DentalBench: Benchmarking and Advancing LLMs Capability for Bilingual Dentistry Understanding

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) and medical LLMs (Med-LLMs) have demonstrated strong performance on general medical benchmarks. However, their capabilities in specialized medical fields, such as dentistry which require deeper domain-specific knowledge, remain underexplored due to the lack of targeted evaluation resources. In this paper, we introduce **DentalBench**, the first comprehensive bilingual benchmark designed to evaluate and advance LLMs in the dental domain. DentalBench consists of two main components: **DentalQA**, an English-Chinese question-answering (QA) benchmark with 36,597 questions spanning 4 tasks and 16 dental subfields; and **DentalCorpus**, a large-scale, high-quality corpus with 337.35 million tokens curated for dental domain adaptation, supporting both supervised fine-tuning (SFT) and retrieval-augmented generation (RAG). We evaluate 14 LLMs, covering proprietary, open-source, and medical-specific models, and reveal significant performance gaps across task types and languages. Further experiments with Qwen-2.5-3B demonstrate that domain adaptation substantially improves model performance, particularly on knowledge-intensive and terminology-focused tasks, and highlight the importance of domain-specific benchmarks for developing trustworthy and effective LLMs tailored to healthcare applications.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in a wide range of domains (OpenAI et al., 2024; Yang et al., 2025; Liu et al., 2024; Team et al., 2025; DeepSeek-AI et al., 2025; Jaech et al., 2024). Especially in the medical field, recent studies have shown that LLMs can achieve expert-level performance on various clinical benchmarks (Wu et al., 2025; Li et al., 2023; Wu et al., 2024; Zhou et al., 2023). However, reliable and fine-grained evaluation of LLM performance in specialized medical subfields-such as dentistry-remain limited, because of the shortage of domain-specific knowledge in general medical corpora or benchmarks.

As an important and highly specialized branch of medicine that spans multiple subfields and involves complex procedures, oral healthcare is in great need of artificial intelligence integration. Although there have been some studies exploring the integration of deep learning techniques into dentistry (Shi et al., 2024; Wei et al., 2020; Xiong et al., 2023; Liu et al., 2023), LLMs remain underevaluated due to the lack of targeted evaluation resources. It hinders not only the understanding of current LLM limitations but also the development of robust systems for clinical applications.

Therefore, in this paper, we introduce **DentalBench**, a comprehensive benchmark and corpus designed for evaluating and advancing LLM performance in the dental domain. We first construct **DentalQA**, an English-Chinese question-answering (QA) benchmark covering 4 task formats and 16 specialized subfields. Then, we develop **DentalCorpus**, a professionally curated bilingual corpus with large-scale and high-quality, aimed at dental-domain adaptation. Using DentalQA, we systematically evaluate various proprietary, open-source and medical-specific LLMs and reveal significant limitations for current models to finish knowledge-intensive tasks in dentistry. Further experiments based on supervised fine-tuning (SFT) and retrieval-augmented generation (RAG) by the DentalCorpus demonstrate that access to in-domain data can substantially improve model performance in specialized oral healthcare tasks, highlighting the importance of benchmarks for domain adaptation in real-world applications. Our main contributions are summarized as follows:

- We introduce **DentalQA**, the first bilingual

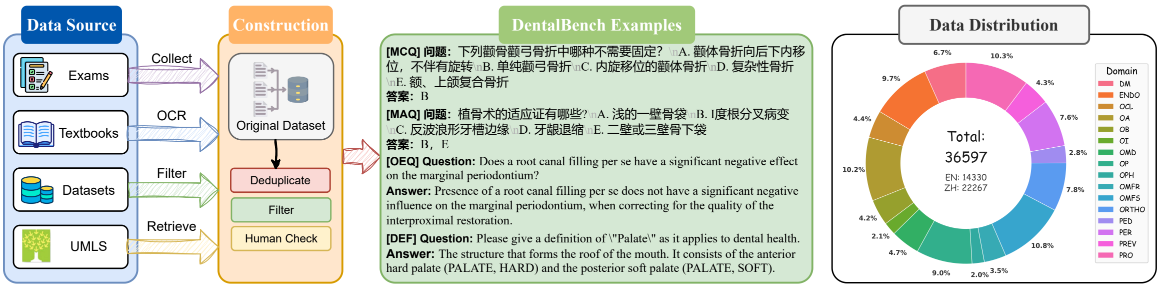

Figure 1: Overview of the DentalBench. It encompasses the following 16 dental specialties and disciplines: dental materials (DM), endodontics (ENDO), occlusion (OCL), oral anatomy (OA), oral biology (OB), oral implantology (OI), oral mucosal diseases (OMD), oral pathology (OP), oral pharmacology (OPH), oral and maxillofacial radiology (OMFR), oral and maxillofacial surgery (OMFS), orthodontics (ORTHO), pediatric dentistry (PED), periodontics (PER), preventive dentistry (PREV), and prosthodontics (PRO).

benchmark for dentistry-specific language understanding, consisting of 36,597 questions across 4 task types and 16 subfields.

- We create **DentalCorpus**, a large-scale, high-quality corpus containing 337.35 million tokens curated for dental domain adaptation with SFT and RAG methods.
- We evaluate 14 LLMs—including proprietary, open-source, and medical-specific models—on DentalQA, revealing clear performance gaps across task types and languages. Through extensive experiments, we further demonstrate that domain adaptation with DentalCorpus significantly improves general LLM performance in the dental domain.

2 DentalBench Dataset

We introduce **DentalBench**, the first comprehensive dataset for evaluating and adapting LLMs in the dental domain, as shown in Figure 1. It consists of: **DentalQA**, a bilingual benchmark for evaluating knowledge-based reasoning in oral heathcare, and **DentalCorpus**, a large-scale and high-quality text corpus curated for dental domain adaptation.

2.1 DentalQA

We construct **DentalQA**, a high-quality English-Chinese benchmark comprising 36,597 questions, covering 4 task formats and 16 dental subfields.

Task Formats. DentalQA includes the following four question types: **(a) MCQ:** Single-answer multiple choice questions (4 in English, 5 options in Chinese), testing factual recall. **(b) MAQ:** Multi-answer multiple choice questions (Chinese only), assessing comprehensive diagnostic knowledge. **(c) OEQ:** Open-ended questions simulating clinical and theoretical scenarios, used to evaluate reasoning and generation. **(d) DEF:** Terminology definition questions, requiring understanding of domain-specific dental terms.

Domain Coverage. Each question is categorized into one of 16 dental subfields (e.g., oral anatomy, periodontics, orthodontics) based on standard textbook classifications of the 8th round of the National Higher Education Curriculum for Five-Year Undergraduate Dental Medicine Programs (e.g., Zhao et al. (2020)). Figure 1 shows examples across task formats and domain data distributions, with additional details provided in Appendix C.1.

Data Sources. The English dataset is curated from seven public medical QA datasets: MMLU (Hendrycks et al., 2021), MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), MedQuAD (Ben Abacha and Demner-Fushman, 2019), PubMedQA (Jin et al., 2019), and iCliniq (Regin, 2017), Medical Meadow Flashcards and Medical Meadow Wikidoc (Yu et al., 2024). Then, we use a keyword list derived from the DentalCorpus filtering process to filter the datasets. Furthermore, we use dental terms from a bilingual glossary compiled from textbooks in the DentalCorpus pipeline and retrieve their definitions from UMLS (U.S. National Library of Medicine, 2025b) to construct English DEF questions. The Chinese dataset includes questions from the China National Dental Licensing Examination (1999–2021), 34 dental textbooks and auxiliary materials, and 181 OEQs derived from real orthodontist-patient interactions.

Construction. We apply a unified pipeline across both languages. MCQs and MAQs are normalized to fixed option counts. DEF questions are generated by filling 50 predefined templates per language (Appendix A.1) with extracted dental terms and their definitions. OEQs are preserved in their original form without modification. To ensure quality and domain relevance, we use GPT-4o to classify all questions into three categories: *oral-related*, *non-oral*, and *insufficient* (Appendix A.2). The *insufficient* category is used for questions with incomplete or corrupted content. We filter and retain only the *oral-related* questions.

Human Validation. To assess classification accuracy, we manually reviewed 300 representative samples—50 for each combination of language and category. The results indicate strong agreement: 100%, 96%, and 94% for English, and 96%, 92%, and 92% for Chinese.

2.2 DentalCorpus

We construct **DentalCorpus**, a bilingual resource designed to support domain adaptation and retrieval-augmented generation in dentistry.

Data Sources. DentalCorpus is built from three major sources: **(a) Textbooks.** We collect 40 Chinese dental textbooks and auxiliary materials, remove non-content sections and apply OCR to obtain 4.1M characters of clean text. We also extract a bilingual glossary of 1,971 dental terms from glossaries. **(b) PubMed Articles.** Using 28 MeSH terms (listed in Appendix B.1), we retrieve 54,651 freely accessible full-text articles from PubMed (U.S. National Library of Medicine, 2025a), published between 2000 and 2024, yielding 983.3M English and 5.4M Chinese characters. **(c) Open Medical Datasets.** We filter MMedC (Qiu et al., 2024) (EN: 10.56B, ZH: 4.35B tokens) and MedRAG (Zhao et al., 2025) (23.9M PubMed snippets) to retain dental-relevant content.

Construction. We implement a rule-based filtering pipeline using keyword lists derived from TF-IDF analysis on dental and general medical corpora. Starting from vocabularies built on PubMed, MedRAG, and textbook texts, we intersect them with the glossary to obtain candidate dental terms. Terms that appear disproportionately in general medical texts are removed. The final filtering lists contain 440 English and 235 Chinese keywords.

Texts from all sources are filtered using these keywords. We apply a keyword density threshold of >1% and require at least two distinct matches per sentence. English is tokenized by spaces; Chinese uses direct string matching.

After filtering, we deduplicate the corpus using MD5 hashes, embed texts with the bge-m3 model, and segment into chunks of up to 512 tokens. The final corpus consists of 1.06M English chunks (319.08M tokens) and 66.3K Chinese chunks (18.27M tokens).

Human Validation. We manually reviewed 100 random samples per language to assess filter quality, confirming domain relevance rates of 99% for English and 96% for Chinese.

3 Experiments

3.1 Experimental Setup

We split DentalQA into training and test sets in a 4:1 ratio while preserving each subfield's proportions, and report all results on the held-out test set. MCQ performance is measured by Accuracy; MAQ by Accuracy, Precision, Recall and F1; and OEQ and DEF by BERTScore F1 (Zhang et al., 2019). We conduct our experiments on multiple popular LLMs. For general LLMs, we select DeepSeek-V3, DeepSeek-R1, GPT-4o, GPT-4o-mini, LLaMA-3.2-3B-Instruct, LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-1.5/3B/7B/14B/32B-Instruct (Qwen et al., 2025). For medical LLMs, we select BioMistral-7B (Labrak et al., 2024), HuatuoGPT2-7B (Chen et al., 2024) and LLaMA-3-8B-UltraMedical (Zhang et al., 2024). We evaluate in a zero-shot setting using task-specific prompt templates (Appendix A.3). Experiments are conducted on eight NVIDIA RTX 3090 GPUs.

3.2 Domain Adaptation on Qwen2.5-3B

To enhance dentistry-specific knowledge and capabilities, we adopt three adaptation strategies based on Qwen-2.5-3B-Instruct. **(a) Supervised Fine-Tuning (SFT):** Full-model fine-tuning on the DentalQA training split for four epochs with a learning rate of 1e-6 and batch size 16 using bfloat16 precision. **(b) Retrieval-Augmented Generation (RAG):** At inference, retrieve the top-5 most relevant passages from DentalCorpus via FAISS with bge-m3 embeddings and prepend them to the prompt (Appendix A.3). **(c) SFT + RAG:** Combine the above supervised fine-tuning with retrieval augmentation during inference.

3.3 Results

The main results are presented in Table 1, where we report the performance of 14 LLMs and our domain adaptation results.

Overall Trends. Performance varies markedly by language. On DentalBench-ZH, DeepSeek-R1 achieves state-of-the-art accuracy on both MCQ

Table 1: Overall Performance on DentalQA. We use Accuracy (ACC), Precistion (P), Recall (R), F1, and BERTScore F1 (BERTScore) as our metrics. **Bold** indicates the best result, and underline indicates the second best.

	DentalQA-ZH							DentalQA-EN		
	MCQ	MAQ				OEQ	DEF	MCQ	OEQ	DEF
Model	ACC	ACC	P	R	F1	BERTScore	BERTScore	ACC	BERTScore	BERTScore
General LLMs										
GPT-4o	64.86	37.30	87.75	81.74	84.63	27.23	21.60	**73.98**	31.28	29.21
GPT-4o-mini	51.65	29.73	81.36	87.37	84.26	26.48	19.50	60.59	34.55	29.42
Deepseek-V3	69.28	41.35	87.22	86.23	86.72	27.79	17.78	68.28	27.65	25.73
Deepseek-R1	**76.06**	**43.51**	**88.64**	86.68	**87.65**	26.77	15.81	60.04	20.91	18.58
Llama-3.2-3B	38.22	7.30	72.24	65.01	68.44	19.88	15.13	48.96	28.13	26.13
Llama-3.1-8B	40.80	10.27	77.45	67.49	72.13	16.69	4.96	55.60	25.31	21.75
Qwen2.5-1.5B	45.58	13.24	76.67	77.55	77.11	21.74	8.57	38.09	26.10	21.59
Qwen2.5-3B	48.63	19.19	77.70	80.37	79.01	20.89	11.16	41.77	34.48	29.62
Qwen2.5-7B	60.29	26.22	83.08	79.22	81.11	26.37	11.59	49.23	26.28	22.12
Qwen2.5-14B	66.48	33.51	84.05	85.01	84.53	25.47	12.69	50.49	26.68	21.93
Qwen2.5-32B	70.86	39.46	85.50	86.15	85.82	26.02	11.65	58.34	26.59	22.78
Medical LLMs										
BioMistral-7B	25.44	5.68	76.33	47.43	58.51	14.48	14.06	34.96	34.50	29.55
HuatuoGPT2-7B	22.51	6.22	74.50	67.54	70.85	25.38	21.04	25.47	15.50	16.55
Llama-3-8B-UltraMedical	30.32	11.08	72.86	81.52	76.95	18.74	9.18	46.10	26.76	24.80
Domain Adaptation on Qwen2.5-3B										
Qwen2.5-3B	48.63	19.19	77.70	80.37	79.01	20.89	11.16	41.77	34.48	29.62
w. SFT	54.58	25.60	75.57	93.24	83.48	22.42	15.29	47.90	**37.74**	**30.79**
w. RAG	54.45	21.35	74.88	91.17	82.22	**30.18**	**22.13**	48.74	36.47	30.04
w. SFT+RAG	60.06	29.07	77.30	**93.46**	84.62	30.06	20.85	52.15	37.68	29.65

and MAQ, with DeepSeek-V3 and Qwen2.5-32B close behind. Conversely, on DentalBench-EN, GPT-4o leads across these tasks. In both languages, however, open-ended tasks (OEQ and DEF) trail far behind MCQ and MAQ, underscoring enduring challenges in domain-specific generative reasoning and terminology.

General Models vs. Medical Models. Although medical LLMs perform relatively well on OEQ and DEF, they fall markedly short of general-purpose models on MCQ and MAQ. For example, Llama-3.1-8B consistently outperforms its medical counterpart across all multiple-choice tasks, suggesting that medical tuning may insufficiently capture dentistry-specific factual knowledge.

Impact of Model Scale. In the Qwen-2.5 family, scaling improves MCQ and MAQ notably but yields limited gains on OEQ and DEF, suggesting factual recall benefits more from model size than generative reasoning does.

Domain Adaptation. Both SFT and RAG improve MCQ and MAQ, but RAG shows a larger impact on open-ended tasks (e.g., OEQ-ZH BERTScore: +9.29 vs. +1.53). Combining both yields additive gains—especially on MCQ and MAQ (+11.43 and +9.88). For OEQ/DEF, SFT+RAG offers clear benefit over SFT alone in Chinese, while in English the effect is less consistent, indicating language sensitivity in retrieval effectiveness.

4 Conclusion

We introduce DentalBench, a comprehensive bilingual benchmark designed for evaluating and enhancing LLMs in the dental domain. It includes 2 main components: DentalQA, the first bilingual high-quality QA dataset for dentistry, and DentalCorpus, a large-scale domain-specific English-Chinese corpus for domain adaptation, such as SFT and RAG. Our experiments across 14 LLMs, covering proprietary, open-source and medical-specific models, reveal significant performance gaps based on task types, language, and model categories. Additionally, through extensive experiments, we demonstrate that domain adaptation using DentalCorpus can significantly improve performance. In general, DentalBench can be served as a valuable resource for evaluating knowledge-grounded language models in dentistry, improving language understanding in oral healthcare, and encouraging more related research.

Limitations

Our work has several limitations. First, the dataset exhibits asymmetry between Chinese and English sources. While both languages are supported throughout DentalQA and DentalCorpus, the distribution, source diversity, and depth of coverage are not fully aligned—potentially contributing to observed cross-lingual performance gaps. Second, the MAQ format is currently only available in Chinese, limiting comprehensive evaluation of multi-answer reasoning capabilities in English. In future work, we aim to construct balanced bilingual resources and expand task coverage across languages.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiaxiang Liu, Jin Hao, Hangzheng Lin, Wei Pan, Jianfei Yang, Yang Feng, Gaoang Wang, Jin Li, Zuolin Jin, Zhihe Zhao, and 1 others. 2023. Deep learning-enabled 3d multimodal fusion of cone-beam ct and intraoral mesh scans for clinically applicable tooth-bone reconstruction. *Patterns*, 4(9).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Lasse Regin. 2017. Medical question answer data. `https://github.com/LasseRegin/medical-question-answer-data`. Accessed: May 15, 2023.

Zefeng Shi, Zijie Meng, Ruizhe Chen, Yang Feng, Zeyu Zhao, Jin Hao, Bing Fang, Zuozhu Liu, and Youyi

Zheng. 2024. Leta: Tooth alignment prediction based on dual-branch latent encoding. *IEEE Transactions on Visualization and Computer Graphics*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

U.S. National Library of Medicine. 2025a. Pubmed: Medline retrieval system. https://pubmed.ncbi.nlm.nih.gov/. Accessed: 2025-05-20.

U.S. National Library of Medicine. 2025b. Unified medical language system (umls). https://www.nlm.nih.gov/research/umls. Accessed: 2025-05-20.

Guodong Wei, Zhiming Cui, Yumeng Liu, Nenglun Chen, Runnan Chen, Guiqing Li, and Wenping Wang. 2020. Tanet: towards fully automatic tooth arrangement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 481–497. Springer.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(1):58.

Huimin Xiong, Kunle Li, Kaiyuan Tan, Yang Feng, Joey Tianyi Zhou, Jin Hao, Haochao Ying, Jian Wu, and Zuozhu Liu. 2023. Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–432. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. *Preprint*, arXiv:2408.04138.

Kaiyan Zhang, Ning Ding, Biqing Qi, Sihang Zeng, Haoxin Li, Xuekai Zhu, Zhang-Ren Chen, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. https://github.com/TsinghuaC3I/UltraMedical.

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

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *Preprint*, arXiv:2502.04413.

Zhihe Zhao, Yanheng Zhou, and Yuxing Bai. 2020. *Orthodontics*. People's Medical Publishing House, Beijing. In Chinese.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, and 1 others. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.

A Dataset Construction Prompts & Templates

A.1 Definition Templates

Fig. 2 and Fig. 3 list the 50 instruction templates used to construct DEF questions from domain terms in English and Chinese.

A.2 Filtering Classification Prompt

Fig. 4 shows the prompt for classifying questions into *oral-related*, *non-oral*, or *insufficient* categories.

A.3 Evaluation and RAG Prompts

Fig. 5 shows the prompt formats used to evaluate different question types in zero-shot settings and the prompt format with RAG.

B Corpus Construction Details

B.1 MeSH Terms for PubMed Query

Fig. 6 is the list of 28 MeSH terms used to retrieve relevant dental articles from PubMed.

C Dataset Statistics and Visualizations

C.1 Distribution by Task and Subfield

Fig. 7 shows the distribution of DentalQA by task and subfield.

C.2 Answer Properties and Input Lengths

Fig. 5 shows the prompt formats used to evaluate different question types in zero-shot settings and the prompt format with RAG.

C.3 Supplementary Performance Figures

Extended plots (8, 9, 10, 11, 12, 13, 14, 16, 17) complementing Section 1, including per-model and per-task visual comparisons.

Prompt Template for DEF-EN

In the context of dental medicine, what does {noun} mean?

Please explain the meaning of {noun} in oral health sciences.

What is the definition of {noun} in the field of dental medicine?

Could you explain what {noun} refers to in dentistry?

In dentistry, how is the term {noun} understood?

How is {noun} defined within the dental field?

What does {noun} stand for in dental terminology?

From a dental medicine perspective, what does {noun} mean?

Can you describe what {noun} refers to in dental science?

Please clarify the concept of {noun} as used in oral medicine.

What is the basic idea behind {noun} in dental health?

In the context of dental science, what does {noun} refer to?

How would a dental professional define {noun}?

What is meant by the term {noun} in the context of oral health?

In dental practice, what does {noun} represent?

Could you provide a simple explanation of {noun} in dentistry?

Explain the meaning of {noun} from a dental perspective.

What is the significance of {noun} in oral medicine?

In dental terms, how is {noun} described?

How is the term {noun} used in the dental profession?

Can you explain the use of {noun} in oral health care?

What role does {noun} play in dentistry?

How should we interpret {noun} in dental research?

Describe what {noun} means in a dental setting.

In dental studies, what is meant by {noun}?

What does {noun} mean in terms of oral anatomy or pathology?

How do we understand {noun} in the context of dental education?

What is the dental meaning or implication of {noun}?

From a dental point of view, what does {noun} signify?

Can you define {noun} in simple dental terms?

How would you explain {noun} to a dental student?

What concept does {noun} convey in dentistry?

Could you elaborate on the meaning of {noun} in oral sciences?

How does dentistry define the concept of {noun}?

In oral medicine, how is {noun} typically understood?

What is the interpretation of {noun} in dental terminology?

Can you summarize the definition of {noun} in dentistry?

What does the term {noun} refer to in the context of oral biology?

Please give a definition of {noun} as it applies to dental health.

What is {noun} from the perspective of oral medicine?

How is the term {noun} applied in the field of dentistry?

What does {noun} mean when used in dental contexts?

What is considered the meaning of {noun} in oral healthcare?

Can you define {noun} as it is used in oral anatomy or treatment?

What understanding of {noun} is common in dental literature?

How is {noun} interpreted in the context of clinical dentistry?

Explain the medical significance of {noun} in dental care.

From an oral medicine viewpoint, what is {noun}?

Could you define the term {noun} as it applies to dentistry?

In oral health practice, what does {noun} stand for?

Figure 2: Templates for DEF-EN

Prompt Template for DEF-ZH

基于口腔医学相关领域，请解释一下{noun}的含义。
{noun}在口腔医学相关领域中是什么意思?
在口腔医学相关领域中，{noun}的基本含义是什么?
你能用简单的语言描述一下在口腔医学相关领域中的{noun}这个概念吗?
如何理解在口腔医学相关领域中{noun}这个概念?
在口腔医学中，如何定义{noun}?
什么是{noun}可以在口腔医学的背景下解释一下吗?
请描述一下在口腔医学中{noun}的定义。
口腔医学中使用的{noun}指的是什么?
{noun}的医学含义是什么，尤其是在口腔医学中?
从口腔医学的角度，如何解释{noun}?
口腔医学中{noun}指的是什么?
在口腔医学领域，{noun}具体是指什么?
{noun}在口腔医学中的具体含义是什么?
口腔医学领域中的{noun}是什么意思?
你能详细说明一下在口腔医学中，{noun}代表什么吗?
请解释一下口腔医学领域中的{noun}的基本概念。
在口腔医学领域中，如何理解{noun}?
{noun}在口腔医学领域有何特定含义?
口腔医学上{noun}的意义是什么?
在口腔医学的语境下，{noun}是指什么?
解释一下口腔医学中{noun}的基本意义。
{noun}在口腔医学中如何定义?
请给出{noun}在口腔医学方面的解释。
口腔医学中，如何理解{noun}这一概念?
从口腔医学的视角解释{noun}的意思。
在口腔医学背景下，{noun}是指什么?
什么是{noun}在口腔医学中的含义是什么?
在口腔医学中，如何解释{noun}?
口腔医学领域中，怎么理解{noun}?
你能解释下在口腔医学中{noun}的具体含义吗?
口腔医学相关领域中{noun}的含义是什么?
{noun}在口腔医学方面指什么?
口腔医学上，如何描述{noun}的定义?
请在口腔医学的背景下描述{noun}的意思。
在口腔医学相关学科中，如何定义{noun}?
{noun}的定义是什么，从口腔医学的角度来说?
能在口腔医学语境下解释一下{noun}的意思吗?
口腔医学中提到的{noun}是指什么?
在口腔医学领域中，{noun}一般指什么?
你能解释在口腔医学相关学科中{noun}的定义吗?
在口腔医学语境中，如何理解{noun}?
口腔医学中，{noun}代表的是什么概念?
{noun}在口腔医学里有何特别的含义?
你可以解释在口腔医学中提到的{noun}吗?
在口腔医学领域中，如何具体解释{noun}?
口腔医学学科中{noun}的意义是什么?
请描述一下{noun}在口腔医学中的作用。
在口腔医学的理解中，如何定义{noun}?
{noun}一词在口腔医学中是什么意思?

Figure 3: Templates for DEF-ZH

Prompt for Filtering Classification

You are an expert in dental education.
Classify the following item into ONE of the following three categories:
1. Insufficient: The input is clearly incomplete, malformed, or not a valid question/statement.
2. Oral: The question is related to dentistry or oral health (including clinical, basic, or preventive dental science like orthodontics, periodontics, oral surgery, even dental materials).
3. Non-oral: The question is clearly unrelated to dentistry or oral health.
Evaluation steps:
- FIRST check if the input is insufficient
- THEN determine if it's oral or non-oral
Return ONLY the label: 'insufficient', 'oral', or 'non oral'
Question:
{question}
Choices:
{choices}
Correct answer:
{answer}

Figure 4: Filtering Classification Prompt

Evaluation Prompts

EN:

MCQ: You are an experienced dentist. Based on your professional knowledge, read the following question and select the most appropriate answer. Only output the option letter.
Question: {question}
Options:
{options}
Answer:

DEF: You are a dentist. Please answer the following short medical question clearly and accurately.
Term: {question}
Answer:

OEQ: You are a dentist. Please answer the following short medical question clearly and accurately.
Question: {question}
Answer:

RAG Prompts: You are a dental medical expert. Please answer the following question based on the provided documents:
Context: {context}
Question: {query}

ZH:

MCQ: 你是一名经验丰富的口腔科医生。请根据专业知识，阅读以下问题并选择最合适的一个选项作答。仅输出选项字母。
问题：{question}
选项：
{options}
答案：

MAQ: 你是一名经验丰富的口腔科医生。请根据专业知识，阅读以下问题并选择所有正确的选项。仅输出选项字母。
问题：{question}
选项：\n{options}
答案：

DEF: 你是一名口腔科医生。请清晰准确地回答以下医学问题。
问题：{question}
答案：

OEQ: 你是一名口腔科医生。请清晰准确地回答以下医学问题。
问题：{question}
答案：

RAG Prompts: 你是一位口腔医疗专家，请结合以下资料回答问题：
资料：{context}
问题：{query}

Figure 5: Evaluation and RAG Prompts

MeSH Terms

The following Medical Subject Headings (MeSH) were employed in the literature retrieval process:
1. dental implantation[mesh]
2. dental materials[mesh]
3. dental occlusion[mesh]
4. dentistry[mesh]
5. drugs, dental[mesh]
6. endodontics[mesh]
7. geriatric dentistry[mesh]
8. mouth anatomy[mesh]
9. mouth diseases[mesh]
10. oral[mesh]
11. oral biology[mesh]
12. oral cancer[mesh]
13. oral diseases[mesh]
14. oral medicine[mesh]
15. oral pathology[mesh]
16. oral surgery[mesh]
17. orthodontics[mesh]
18. pediatric dentistry[mesh]
19. periodontics[mesh]
20. preventive dentistry[mesh]
21. prosthodontics[mesh]
22. radiology, oral[mesh]
23. stomatitis[mesh]
24. surgery, oral[mesh]

Figure 6: MeSH terms

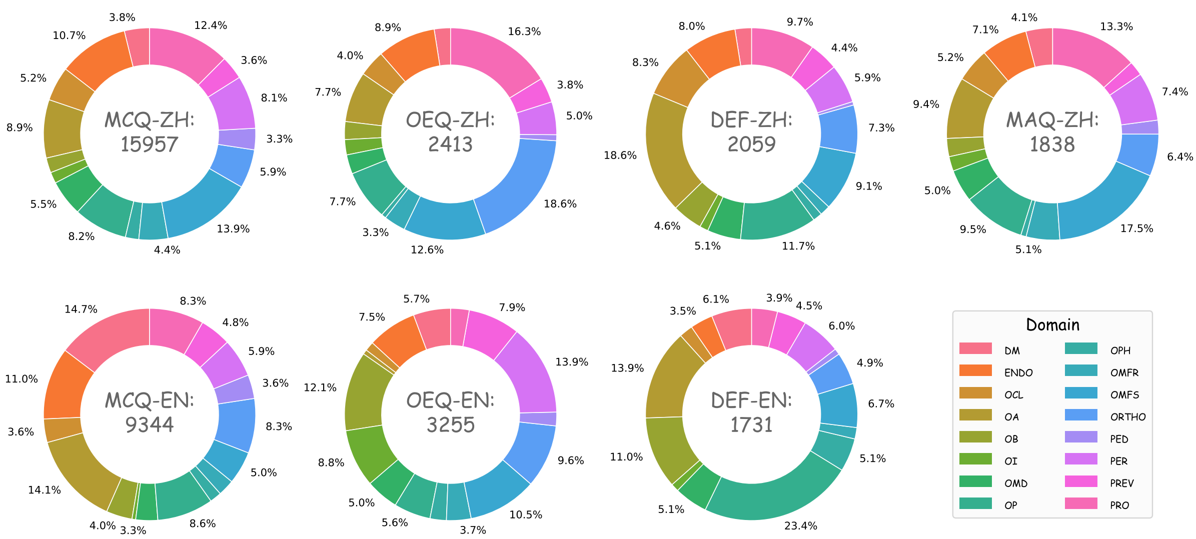

Figure 7: Distribution by Task and Subfield

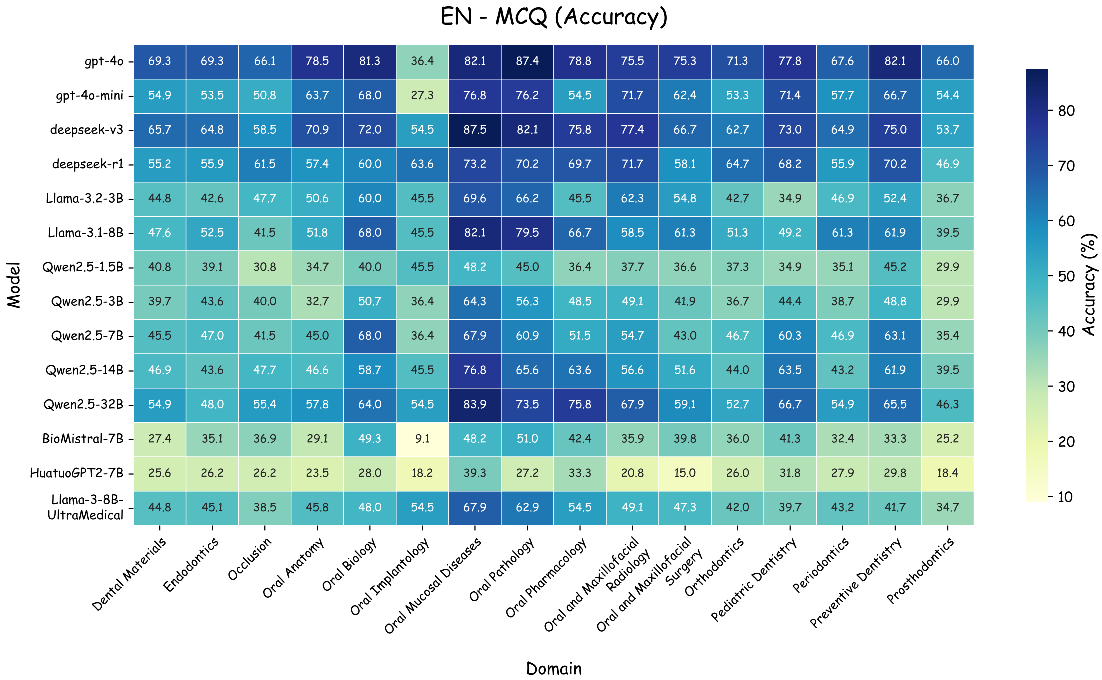

Figure 8: MCQ-EN-Accuracy

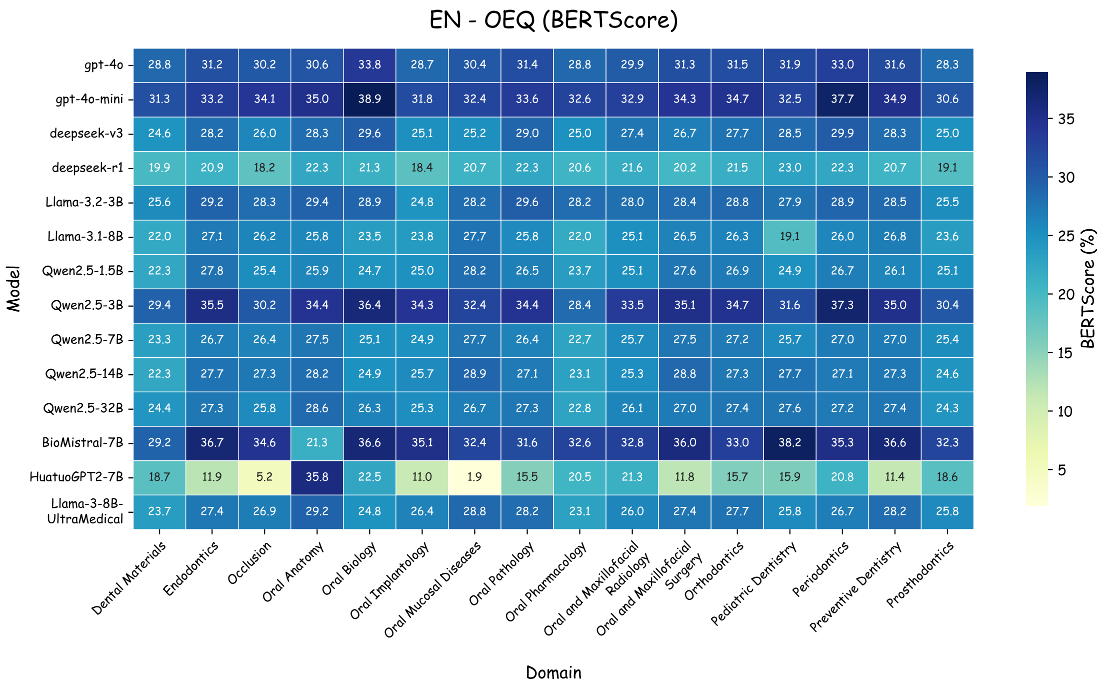

Figure 9: OEQ-EN-BERTScore

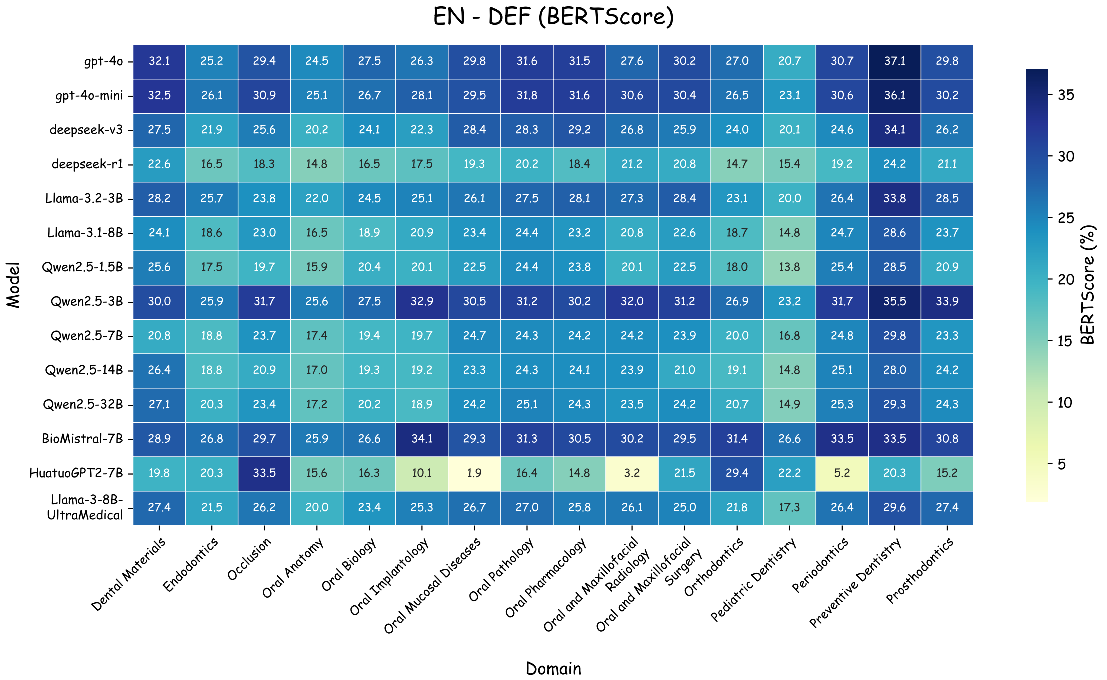

Figure 10: DEF-EN-BERTScore

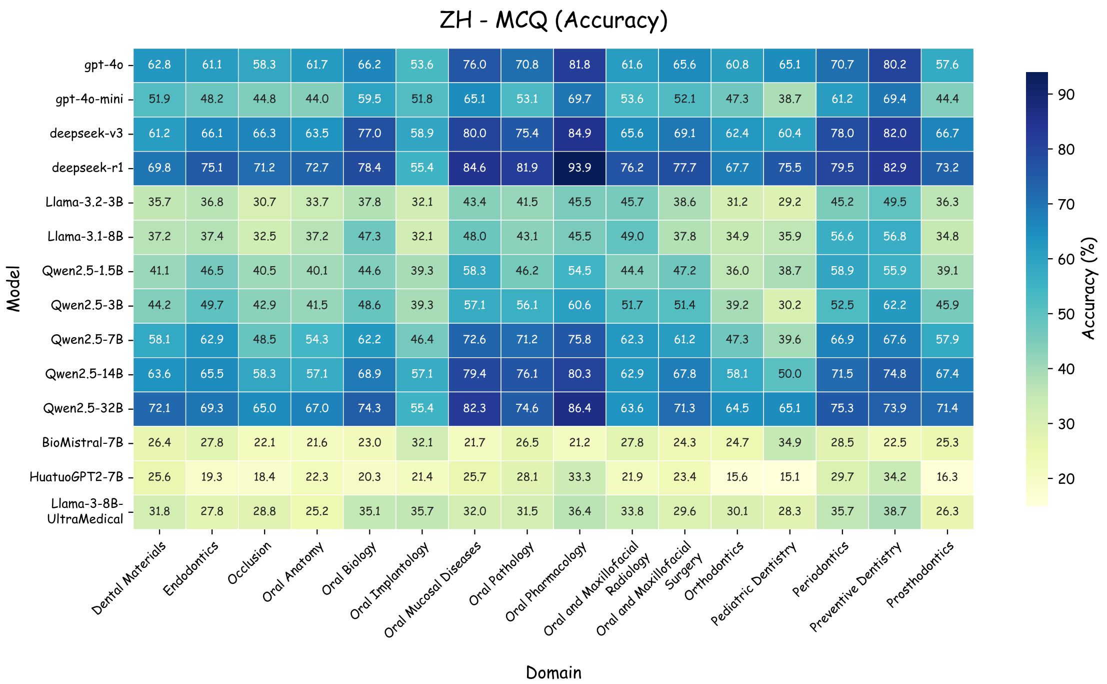

Figure 11: MCQ-ZH-Accuracy

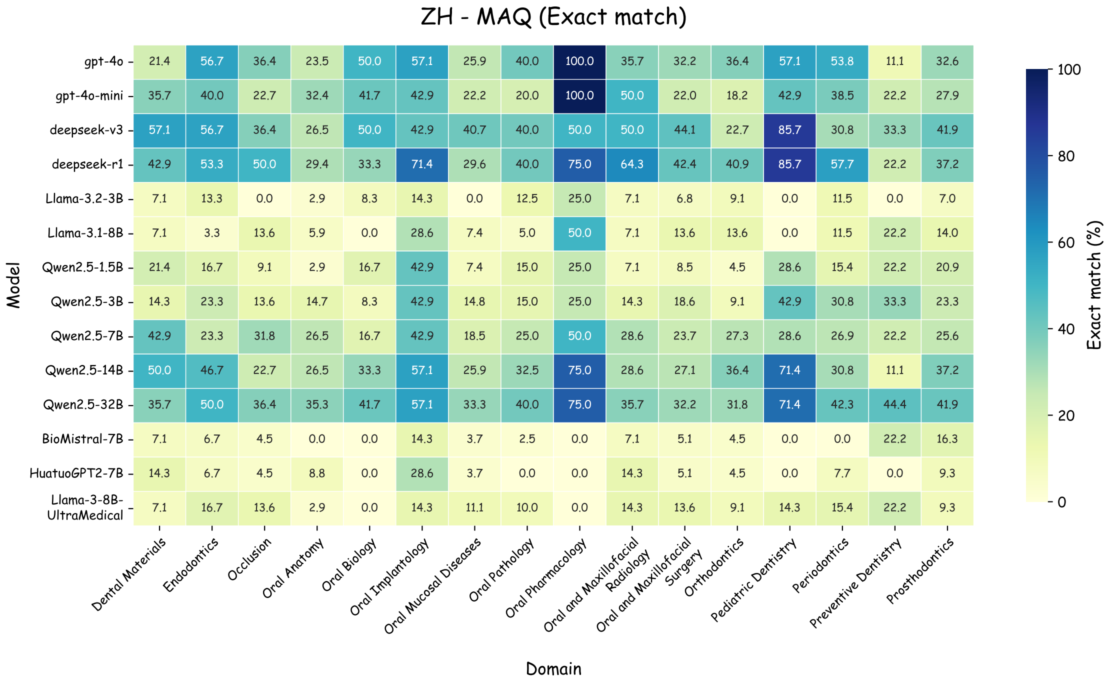

Figure 12: MAQ-ZH-Accuracy

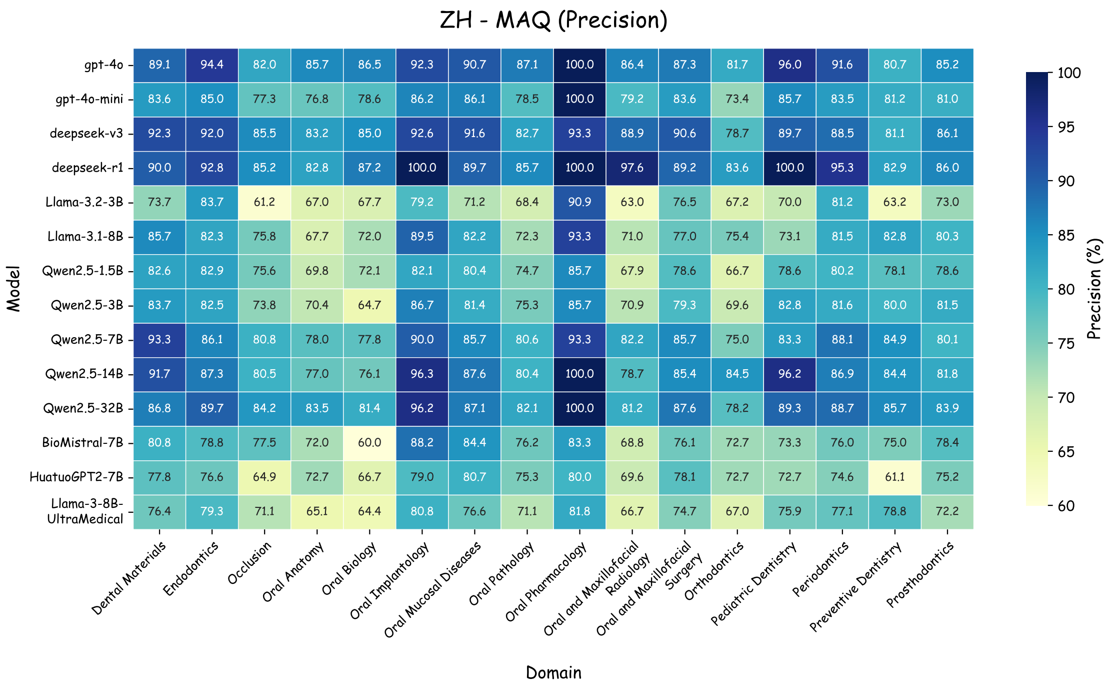

Figure 13: MAQ-ZH-Precision

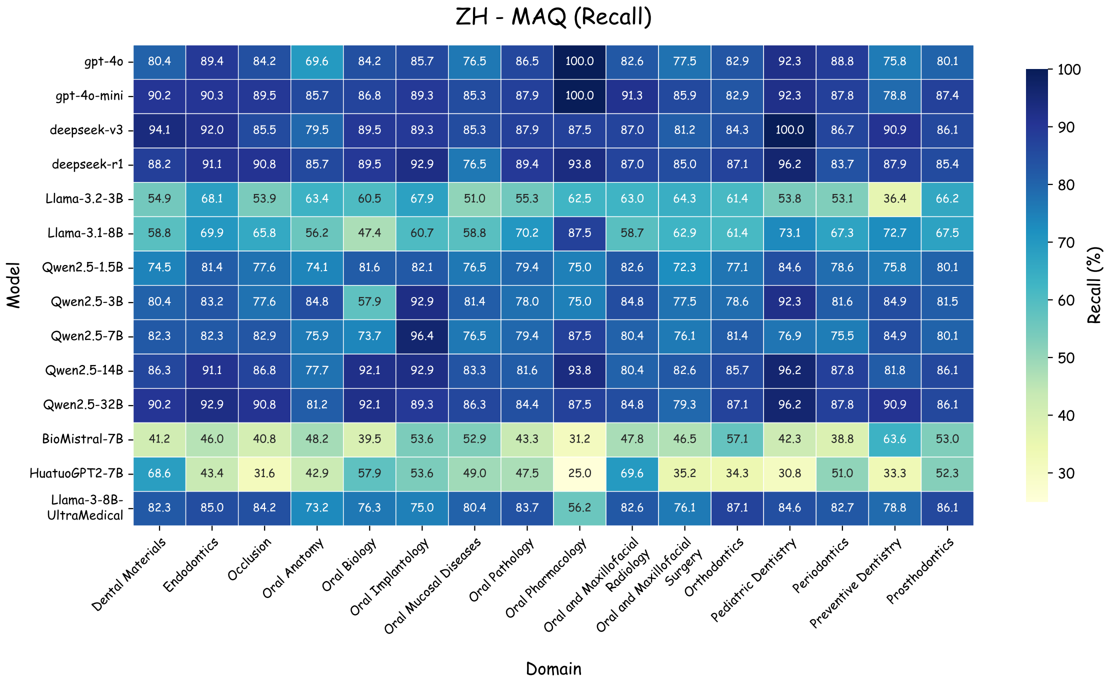

Figure 14: MAQ-ZH-Recall

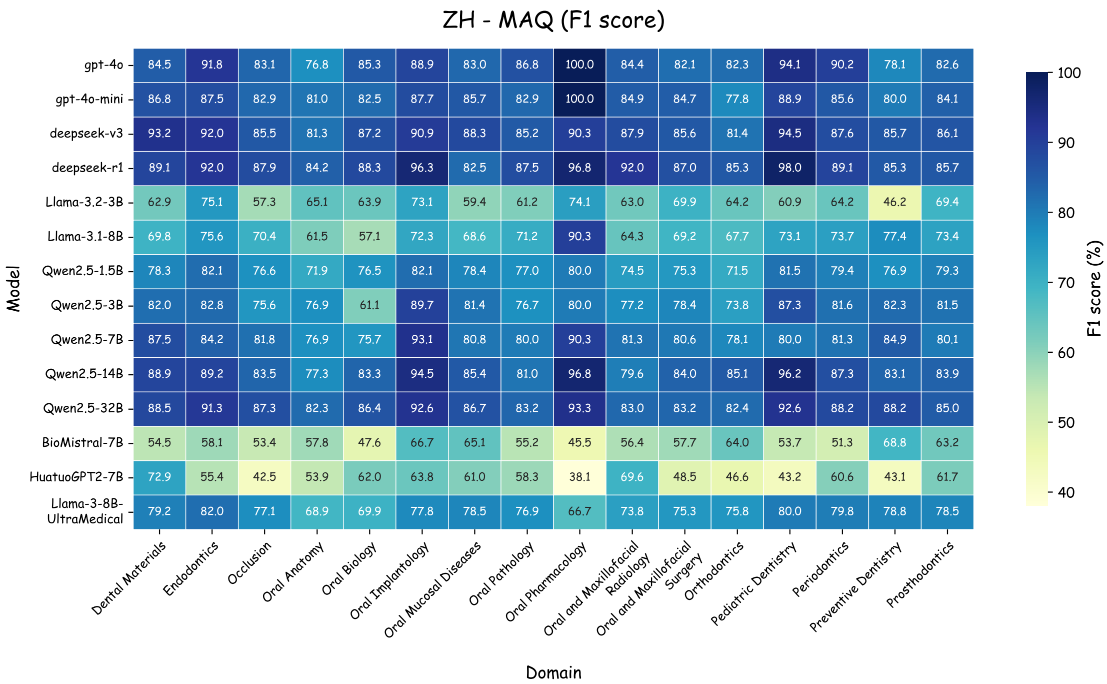

Figure 15: MAQ-ZH-F1

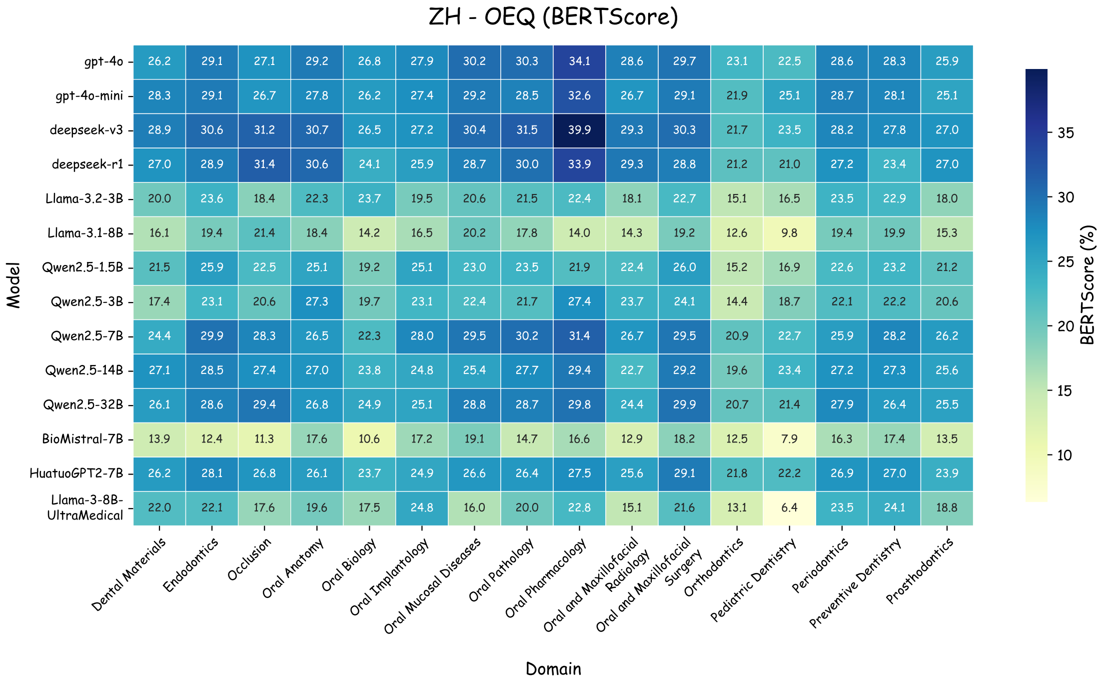

Figure 16: OEQ-ZH-BERTScore

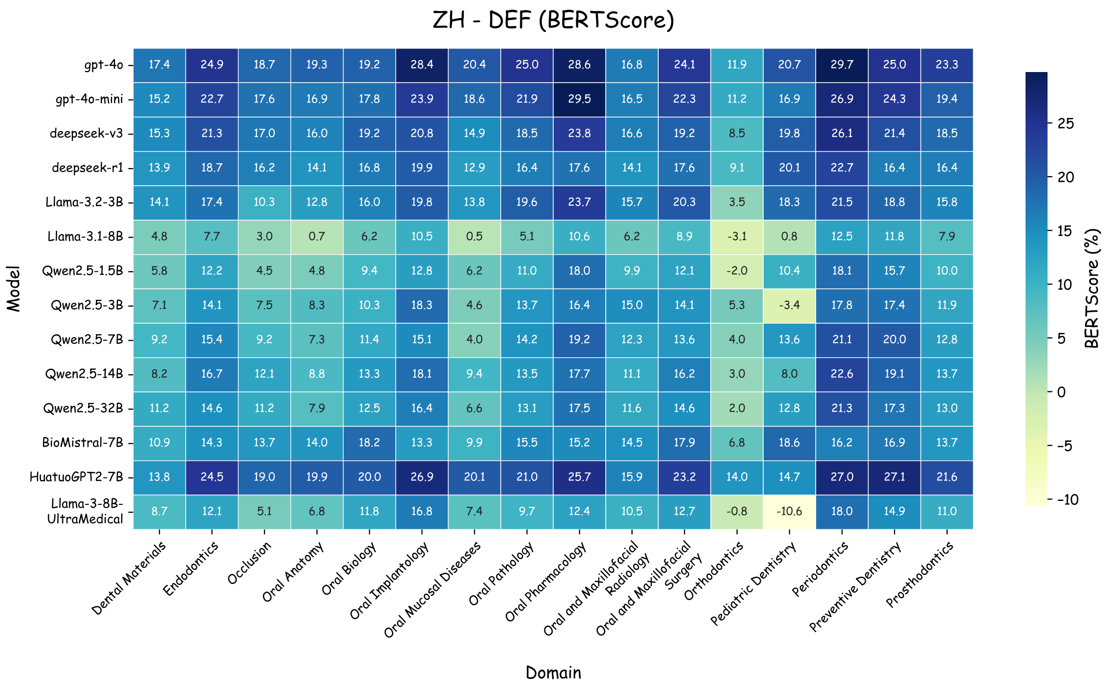

Figure 17: DEF-ZH-BERTScore

Table 2: The average length of questions and answers in OEQ and DEF

Question Format	Content	Mean	Median	Min	Max
OEQ-EN	answer	331.44	263	58	1941
OEQ-EN	question	147.90	108	20	1498
OEQ-ZH	answer	182.88	149	7	1321
OEQ-ZH	question	25.21	16	6	326
DEF-EN	answer	312.10	193	46	5249
DEF-EN	question	77.84	75	24	234
DEF-ZH	answer	59.79	52	2	254
DEF-ZH	question	22.63	22	6	64