时候，简直就是行走的蒲公英！😭 你们有什么好办法吗？在线等，挺急的！

月亮上的海：试过好多种猫粮了，感觉效果都不太明显，每天都得吸好多毛，心累... 😩

Reference
月亮上的海：
很发愁布偶猫掉毛严重的问题，想寻求解决办法。

Figure 3: Example of the Message Summarization task in SmartBench (English translated version in Fig. 16).

评价维度：

1. 连贯性：检查续写部分是否自然地与前文衔接，保持一致的主题和情境，避免了突然转折或引入不相关信息。
2. 一致性：评估续写是否符合前文设定的风格、语气及人物特征，确保整个故事或论述的声音统一。
3. 创造性：在维持连贯性和一致性的基础上，考察续写是否展示了新颖的观点或是有趣的情节发展，而非仅仅是简单重复已有的信息。
4. 语言质量：分析续写内容的语言表达是否清晰流畅，无语法、拼写或标点错误，使用了丰富的词汇以及良好的句子结构来提高文章的可读性和吸引力。

评分标准：

1. 将AI助手生成的答案与参考答案进行比较，指出AI助手生成的答案有哪些不足，并进一步解释。
2. 从不同维度对AI助手生成的答案进行评价，在每个维度的评价之后，给每一个维度一个1～10的分数。
3. 最后，综合每个维度的评估，对AI助手生成的答案给出一个1～10的综合分数。
4. 你的打分需要尽可能严格，并且要遵守下面的评分规则：总的来说，AI助手的答案质量越高，则分数越高。
   - ✓当AI助手的答案存在明显的逻辑断裂、严重偏离主题或包含大量无关信息时，总分必须是1到2分；
   - ✓当AI助手的答案虽然没有严重偏离主题但质量较低，未能很好地延续前文的风格或情节时，总分为3到4分；
   - ✓当AI助手的答案基本满足了连贯性和一致性要求，但在创造性和/或语言质量上表现较差，总分可以得5到6分；
   - ✓当AI助手的答案质量与参考答案相近，在所有维度上表现良好，总分得7到8分；
   - ✓只有当AI助手的答案质量显著超过参考答案，不仅完美地延续了前文，而且在创造性和语言质量方面表现出色的情况下，才能得9到10分。

Figure 4: Evaluation Dimension & Scoring Standard for the text continuation task (English version in Fig. 17).

minimize distractions, enhance productivity, and ensure timely access to important information. Currently, LLMs deployed on smartphones primarily support two functions: Notification Sorting, which organizes and prioritizes notifications based on degree of urgency or chronological order; and Message Summarization, which condenses lengthy notifications or messages into concise summaries for quick understanding. By intelligently sorting and summarizing information, smartphones equipped with such features can significantly improve efficiency and reduce cognitive overload in our increasingly connected world.

To facilitate a clear comparison between SmartBench and other LLM benchmarks, we analyze their data composition. For the Chinese benchmark, we compare with AlignBench (Liu et al., 2023), while we select AitZ (Zhang et al., 2024) for the mobile agent benchmark, as illustrated in Fig. 2. As shown, AlignBench serves as a more general benchmark for evaluating Chinese LLMs, AitZ focuses more on automated operations on mobile devices, while SmartBench emphasizes common on-device LLM functionalities. Additionally, we provide the number of QA pairs for each category in SmartBench in Tab. 2, along with the average input (query) length and the average target (reference answer) length for each category. Furthermore, to better illustrate the essence of SmartBench for mobile scenarios, we offer an example of the Message Summarization task in Fig. 3.

3.2 Data Source

The data for SmartBench is primarily derived from three sources. 1) We screen open-source datasets to select QA pairs that align with smartphone application scenarios. 2) For datasets that provide contextual information but lack appropriate questions and answers, we utilize advanced LLMs, e.g., Qwen-Max (Yang et al., 2024a), Gemini Pro (Reid et al., 2024), to generate corresponding answers for each task. 3) For the lack of open-source data in certain categories, we employ human collection and LLMs to generate QA pairs, followed by manual screening and editing to curate high-quality data.

For Text Summarization, we primarily use content from open-source datasets. For document data, we utilize the dataset from (Xu, 2019), which comprises a substantial Chinese corpus including content from Wikipedia, news reports, etc. For summarizing calls, recordings, and meetings, we draw data from Alimeeting4MUG (Zhang et al., 2023b), LCCC (Wang et al., 2020), VCSum (Wu et al., 2023), WenetSpeech (Zhang et al., 2022), etc. Speech content is converted to text transcriptions using speech-to-text converters in our benchmark, and the reference summaries for the summarization tasks are generated by Qwen-Max.

For Content Creation, we leverage QA pairs from CSCD-NS (Hu et al., 2024b) for text correction. For other tasks, e.g., polishing, abbreviation, expansion, etc, we manually collect and design examples, and then use Gemini Pro and Qwen-Max to generate QA pairs tailored to meet the require-

Category	Task	GPT-4o	BlueLM-3B	InternVL2.5-4B	MiniCPM3-4B	Qwen2.5-3B	Qwen2-VL-2B
Text Summarization	Document Summ	7.05	**7.56**	6.89	7.40	7.21	4.37
	Call Summ	7.03	**7.22**	5.43	6.88	6.35	3.48
	Recording Summ	7.78	**7.63**	6.38	7.45	7.07	4.17
	Meeting Summ	7.73	**7.09**	6.23	6.98	6.67	3.75
Text Q&A	Document Q&A	8.83	**9.37**	9.36	8.39	9.34	9.15
	Retrieval Q&A	6.91	5.89	5.81	**6.76**	6.25	4.77
	Personal Q&A	9.78	9.36	8.89	8.87	**9.39**	8.83
Content Creation	Text Polishing	7.49	**7.55**	6.17	7.53	7.42	6.19
	Text Continuation	7.63	7.45	6.89	7.52	**7.72**	5.96
	Text Abbreviation	7.49	8.23	7.43	8.17	**8.51**	7.51
	Text Expansion	8.42	7.44	8.04	**8.74**	8.07	6.04
	Text Creation	7.55	**6.93**	6.16	6.89	6.68	5.26
	Text Formatting	6.15	6.03	5.10	**6.80**	3.69	1.20
	Instant Reply	7.62	**6.70**	5.90	6.28	6.44	3.14
	Text Correction	7.03	**3.69**	2.46	3.24	2.38	1.17
Information Extraction	Entity Extraction	8.36	7.82	**8.13**	7.58	6.35	5.00
	Relation Extraction	3.54	**5.55**	3.58	4.15	3.54	3.04
	Event Extraction	7.86	6.79	**7.09**	6.20	6.75	4.66
Notification Management	Message Summ	7.24	7.45	7.29	**8.08**	7.86	5.90
	Notification Sorting	5.97	4.78	4.19	**4.85**	4.51	2.14
AVG		7.37	**7.03**	6.37	6.94	6.61	4.79

Table 3: Evaluation results using GPT-4 Turbo (`gpt-4-turbo-04-09`) as the judge LLM. We compare BlueLM-3B, InternVL2.5-4B, MiniCPM3-4B, Qwen2.5-3B, and Qwen2-VL-2B (on-device models) on the whole SmartBench in BF16 precision with GPT-4o (on-cloud model) for reference. The scores assessed by Qwen-Max as the judge LLM are also provided in Tab. 8 in the Appendix.

ments of daily mobile usage scenarios.

For text Q&A, we select document Q&A pairs from the CMRC (Cui et al., 2019) dataset. For retrieval-based Q&A, the textual sources are from DuReader 2.0 (He et al., 2017), and the answers are generated by Qwen-Max. For personal Q&A, we design human examples and construct QA pairs (e.g., memos or personal notes) using Qwen-Max.

For Information Extraction, we source textual data for entity extraction from MSRA (Levow, 2006), OntoNotes Release 4.0 (Weischedel et al., 2011), and Weibo (Peng and Dredze, 2016). We use Qwen-Max to generate the corresponding answers. For relation and event extraction, we manually collect example data and generate textual information using Gemini Pro, then produce the corresponding answers with GPT-4 Turbo.

For Notification Management, we find that there is currently no suitable open-source data available for the smartphone platform. Therefore, we create human-designed examples and then use Gemini Pro to generate QA pairs for both notification sorting and message summarization.

3.3 Data Screening

After initially collecting all the data for each task, we implement a rigorous screening process involving six domain experts with over five years of mobile AI experience. These specialists evaluate the dataset through dual-layer verification, primarily focusing on five core criteria: alignment with real-world smartphone interaction scenarios, detection of toxic or harmful information, identification of potential privacy leakage risks, flagging of socially controversial or polarizing topics, and comprehensive assessment of instruction-following capabilities of the reference answers.

To be specific, we implement a five-point scoring system (1–5) for each criterion, where 1 = Unacceptable, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Very Good. Human experts score each item across all criteria, and we retain only those with an average score of ≥ 3.5, reducing the original 30k dataset to roughly 3k high-quality entries. For items scoring between 3.5 and 4, we perform manual re-labeling to guarantee accuracy and alignment with human judgment. All items are further refined and subjected to dual-layer verification, ensuring rigorous quality control throughout the dataset.

3.4 Evaluation Protocol

Since subjective questions often lack an absolutely correct answer and involve multifaceted scoring dimensions, current subjective question evaluation

Category	Task	BlueLM-3B			Qwen2.5-3B		
Precision		BF16	INT4	Retention (%)	BF16	INT4	Retention (%)
Text Summarization	Document Summ	7.22	4.98	68.92	6.89	4.44	64.52
	Call Summ	7.00	6.77	96.73	6.86	6.29	91.67
	Recording Summ	7.15	6.53	91.24	6.94	5.63	81.12
	Meeting Summ	6.85	5.25	76.65	6.67	4.84	72.61
Text Q&A	Document Q&A	9.77	9.54	97.64	9.77	9.38	96.06
	Retrieval Q&A	6.13	5.38	87.76	6.38	5.88	92.16
	Personal Q&A	8.71	7.58	87.04	9.29	9.15	98.54
Content Creation	Text Polishing	7.57	7.18	94.81	7.54	7.07	93.84
	Text Continuation	7.50	7.13	95.06	7.70	7.27	94.37
	Text Abbreviation	7.81	7.06	90.50	8.23	7.27	88.33
	Text Expansion	8.18	8.12	99.28	8.47	8.41	99.31
	Text Creation	6.82	6.55	96.16	6.50	6.42	98.82
	Text Formatting	6.10	5.67	92.97	4.33	3.99	91.97
	Instant Reply	6.55	6.30	96.18	6.20	5.94	95.84
	Text Correction	2.83	2.24	78.89	1.67	1.17	70.00
Information Extraction	Entity Extraction	7.15	7.05	98.49	6.13	6.08	99.06
	Relation Extraction	5.73	5.01	87.48	4.64	3.62	78.06
	Event Extraction	7.00	6.06	86.68	7.06	6.05	85.62
Notification Management	Message Summ	7.92	7.80	98.48	8.00	7.88	98.50
	Notification Sorting	5.13	4.83	94.14	4.90	4.74	96.71
AVG		6.96	6.35	91.31	6.71	6.08	90.58

Table 4: W4A16 evaluation results with 50 questions per task using GPT-4 Turbo as the judge LLM. We deploy BlueLM-3B and Qwen2.5-3B on the NPU of the vivo iQOO 12 smartphone, which is equipped with the Snapdragon 8 Gen 3 SoC. The quantized models are able to maintain an average performance of around 90%.

datasets always adopt the "LLM-as-a-Judge" approach for assessment (Liu et al., 2023; Zheng et al., 2023; Li et al., 2024b). In SmartBench, we meticulously design different LLM evaluation prompts for each function category. For Content Creation, Information Extraction, and Notification Management, we especially design distinct scoring prompts for each task. This targeted design makes the scoring more aligned with human perceptions.

In SmartBench, each question is assigned a total score of 10 points. For the evaluation prompt of each task, in addition to providing reference answers for each question, we also include detailed scoring guidelines. We first outline the scoring dimensions; for example, in the text continuation task (as in Fig. 4), we assess the answer's coherence, language quality, creativity, and consistency with the original text. Next, we develop comprehensive scoring standards for each dimension to ensure accurate and consistent grading. The judge LLM first assigns separate scores for each dimension and then provides an overall aggregate score. Especially, for the Text Correction task, which has clearly defined correction answers, the evaluation criterion focuses on the accuracy of the modifications made.

4 Experiment

In this section, a series of experiments are conducted. We evaluate the performance of representative on-device LLMs and MLLMs on SmartBench (Sec. 4.1) and conduct human tests to assess the effectiveness of the LLM-as-a-Judge evaluation method (Sec. 4.3). To better align with practical on-device deployment, we also analyze the model performance after quantized inference on the NPU in actual smartphones (Sec. 4.2).

4.1 BF16 Precision Evaluation

In this subsection, we evaluate representative on-device LLMs/MLLMs on SmartBench (BF16 parameter precision). We select BlueLM-3B (Lu et al., 2024b), InternVL2.5-4B (Chen et al., 2024), MiniCPM3-4B (Hu et al., 2024a), Qwen2.5-3B (Yang et al., 2024b), and Qwen2-VL-2B (Wang et al., 2024b). GPT-4 Turbo (`gpt-4-turbo-04-09`) is utilized as the judge LLM. We also provide the scores of GPT-4o to compare on-cloud models with on-device models. The results are summarized in

Tab. 3, where BlueLM-3B achieves the highest average score among on-device models. Additionally, we can observe the following trends from the table:

1) For common text-based tasks on mobile devices, such as summarization and question-answering, existing on-device models have shown satisfactory performance. However, when dealing with tasks that require more rigorous logical reasoning, such as Text Correction, Relation Extraction, and Notification Sorting, the performance of on-device models still lags behind. We provide several examples in the Appendix. For instance, Fig. 10 demonstrates that all models struggle to identify subtle typos within sentences.

2) Integrating multimodal capabilities into MLLMs might result in a reduction of pure language performance. Specifically, the InternVL2.5-4B model is developed based on Qwen2.5-3B. While InternVL2.5-4B successfully acquires multimodal functionalities, this enhancement leads to a partial decline in its pure language performance.

3) On-device models still exhibit notable performance gaps compared to the on-cloud model, particularly on the Text Correction task, where they achieve only about half the score of GPT-4o. This suggests that enhancing the reasoning capability of on-device models remains an important area.

4) For a more comprehensive evaluation, we also present the scores assessed by Qwen-Max (`qwen-max-longcontext`) as the judging LLM in Tab. 8 in the Appendix. It can be observed that although there are slight differences in the average scores, both Qwen-Max and GPT-4 Turbo rank the models in the same order. This demonstrates the robustness of our LLM-as-a-Judge approach.

Remark: We evaluate MLLMs on SmartBench because in on-device deployment scenarios on real smartphones, memory limitations often prevent us from deploying both an LLM and an MLLM on the device. Consequently, this on-device model must simultaneously handle both pure language tasks and multimodal tasks effectively.

4.2 INT4 Precision Evaluation on NPU

On-device LLMs are often deployed on the smartphone's Neural Processing Unit (NPU) to leverage its specialized parallel computational capabilities. In our experiment, we deploy the BlueLM-3B and Qwen2.5-3B models on the NPU of the vivo iQOO 12 smartphone equipped with the Snapdragon 8 Gen 3 SoC. To be specific, we quantize the models to W4A16 using the Qualcomm QNN SDK. Due to the inference speed limitations on the mobile NPU, we select 50 questions per task for inference. The results are shown in Tab. 4. We present the scores for each task (BF16 and INT4) and the capability retention of the INT4 models. Additionally, we provide the evaluation results using Qwen-Max as the judge LLM in Tab. 9 in the Appendix.

1) For most tasks, the quantized models retain over 80% of their original capabilities, with an overall average retention rate of approximately 90%.

2) Although the models can achieve, on average, 90% of the original score on the edge side, they may still generate incorrect responses after quantization. We provide two failure cases of BlueLM-3B in Sec. A.5, where the model exhibits fluency degradation and reduced understanding capability.

3) To offer deeper insights into real-world edge-side deployment, we report the prefilling speed, output token generation speed, and power consumption on the iQOO 12 smartphone using the Qualcomm QNN SDK, as shown in Tab. 5.

4.3 Human Test

We use the LLM-as-a-Judge method to assess different on-device models. Therefore, it is important to examine the consistency between the scores given by the judge LLM and those given by humans. We carry out a human test with six human experts in this subsection.

During the auto-evaluation process, the judge LLM assigns a score between 0 and 10 to the output of each model response. Considering that humans might find it challenging to directly score subjective questions, especially tasks like text polishing, we ask human experts to rank the outputs generated by different on-device models (i.e., BlueLM-3B, InternVL2.5-4B, MiniCPM3-4B, Qwen2.5-3B, and Qwen2-VL-2B) for each question. We then use the scores from the judge LLM (Qwen-Max in our setting) to compute model rankings for each question. Finally, we calculate the Pearson correlation between the rankings from the judge LLM and those provided by human experts.

In SmartBench, we meticulously design evaluation dimensions and scoring standards for each task/category. To establish a baseline, we compare our evaluation prompts with those used in MT-Bench. We randomly select 50 questions for each task, with each question containing responses from 5 on-device models. This results in a total of $50 \times 20 \times 5 = 5000$ samples. We conduct human ranking and calculate the Pearson correlation with

	#Params	Context Length	Prefilling Speed (token/s)	Output Speed (token/s)	Power (W)
BlueLM-3B	2.7B	2048	930.9	27.1	6.4
Qwen2.5-3B	3.1B	2048	873.4	24.9	6.8

Table 5: Inference speed and power usage of BlueLM-3B and Qwen2.5-3B on iQOO 12 with Qualcomm QNN SDK. Due to its larger parameter size, Qwen2.5-3B exhibits slower inference speed and higher power consumption.

	Text Summarization	Text Q&A	Content Creation	Information Extraction	Notification Management	AVG
MT-Bench	0.8412	0.8025	0.6998	0.7894	0.8467	0.7959
SmartBench	0.8823	0.8151	0.7289	0.8396	0.8742	0.8280

Table 6: We compare our LLM-as-a-Judge evaluation method with MT-Bench's evaluation method using the Pearson correlation score with human rankings. Our evaluation method demonstrates higher consistency with humans.

the judge LLM ranking (our prompt versus MT-Bench prompt), and the results are shown in Tab. 6. Our designed prompt excels in all categories.

5 Conclusion

In this paper, we present SmartBench, the first benchmark designed to evaluate the capabilities of on-device LLMs in Chinese mobile contexts. By analyzing functionalities offered by leading smartphone manufacturers, we create a standardized framework divided into five key categories and 20 specific tasks, complete with high-quality datasets and tailored evaluation criteria. Our comprehensive evaluations of on-device LLMs and MLLMs using SmartBench highlight the strengths and weaknesses of current models in real-world mobile scenarios. This work fills a critical gap in benchmarking tools for Chinese users, promoting further development and optimization of on-device LLMs in practical mobile applications.

Limitations

In this paper, we provide SmartBench, the first benchmark designed to evaluate the capabilities of on-device LLMs in Chinese mobile contexts. Our work still has some limitations: 1) With the advancement of technology, the functions of on-device LLMs will continually evolve. Our investigation only covers up to December 2024. We will continue to update the dataset in line with the release of new features. 2) We have developed SmartBench specifically for the usage scenarios of Chinese users. The usage habits and methods of smartphone users may vary significantly across different countries. Moving forward, we will continue to support multiple languages. 3) The current benchmark focuses on the text modality, whereas mobile applications may also involve vision and audio modalities (e.g., camera input and voice recognition). We plan to incorporate these additional modalities in future versions.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

Anthropic. 2023. Claude 3. https://www.anthropic.com. Large Language Model.

Saleh Ashkboos, Iman Mirzadeh, Keivan Alizadeh, Mohammad Hossein Sekhavat, Moin Nabi, Mehrdad Farajtabar, and Fartash Faghri. 2024. Computational bottlenecks of training small-scale large language models. *arXiv preprint arXiv:2410.19456*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. 2024. Amex: Android multi-annotation expo dataset for mobile gui agents. *arXiv preprint arXiv:2407.17490*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and 1 others. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and 1 others. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

Yucheng Ding, Chaoyue Niu, Fan Wu, Shaojie Tang, Chengfei Lyu, and Guihai Chen. 2024. Enhancing on-device llm inference with historical cloud-based llm interactions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 597–608.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, and 1 others. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024a. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Yong Hu, Fandong Meng, and Jie Zhou. 2024b. Cscdns: a chinese spelling check dataset for native speakers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–159.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024a. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024b. From live data to high-quality benchmarks: The arena-hard pipeline.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, and 1 others. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024a. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, and 1 others. 2024b. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Seyed Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and 1 others. 2024. Openelm: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

OpenAI. 2024. Hello GPT-4o.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 149–155.

Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. 2024. Mobile edge intelligence for large language models: A contemporary survey. *arXiv preprint arXiv:2407.18921*.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, and 1 others. 2024. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Statista. 2025. Number of smartphone users in china from 2018 to 2022 with a forecast until 2027. https://www.statista.com/statistics/467160/forecast-of-smartphone-users-in-china/.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, and 1 others. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024a. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and 1 others. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

Han Wu, Mingjie Zhan, Haochen Tan, Zhaohui Hou, Ding Liang, and Linqi Song. 2023. Vcsum: A versatile chinese meeting summarization dataset. *arXiv preprint arXiv:2305.05280*.

Liangxuan Wu, Yanjie Zhao, Chao Wang, Tianming Liu, and Haoyu Wang. 2024. A first look at llm-powered smartphones. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering Workshops*, pages 208–217.

XiaomiTime. 2024. Xiaomi milm2: How xiaomi's giant language model evolves to a super intelligent ecosystem by itself.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, and 1 others. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

Zhenliang Xue, Yixin Song, Zeyu Mi, Le Chen, Yubin Xia, and Haibo Chen. 2024. Powerinfer-2: Fast large language model inference on a smartphone. *arXiv preprint arXiv:2406.06282*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, and 1 others. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023a. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. 2024. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*.

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023b. Mug: A general meeting understanding and generation benchmark. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Appendix

A.1 Data License

Dataset	Source	License
nlp_chinese_corpus	https://github.com/brightmart/nlp_chinese_corpus	MIT License
WenetSpeech	https://wenet.org.cn/WenetSpeech/	CC BY 4.0
LCCC	https://github.com/thu-coai/CDial-GPT	MIT License
Alimeeting4MUG	https://modelscope.cn/datasets/modelscope/Alimeeting4MUG/	CC BY 4.0
VCSum	https://github.com/hahahawu/VCSum	MIT License
CMRC 2018	https://ymcui.com/cmrc2018/	CC BY-SA 4.0
DuReader-2.0	https://github.com/baidu/DuReader/tree/master/DuReader-2.0	Apache License 2.0
Weibo	https://github.com/hltcoe/golden-horse	CC BY-SA 3.0
MSRA	https://tianchi.aliyun.com/dataset/144307	CC BY 4.0
OntoNotes Release 4.0	https://www.modelscope.cn/datasets/yingxi/cross_ner	Apache License 2.0
CSCD-NS	https://github.com/nghuyong/cscd-ns	MIT License

Table 7: Data license of the open-source datasets used in SmartBench.

A.2 More Evaluation Results

	Task	BlueLM-3B	InternVL2.5-4B	MiniCPM3-4B	Qwen2.5-3B	Qwen2-VL-2B
Text Summarization	Document Summ	7.20	6.98	6.95	7.09	4.74
	Call Summ	7.27	5.97	6.94	6.58	4.12
	Recording Summ	7.13	6.56	7.02	7.00	4.58
	Meeting Summ	7.22	6.70	7.10	7.07	4.36
Text Q&A	Document Q&A	8.45	8.46	7.79	8.63	8.34
	Retrieval Q&A	6.14	5.95	6.83	6.33	4.92
	Personal Q&A	8.37	8.30	8.04	8.57	8.16
Content Creation	Text Polishing	6.93	5.78	6.90	6.86	5.91
	Text Continuation	7.13	6.56	7.19	7.31	5.60
	Text Abbreviation	7.23	6.72	7.40	7.63	6.52
	Text Expansion	6.79	7.04	7.23	7.31	5.72
	Text Creation	6.73	5.78	6.67	6.63	5.02
	Text Formatting	6.35	5.93	7.03	5.25	2.93
	Instant Reply	5.60	5.09	5.25	5.26	3.17
	Text Correction	3.53	2.39	3.48	2.06	1.24
Information Extraction	Entity Extraction	7.77	7.40	7.44	6.21	5.00
	Relation Extraction	5.73	4.13	4.77	3.97	3.65
	Event Extraction	7.14	7.32	6.85	7.21	5.17
Notification Management	Message Summ	6.92	6.96	7.47	7.62	5.64
	Notification Sorting	5.38	4.63	5.56	5.21	2.93
AVG		6.75	6.23	6.70	6.49	4.89

Table 8: Evaluation results using Qwen-Max (`qwen-max-longcontext`) as the judge LLM with BF16 precision.

We present the scores evaluated by Qwen-Max as the judging LLM in Tab. 8 with BF16 precision. When compared to the GPT-4 Turbo results shown in Tab. 3, both Qwen-Max and GPT-4 Turbo rank the models in the same order. This demonstrates the robustness of our LLM-as-a-Judge approach.

We also include the INT4 precision inference performance (evaluated by Qwen-Max) of BlueLM-3B and Qwen2.5-3B on the vivo iQOO 12 smartphone (50 questions per task), along with the performance retention compared to the original BF16 models. As shown in Tab. 9.

Category	Task	BlueLM-3B			Qwen2.5-3B		
Precision		BF16	INT4	Retention (%)	BF16	INT4	Retention (%)
Text Summarization	Document Summ	6.89	5.62	81.61	6.67	4.00	60.00
	Call Summ	6.71	5.89	87.66	6.57	5.43	82.61
	Recording Summ	6.95	6.63	95.38	6.95	6.53	94.03
	Meeting Summ	7.39	5.57	75.29	7.10	5.10	71.84
Text Q&A	Document Q&A	8.50	8.32	97.83	8.85	8.77	99.13
	Retrieval Q&A	6.19	5.75	92.93	6.31	6.13	97.03
	Personal Q&A	8.06	7.16	88.80	8.42	8.26	98.08
Content Creation	Text Polishing	6.89	6.75	97.93	6.82	6.39	93.72
	Text Continuation	7.17	7.07	98.60	7.17	7.03	98.14
	Text Abbreviation	7.26	7.10	97.78	7.61	7.58	99.58
	Text Expansion	7.24	7.13	98.54	7.65	7.29	95.38
	Text Creation	6.96	6.48	93.15	6.62	6.23	94.19
	Text Formatting	6.45	6.38	98.93	5.34	5.00	93.55
	Instant Reply	5.60	5.21	93.04	4.95	4.15	83.84
	Text Correction	3.17	2.00	63.16	1.83	1.17	63.64
Information Extraction	Entity Extraction	6.98	6.66	95.46	5.65	5.34	94.40
	Relation Extraction	5.64	4.77	84.63	4.12	3.81	92.50
	Event Extraction	6.90	6.10	88.32	7.32	6.80	92.83
Notification Management	Message Summ	7.16	7.12	99.44	7.60	7.52	98.95
	Notification Sorting	5.50	5.34	96.95	5.35	5.15	96.20
AVG		6.68	6.15	92.08	6.45	5.88	91.29

Table 9: Evaluation results using Qwen-Max (`qwen-max-longcontext`) as the judge LLM with INT4 precision.

A.3 Details of Human Annotators

In the Data Screening and Human Test stages, we hire six domain experts with over five years of mobile AI experience. These experts have at least a master's degree. We pay them a labeling fee of $20 per hour.

A.4 Comparison with Traditional Benchmarks

Traditional benchmarks (e.g., MMLU) mainly evaluate the model's objective knowledge, while SmartBench focuses on subjective data aligned with end-side smartphone scenarios, assessing the degree of alignment with human preferences. In practice, due to memory and storage constraints on mobile devices, only a single MLLM can be deployed on the device. We consider the training process from LLM to MLLM. For example, InternVL2.5-4B is trained from Qwen2.5-3B-Instruct. During this process, objective knowledge tends to be preserved, but the subjective (human alignment) performance often degrades.

	MMLU	SmartBench (GPT-4)	SmartBench (Qwen-Max)
Qwen2.5-3B	66.31	6.61	6.49
InternVL2.5-4B	68.35	6.37	6.23

MMLU measures objective knowledge, while SmartBench evaluates subjective alignment. We test SmartBench performance using both GPT-4 and Qwen-Max as evaluators. As can be seen, during the training process from LLM to multimodal models, MMLU performance improves due to the injection of new knowledge. However, the subjective performance represented by SmartBench decreases.

A.5 Failure Cases on NPU

We here provide two failure cases of BlueLM-3B after quantized deployment on the NPU.

Example 1: Fluency Degrades after Quantization

Prompt:
你是一位优秀的文本扩写助手，请对给定内容进行适度扩写，每句不超过原句三倍长度。若原句为疑问句，扩写也应保持疑问形式。内容如下：

美丽的花园中蝴蝶在花丛中飞舞

Result Before Quantization:
美丽的花园里，五颜六色的鲜花开得正艳丽，蝴蝶们在它们中间自由飞舞。它们时而停在花瓣上，时而在花丛中翩翩起舞，如同花园里的一幅幅生动的图画。这一幕让人不禁感叹大自然的奇妙和生命的美好。

Result After Quantization:
蝴蝶在花丛中。花园很美丽。翅膀在阳光下闪烁。犹如微小的的艺术品。花儿开放，芬芳。蝴蝶流连忘返。

Prompt:
You are an excellent text expansion assistant. Please moderately expand the given content, with each expanded sentence not exceeding three times the length of the original. If the original sentence is a question, the expansion should maintain the interrogative form. The content is as follows:

In the beautiful garden, butterflies are fluttering among the colorful flowers.

Result Before Quantization:
In the beautiful garden, colorful flowers are in full bloom, and butterflies are freely fluttering among them. Sometimes they land on the petals, and sometimes they dance gracefully among the blossoms, like vivid paintings in the garden. This scene makes one marvel at the wonder of nature and the beauty of life.

Result After Quantization:
Butterflies are among the flowers. The garden is very beautiful. Wings shimmer in the sunlight. Like tiny artworks. Flowers bloom, fragrant. Butterflies linger, reluctant to leave.

Figure 5: Failure case after quantization on the NPU.

Example 2: Weakened Contextual Understanding Capability

Prompt:
请将一系列消息精炼地总结成一句话。确保总结准确、反映消息的主要内容且总结内容的长度不要超过消息内容的长度，输出格式为：

消息发送人姓名或昵称：
总结内容

请严格按照输出格式总结以下消息：
吕娜：我也这么感觉。就是那边拿房子吧没熟人好像不好买
吕娜：我家隔壁好像又盖了，看广告应该是绿地的
吕娜：你可以问问，淮中东边的那块地
吕娜：也不是啥好价，不过瓶子大气呀。
吕娜：我也感觉会在8500左右

Result Before Quantization:
吕娜：
感觉买房子需要熟人，询问淮东边的地，预计价格在8500左右。

Result After Quantization:
吕娜：
认为你买淮中东边的地可能会遇到熟人不好买的问题，建议你可以问问。

Prompt:
Please concisely summarize a series of messages into one sentence. Ensure the summary is accurate, reflects the main content of the messages, and does not exceed the length of the original messages. The output format should be:

Sender's name or nickname:
Summary content

Strictly follow the output format to summarize the following messages:
Lü Na: I feel the same way. It seems hard to buy a house there without connections.
Lü Na: Looks like they're building next to my place again. Judging by the ads, it should be Greenland's project.
Lü Na: You could ask about the plot of land east of Huai Middle School.
Lü Na: The price isn't great either, but the developer is generous.
Lü Na: I also think it'll be around 8500.

Result Before Quantization:
Lü Na:
Feels you need connections to buy property, inquires about land east of Huai, estimates price around 8500.

Result After Quantization:
Lü Na:
Thinks you might have trouble buying the land east of Huai Middle School with connections and suggests you ask about it.

Figure 6: Failure case after quantization on the NPU.

A.6 Example Prompt for Generating Task Queries

Help me generate a series of mobile notifications, each in the following format:

Subject (YYYY-MM-DD-HH-MM) [Specific message content]

You can use subjects including but not limited to:
{'Email from Zhang Wei', 'Delivery Reminder', 'WeChat message from Wang Li', 'Bank of China', 'Bank Account Alert', 'Health App Notification', 'Breaking News Push', 'Traffic Notification', 'Alipay App', 'SMS from Zhang Hua', 'Emergency Weather Alert', 'Traffic Reminder', 'Tencent Meeting', 'Online Meeting Reminder', 'JD Logistics', 'JD Express', 'Delivery Notification', 'Meituan Takeaway', 'Battery Status', 'Taobao', 'Fitness App', 'E-commerce Platform', 'Health Assistant', 'Health APP', 'System Update Prompt', 'Tencent News', 'Takeaway App', 'Security Alert', 'Traffic Fine', 'SMS from Li Lei', 'QQ Mail', 'DingTalk Reminder', 'Bank of Communications', 'Email', 'Bank Card Consumption Alert', 'DingTalk', 'Ctrip Travel', 'China Merchants Bank', 'NetEase Cloud Music', 'Power Company App', 'SF Express', 'Email Notification', 'Schedule Reminder', 'Mobile System Update', 'Meeting Reminder', 'Email Notification', 'WeChat Voice Call from Zhang Wei', 'Shopping Platform', 'Health Code', 'Xiaomi Sports', 'Douban Movie', 'JD Finance', 'Weibo', 'JD', 'Security Center', 'Taobao Notification', 'WeChat message from Wang Qiang', 'Weather', 'SMS from Li Xiao', 'Shopping Discount', 'E-commerce Shopping', 'Health App', 'Sports Health App', 'WeChat Friend Request', 'WeChat', 'Health Tip', 'Weather Forecast', 'Sports Health', 'TikTok', 'Takeaway APP', 'Takeaway/Ele.me', 'Low Battery Alert', 'WeChat message from Li Ming', 'Takeaway', 'Meituan Taxi', 'Health Reminder', 'SF Express', 'SMS/Carrier', 'Bank Notification', 'Missed Call', 'Mailbox', 'NetEase Mail', 'WeChat message from Li Na', 'Call Reminder', 'Bank Service Alert', 'Bank', 'Logistics Notification', 'Zhihu', 'App Update Notification', 'Power Company Notification', 'Ximalaya', 'Weather Alert', 'Bank Alert', 'Fitness APP', 'Health App Reminder', 'Bank Card', 'Health App', 'Taobao', 'Software Update Reminder', 'Mailbox/QQ Mailbox', 'Social Media Notification', 'Weather Reminder', 'Emergency Alert', 'Mobile Banking', 'New Email Alert', 'Travel Information', 'Network Connection', 'JD Home', 'Music Platform', 'Flight Status', 'Message from Wang Li', 'WeChat Message', 'WeChat Official Account', 'Social App Message', 'Mobile Phone Bill Reminder', 'Takeaway Order', 'Voice/Video Call Invitation', 'Movie Ticketing', 'Keep', 'Taobao Express', 'Didi Chuxing', 'Alipay Notification', 'Music APP', 'JD Mall', 'Alarm Clock', 'Video Conference Reminder', 'SMS/MMS', 'China Mobile', 'System Prompt', 'Tmall Supermarket', 'Breaking News', 'Bank SMS', 'System Reminder', 'Zhihu Daily', 'Alipay', 'System Update', 'SMS', 'WeChat Voice Call from Zhang San'}

Please choose a different subject for each notification. You can generate two types of notifications: categorize them by whether they require user action; for those that need to be handled, please clearly distinguish their urgency; for purely informational notifications that do not require action, feel free to generate them.

Please also generate the correct sorting. The sorting rules are:

Sorting Rules:
1. **Sort by Urgency and Importance**: All notifications and events should first be sorted based on their urgency and importance, with urgent and important items given priority. For example, if your data is about to run out completely, a family member happens to be ill, or your boss notifies you of an immediate meeting.
2. **Events Completed or Needing Early Completion Have Priority**: If Event A has been completed or needs to be completed earlier than Event B, then Event A has a higher priority and should be addressed first. For example, tasks that need to be completed now are more urgent than those due tomorrow; tasks due tomorrow are more urgent than those due the day after.
3. **When Urgency Is Equal, Sort by Time Order**: When it's impossible to distinguish the urgency between events, they should be sorted according to the time they occur, with earlier events given priority.
4. **Notifications Requiring Immediate Action Have Priority**: For notifications that require immediate user action, such as meal pick-up notifications or vehicle arrival reminders, they should be given the highest priority to avoid missing important opportunities.
5. **Notifications Related to Basic Physiological Needs Are More Important**: Notifications involving the user's basic physiological needs, such as eating or health reminders, should be considered more important and given higher priority over other general notifications.
6. **Notifications Related to Upcoming Schedules Are Next**: Notifications directly related to the user's short-term itinerary or arrangements, such as upcoming meetings or flight reminders, should be considered after urgent matters.
7. **Notifications Affecting Future Plans Come Next**: Notifications about future matters that need attention, such as weather forecasts or future activity arrangements, have less impact on the current situation and can be handled later.
8. **Social Interaction Notifications Are Last**: Non-urgent social media reminders, such as new friend requests or like notifications, do not require immediate action and can be handled last.
9. **Final Confirmation Based on Event Timeliness**: If the notification involves an event that is about to expire or has already occurred, it should be handled promptly to avoid missing important information, ensuring all matters are attended to within an appropriate time frame.

In short: First sort by urgency and importance (whether they require user action); when unable to distinguish urgency, sort according to time order. Please generate a series of notifications and the sorted results.

Generate 5-6 notifications; please strictly follow the template below:

Notifications:

Sorting:

Explanation:

An example for reference:

Notifications:

2024.2.1-10:20 Meituan Takeaway: Your takeaway has been delivered, please pick it up promptly.

2024.2.1-11:20 Email from Zhang Wei: Regarding the scheduling of next week's project meeting, please check the attachment.

2024.2.1-11:30 Weibo: @Your friend liked your post

2024.2.1-12:00 JD: Your purchased item has been signed for, welcome to shop again.

2024.2.1-11:50 NetEase Cloud Music: Your favorite singer has released a new album, go have a listen.

2024.2.1-10:30 Health App: Your daily steps have met the goal, you've earned a badge.

2024.2.1-10:20 High-Speed Rail Assistant: Your reserved ticket has been successfully purchased, please pay promptly.

2024.2.1-11:00 Fitness App: There's a new training plan today, don't forget to complete it.

Sorting:

2024.2.1-10:20 High-Speed Rail Assistant: Your reserved ticket has been successfully purchased, please pay promptly.

2024.2.1-10:20 Meituan Takeaway: Your takeaway has been delivered, please pick it up promptly.

2024.2.1-11:20 Email from Zhang Wei: Regarding the scheduling of next week's project meeting, please check the attachment.

2024.2.1-10:30 Health App: Your daily steps have met the goal, you've earned a badge.

2024.2.1-11:00 Fitness App: There's a new training plan today, don't forget to complete it.

2024.2.1-11:30 Weibo: @Your friend liked your post

2024.2.1-11:50 NetEase Cloud Music: Your favorite singer has released a new album, go have a listen.

2024.2.1-12:00 JD: Your purchased item has been signed for, welcome to shop again.

Explanation:

[Explanation about this sorting]

Please expand your imagination, generate diverse subjects, be imaginative but consistent with real life by creating 5-6 notifications and their sorting. Try to make them easily distinguishable. Please randomly shuffle the times in the **Notifications** section so they are not displayed in time order before sorting. Additionally, please pay attention to logical issues, ensuring that the sender matches the content of the notification; the sorted results must correspond with your explanation.

Figure 7: Prompt for generating mobile notification sorting pairs (en).

A.7 More Example Cases

We provide examples of SmartBench, along with the inference results of different models using BF16 precision (zh). The corresponding English translation is also provided (en). The numbers in red represent the scores given by Qwen-Max (`qwen-max-longcontext`).

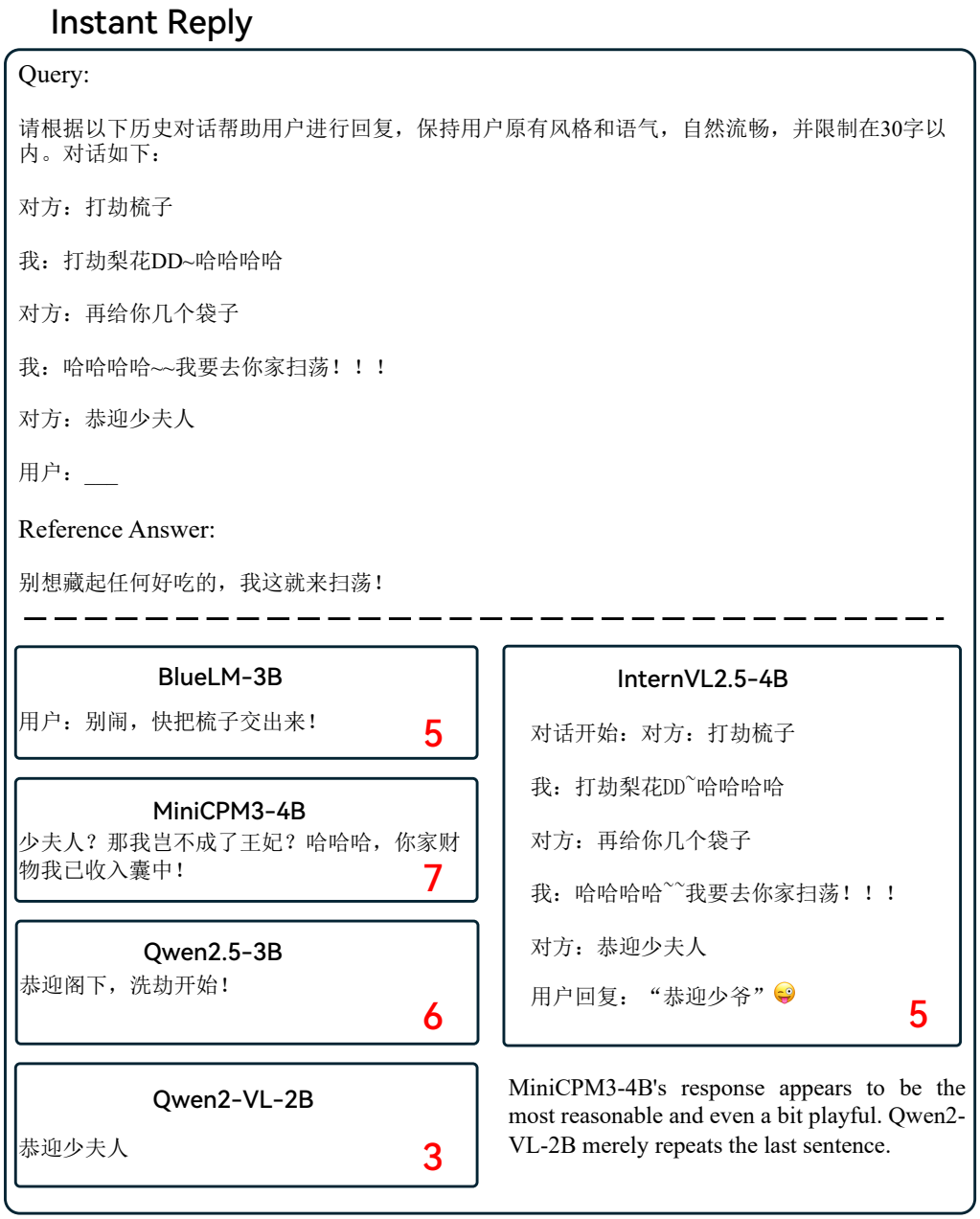


Figure 8: Instant reply case (zh).

Instant Reply

Query:

Based on the historical dialogue below, help the user craft a reply that maintains their original style and tone, sounds natural, and stays under 30 words.

Dialogue:
Them: Robbing combs

Me: Robbing Pear Blossom DD~ Hahaha

Them: Here are more bags for you

Me: Hahaha~~ I'm coming to raid your place!!!

Them: Greetings, Young Madam

User: ___

Reference Answer:

Don't think you can hide any snacks—I'm raiding them all!

BlueLM-3B

User: Stop messing around, give me the comb!

5

MiniCPM3-4B

Young Madam? Then wouldn't I become the princess consort? Hahaha, I've already pocketed all your treasures!

7

Qwen2.5-3B

Welcome, Your Excellency—the raid begins!

6

Qwen2-VL-2B

Greetings, Young Madam

3

InternVL2.5-4B

Them: Robbing combs

Me: Robbing Pear Blossom DD~ Hahaha

Them: Here are more bags for you

Me: Hahaha~~ I'm coming to raid your place!!!

Them: Greetings, Young Madam

User: "Welcome, Young Master" 😜

5

MiniCPM3-4B's response appears to be the most reasonable and even a bit playful. Qwen2-VL-2B merely repeats the last sentence.

Figure 9: Instant reply case (en).

Text Correction

Query:

待纠正的文本如下：

这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

请直接输出纠正后结果：

Reference Answer:

纠正后文本：这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院等，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

纠正内容：“灯”纠正为“等”

BlueLM-3B

纠正后文本：这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

纠正内容：无 0

InternVL2.5-4B

这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

0

MiniCPM3-4B

纠正后文本：这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

纠正内容：无 0

Qwen2.5-3B

纠正后文本：这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

纠正内容：无 0

Qwen2-VL-2B

纠正后文本：这座位于南四教学楼西北角的竹房建造面积60平方米，设有客厅、书房、露台、庭院灯，全部由环保绿色新材料胶合竹制作，总费用10余万元，可使用50年

纠正内容：无 0

The living room, study, and terrace all refer to rooms, whereas the courtyard light is an object, which is illogical. None of the models detected this error.

Figure 10: Text correction case (zh).

Text Correction

Query:

The text to be corrected is as follows:

This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Please output the corrected result directly:

Reference Answer:

The corrected text: This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard, etc, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Corrected content: Change “light” to “etc”.

BlueLM-3B

The corrected text: This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Corrected content: None. 0

InternVL2.5-4B

This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

0

MiniCPM3-4B

The corrected text: This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Corrected content: None. 0

Qwen2.5-3B

The corrected text: This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Corrected content: None. 0

Qwen2-VL-2B

The corrected text: This bamboo structure, located in the northwest corner of South 4th Teaching Building, covers a construction area of 60 square meters. It includes a living room, study, terrace, and courtyard light, all made of eco-friendly green new material—glued laminated bamboo. The total cost exceeds 100,000 yuan, and it has a lifespan of 50 years.

Corrected content: None. 0

The living room, study, and terrace all refer to rooms, whereas the courtyard light is an object, which is illogical. None of the models detected this error.

Figure 11: Text correction case (en).

Event Extraction

Query:

以下输入文本描述的是与“飞机票”相关的场景，请根据文本内容进行信息抽取，请根据文本内容进行信息抽取。

【抽取要求】：

1.请根据输入文本判断每件事件的类别并提取相应的属性。

2.文本中可能包含购买机票、航班改签、航班调整、航班延误、航班取消、航班值机、乘机提醒、退票成功八种事件；

3.要求2中的事件属性可能包含以下属性：起飞时间、出发地、到达地、乘机人、航班、机场、票价、座位号、登机口、航站楼、购票状态、支付状态、到达时间、改签费、新航班、延误时间、退款金额、退票费、预计到账时间、退票受理时间、订单号、票号、电子登机牌、值机验证码、值机方式、原订航班、原航班、原起飞时间、原订起飞时间、原订到达时间、原订出发地、原订到达地、原登机口、调整原因、原出发地、延误提示航空公司、退款到账时间、客服电话属性。

4.请注意，未提及的属性严禁抽取；

5.输出格式一定要严格按照规定的可解析的json格式，如 [{"事件名":"xxx","属性":{"属性1名称":"xxx","属性2名称":"xxx"}}]，严禁输出多余内容。

6.若文本未涉及上述事件类别，则输出内容为[{"事件名":"无规定事件","属性":"无"}]。

【输入文本】：

【畅享旅程】通知：您的拉萨贡嘎至兰州中川航班（JH234）已调整至2023-10-25 13:00起降，请及时关注最新动态。

【抽取结果】：

Reference Answer:

[{"事件名": "航班调整", "属性": {"航班": "JH234", "起飞时间": "2023-10-25 13:00", "出发地": "拉萨贡嘎", "到达地": "兰州中川"}}]

BlueLM-3B

[{"事件名":"航班调整","属性":{"原起飞时间":"无","原到达时间":"无","原出发地":"拉萨贡嘎","原订起飞时间":"无","原订到达地":"兰州中川","原订到达时间":"无","原订出发地":"拉萨贡嘎","原登机口":"无","调整原因":"无","原出发地":"拉萨贡嘎"}}]

3

Qwen2.5-3B

[{"事件名":"航班调整","属性":{"航班":"JH234","原订起飞时间":"无","原订到达时间":"无","原出发地":"拉萨贡嘎机场","原订出发地":"兰州中川机场","原订到达地":"无","到达地":"兰州中川机场","出发地":"拉萨贡嘎机场","起飞时间":"2023-10-25 13:00","调整原因":"无"}}]

6

Figure 12: Event extraction case (zh).

Event Extraction

Query:

The following input text describes a scenario related to "airline tickets." Please extract the relevant information based on the text content.

[Extraction Requirements]:
1.Please determine the category of each event based on the input text and extract the corresponding attributes.

2.The text may include the following eight types of events: ticket purchase, flight rescheduling, flight adjustment, flight delay, flight cancellation, check-in, boarding reminder, and ticket refund success.

3.The event attributes in requirement 2 may include the following: departure time, departure location, arrival location, passenger, flight number, airport, ticket price, seat number, gate, terminal, ticket status, payment status, arrival time, rescheduling fee, new flight, delay time, refund amount, refund fee, expected refund time, refund processing time, order number, ticket number, e-boarding pass, check-in verification code, check-in method, original flight, original departure time, original scheduled departure time, original scheduled arrival time, original departure location, original arrival location, original gate, adjustment reason, delayed flight airline, refund arrival time, customer service phone number.

4.Please note, attributes that are not mentioned in the text must not be extracted.

5.The output format must strictly follow the specified parsable JSON format, such as [{"Event name": "xxx", "Attributes": {"Attribute 1 name": "xxx", "Attribute 2 name": "xxx"}}], and no extra content is allowed.

6.If the text does not involve any of the aforementioned event categories, the output should be [{"Event name": "No specified event", "Attributes": "None"}].

[Input Text]:

[Enjoy Your Journey] Notification: Your flight from Lhasa Gongga to Lanzhou Zhongchuan (JH234) has been adjusted to 2023-10-25 13:00 for departure. Please stay updated with the latest news.

[Extraction Result]:

Reference Answer:

[{"Event name": "Flight Adjustment", "Attributes": {"Flight": "JH234", "Departure Time": "2023-10-25 13:00", "Departure Location": "Lhasa Gongga", "Arrival Location": "Lanzhou Zhongchuan"}}]

BlueLM-3B

[{"Event name": "Flight Adjustment", "Attributes": {"Original Departure Time": "None", "Original Arrival Time": "None", "Original Departure Location": "Lhasa Gongga", "Original Scheduled Departure Time": "None", "Original Scheduled Arrival Location": "Lanzhou Zhongchuan", "Original Scheduled Arrival Time": "None", "Original Scheduled Departure Location": "Lhasa Gongga", "Original Gate": "None", "Adjustment Reason": "None", "Original Departure Location": "Lhasa Gongga"}}] 3

Qwen2.5-3B

[{"Event name": "Flight Adjustment", "Attributes": {"Flight": "JH234", "Original Scheduled Departure Time": "None", "Original Scheduled Arrival Time": "None", "Original Departure Location": "Lhasa Gongga Airport", "Original Scheduled Departure Location": "Lanzhou Zhongchuan Airport", "Original Scheduled Arrival Location": "None", "Arrival Location": "Lanzhou Zhongchuan Airport", "Departure Location": "Lhasa Gongga Airport", "Departure Time": "2023-10-25 13:00", "Adjustment Reason": "None"}}] 6

Figure 13: Event extraction case (en).

Text Polishing

Query:

请在保持原意的基础上，对以下文本进行润色，使其更加通顺、连贯、自然。文本内容如下：

我也更喜欢篮球，下次让他带着我去玩儿～～

润色后文本：

Reference Answer:

篮球真是我的心头好，下次希望能有幸跟他一起切磋呢！

BlueLM-3B

我个人对篮球更感兴趣，下次有机会的话，让他带我去体验一下篮球的魅力吧。

8

InternVL2.5-4B

我同样也偏好篮球，下次让他带我去体验一回。

6

MiniCPM3-4B

在我的兴趣爱好中，篮球占据了相当大的比重，我对其的热爱不言而喻。因此，我非常期待下一次的机会，能够与他一同踏上篮球场，共同体验投篮、运球、防守的乐趣。我相信，这将是一次非常愉快的体验，我期待着与你一起在篮球场上挥洒汗水，享受运动带来的快乐。

5

Qwen2.5-3B

我也更喜欢篮球，下次可以让他带我去玩儿。

7

Qwen2-VL-2B

我也更喜欢篮球，下次带他一起去玩儿～～

4

The refinements of BlueLM-3B are reasonable and readable; the refinements of MiniCPM3-4B are too long; Qwen2-VL-2B reverses the meaning, it should be that he takes me out to play.

Figure 14: Text polishing case (zh).

Text Polishing

Query:

Please polish the following text while maintaining the original meaning to make it more smooth, coherent, and natural. The text is as follows:

I also like basketball more. Next time, let him take me to play~~

Polished text:

Reference Answer:

Basketball is truly my favorite. I hope I'll have the chance to compete with him next time!

BlueLM-3B

I'm personally more interested in basketball. If there's a chance next time, I'd love for him to take me to experience the charm of basketball.

8

InternVL2.5-4B

I also prefer basketball. Next time, let him take me to experience it.

6

MiniCPM3-4B

Basketball occupies a significant place in my hobbies, and my love for it is evident. Therefore, I'm really looking forward to the next opportunity to step onto the basketball court with him, experiencing the joys of shooting, dribbling, and defending together. I believe this will be an incredibly enjoyable experience, and I look forward to sweating it out with you on the court, enjoying the happiness that sports bring.

5

Qwen2.5-3B

I also prefer basketball. Next time, he can take me to play.

7

Qwen2-VL-2B

I also prefer basketball. Next time, I'll take him to play with me~~

4

The refinements of BlueLM-3B are reasonable and readable; the refinements of MiniCPM3-4B are too long; Qwen2-VL-2B reverses the meaning, it should be that he takes me out to play.

Figure 15: Text polishing case (en).

A.8 English Translation of Pictures in the Paper

Message Summarization Query
Sea on the Moon: My Ragdoll cat sheds a lot, especially during seasonal changes—it's like a walking dandelion! 😭 Do you have any good solutions? Waiting online, kinda urgent!

Sea on the Moon: I've tried many types of cat food, but none seem to work well. I have to vacuum so much fur every day, it's exhausting... 😩

Reference
Sea on the Moon:
Worried about the severe shedding of my Ragdoll cat and seeking solutions.

Figure 16: Translated example of the Message Summarization task in SmartBench.

Evaluation Dimensions:

1. Coherence: Check whether the continuation naturally connects with the preceding text, maintaining a consistent theme and context, while avoiding abrupt shifts or the introduction of irrelevant information.
2. Consistency: Assess whether the continuation aligns with the style, tone, and character traits established in the preceding text, ensuring a unified voice throughout the narrative or discussion.
3. Creativity: Evaluate whether the continuation demonstrates novel ideas or interesting plot developments while maintaining coherence and consistency, rather than simply repeating existing information.
4. Language Quality: Analyze whether the language in the continuation is clear, fluent, and free of grammatical, spelling, or punctuation errors. It should also use rich vocabulary and well-structured sentences to enhance readability and appeal.

Scoring Criteria:

1. Compare the AI assistant's response with the reference answer, identifying any shortcomings in the AI's response and providing further explanation.
2. Evaluate the AI assistant's response across the different dimensions, assigning a score of 1 to 10 for each dimension.
3. Based on the evaluation of each dimension, provide an overall score of 1 to 10 for the AI assistant's response.
4. Your scoring should be as strict as possible. In general, the higher the quality of the AI assistant's response, the higher the score.
   - ✓When the AI assistant's answer exhibits obvious logical gaps, severe deviation from the topic, or contains a large amount of irrelevant information, the total score must be 1 to 2 points;
   - ✓When the AI assistant's answer does not severely deviate from the topic but is of low quality, failing to effectively continue the style or plot of the preceding text, the total score is 3 to 4 points;
   - ✓When the AI assistant's answer basically meets the requirements of coherence and consistency, but performs poorly in creativity and/or language quality, the total score can be 5 to 6 points;
   - ✓When the AI assistant's answer quality is similar to the reference answer, performing well in all dimensions, the total score is 7 to 8 points;
   - ✓Only when the AI assistant's answer quality significantly surpasses the reference answer, perfectly continuing the preceding text and excelling in creativity and language quality, can it receive 9 to 10 points.

Figure 17: Evaluation Dimension & Scoring Standard (in English) for the text continuation task.