FetalExtract-LLM: Structured Information Extraction from Free-Text Fetal MRI Reports Based on Privacy-Ensuring Open-weights Large Language Models

Mingxuan Liu[1†], Yijin Li[2†], Juncheng Zhu[3†], Hongjia Yang[1], Yiming Huang[4], Haoxiang Li[1], Yifei Chen[1], Xuguang Bai[1], Yi Liao[3], Haibo Qu[3], and Qiyuan Tian[1*]

[1] School of Biomedical Engineering, Tsinghua University
[2] School of Biological Science and Medical Engineering, Beihang University
[3] West China Second University Hospital, Sichuan University
[4] Department of Computer Science and Engineering, University of California San Diego

Abstract. Magnetic resonance imaging (MRI) is essential for evaluating fetal abnormality, providing superior soft tissue contrast to ultrasound. Radiologists write unstructured free-text reports documenting findings in MRI data, but this format hinders retrospective access and secondary data usage. Structuring these reports facilitates cohort identification for research, supports the development of medical AI systems, and enables educational applications. However, the diverse linguistic styles and inconsistent formatting of free-text reports pose significant challenges for structured information extraction. Despite the capabilities of proprietary large language models (LLMs) like GPT, Claude and Gemini in automating extraction via zero-shot prompting, their cost and privacy limitations prohibit their use. To address these problems, this study introduces FetalExtract-LLM, the first fetal MRI report structured information extraction model developed through instruction tuning of an open-weights, privacy-preserving LLM. Experimental results show that FetalExtract-LLM achieves 0.987 average F1-score in per-key matching, on par with proprietary models, with additional metrics of 1.000 JSONable accuracy, 0.966 domain compliance accuracy, 0.829 positive finding accuracy, and 0.770 exact matching accuracy.

Keywords: Fetal MRI · Large language models · Instruction tuning · Information extraction · Free-text reports

1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive diagnostic tool for further evaluating suspected fetal abnormalities by prenatal ultrasound [14,15]. With

† Equal contribution.
* Corresponding author: qiyuantian@tsinghua.edu.cn

significant technical advancements over recent decades, fetal MRI offers distinct advantages including superior soft tissue contrast and expanded field of view [12]. For diagnosis, radiologists write a detailed report summarizing their findings after analyzing fetal MR images. In recent publications, the Fetal Task Force of the European Society of Paediatric Radiology (ESPR) [23,22] recommends that the fetal MRI report should begin with gestational age (GA), technical parameters, and the clinical rationale for MRI beyond ultrasound. It should then concisely review all visualized anatomy, including ventricles, midline structures, cortical maturation, posterior fossa, and any hemorrhagic or ischemic lesions for the brain, and the face and neck, thorax, abdomen and pelvis, spine, limbs, and placenta for the rest of the body. Each region should be assessed against gestational norms and ultrasound findings, with abnormal measurements or signals quantified, and associated anomalies identified.

Fetal MRI reports are highly informative but stored as free text with wide stylistic variations across radiologists. This heterogeneity impedes archival case querying and high-quality AI training label generation [4,9,25]. Structured reporting overcomes the aforementioned limitations by converting narrative descriptions into queryable data fields through standardized formats. This is crucial for large-scale data analysis, enabling rapid assembly of patient cohorts for research and the integration of imaging findings with other clinical and pathological data to create unified datasets [9]. Most importantly, structured data provides high-quality and consistent labels essential for developing medical AI systems [19,7]. For example, Bai et al. trained Chest-OMDL, a Y-Mamba network, on 25,692 chest CT datasets using disease labels extracted from text reports by RadBERT, which outperformed fully-supervised baselines in multi-disease detection [2]. Nowak et al. converted ICU chest X-ray reports into structured labels using a transformer to built a decision-support system [18]. Additionally, an abdominal CT study used information extraction of lesion phrases from free-text reports plus multi-instance learning over organ-segmented slices to train a five-organ anomaly detector [24].

Currently, structured information extraction relies on three main methods: rule-based systems [11], supervised deep learning [17], and zero-shot prompting of LLMs [9,4]. Rule-based methods are based on manually defined rules to extract terms and detect negations, but they fail when tasks require more complex analysis than simple mention detection. Supervised learning relies on representative models like BERT and its variants, but suffers from dependence on large, manually annotated datasets. When dealing with large-scale data involving complex textual reports, the training costs become prohibitive. This challenge is particularly pronounced with fetal MRI reports, which typically contain substantially more information than the chest X-ray [17] or breast ultrasound reports [4] studied in previous studies. Since OpenAI released ChatGPT (a chat generative pretrained transformer) on November 30, 2022 , studies have systematically validated pretrained LLMs for structured information extraction. Well-designed prompts enable high-precision extraction without retraining or fine-tuning. For example, Grothey's study demonstrate that GPT-4 achieves >97% accuracy

in pathology report extraction [9]. However, leading proprietary LLM providers (e.g., OpenAI, Anthropic, Google, and X) typically restrict usage and do not provide access to model weights, preventing local deployment. Meanwhile, stringent data protection laws regulate external processing of protected medical information. Consequently, proprietary LLMs pose persistent challenges in terms of cost and data privacy.

To address these challenges, we developed FetalExtract-LLM, a specialized model for extracting structured information from fetal MRI reports, using a privacy-preserving, open-weights LLM for secure in-house deployment. Specifically, we constructed a training dataset using in-context learning with the larger open-source model Qwen-3 235B-A22B [26], then fine-tuned DeepSeek-R1-0528-Qwen3-8B [5] via instruction tuning [27]. On a clinician-annotated subset, FetalExtract-LLM achieves accuracy comparable to SOTA closed-source models, including GPT-4.1 [21], Claude Sonnet 4 [1], and Gemini 2.5 Pro Preview [8].

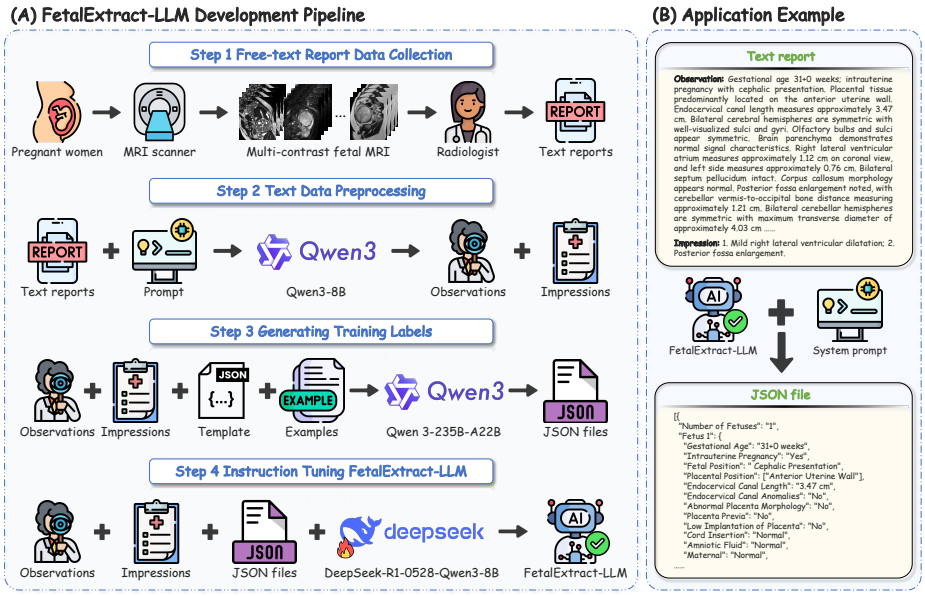

Fig. 1. (A) FetalExtract-LLM Development Pipeline. The workflow comprises four stages: (1) retrospectively collect fetal MRI reports; (2) preprocess reports with Qwen3-8B to extract observations and impressions; (3) using predefined JSON schema and ten curated examples, employ Qwen-3-235B-A22B to generate training corpus; (4) fine-tune DeepSeek-R1-0528-Qwen3-8B on this corpus to obtain FetalExtract-LLM. (B) Application Example of FetalExtract-LLM.

Table 1. Keys and Values of the Structured JSON Object.

Category	Key	Value
Basic Information	Number of Fetuses	Integer (1, 2, 3...)
	Gestational Age	"m+n weeks" Format
	Intrauterine Pregnancy	Yes/No
	Fetal Position	Cephalic Presentation/Breech Presentation/Transverse Presentation/Other Position
Placental & Maternal	Placental Position	Anterior Uterine Wall/Posterior Uterine Wall/Left Lateral Uterine Wall/Right Lateral Uterine Wall/Fundal
	Endocervical Canal Length	"x cm" Format
	Endocervical Canal Anomalies	Yes/No
	Abnormal Placenta Morphology	Yes/No
	Placenta Previa	Yes/No
	Low Implantation of Placenta	Yes/No
	Cord Insertion	Normal/Marginal/Velamentous
	Amniotic Fluid	Normal/Polyhydramnios/Oligohydramnios
	Maternal	Normal/Abnormal
Brain Structure	Cerebral Hemisphere	Normal/Abnormal
	Cerebral Parenchyma	Normal/Abnormal
	Widened Extracerebral Space	Yes/No
	Central Sulcus	Visible/Not Visible
	Cerebral Sulci and Gyri	Visible/Not Visible
	Olfactory Sulcus and Olfactory Bulb	Normal/Abnormal
	Choroid Plexus Cyst	None/In the left lateral ventricle/In the right lateral ventricle/In both lateral ventricles
	Subependymal Cyst	None/Adjacent to the left lateral ventricle/Adjacent to the right lateral ventricle/Adjacent to both lateral ventricles
	Choroidal Fissure Cyst	None/In the left lateral ventricle/In the right lateral ventricle/In both lateral ventricles
	Lateral Ventricles	Normal/Left lateral ventriculomegaly/Right lateral ventriculomegaly/Bilateral ventriculomegaly/Other Abnormalities
	Third Ventricle	Normal/Dilated/Stenosis
	Fourth Ventricle	Normal/Dilated/Stenosis
	Septum Pellucidum	Normal/Abnormal
	Optic Chiasm	Visible/Not Visible
	Corpus Callosum	Normal/Abnormal
	Enlarged Cisterna Magna	Yes/No
	Cerebellar Hemispheres	Normal/Abnormal
	Cerebellar Vermis	Normal/Abnormal
	Primary Fissure	Visible/Not Visible
	Brainstem	Normal/Abnormal
	Brain Midline Shift	Yes/No
	Skull	Normal/Abnormal
Spine & Spinal Cord	Scoliosis	Yes/No
	Vertebrae	Normal/Abnormal
	Spinal Canal	Normal/Abnormal
	Conus Medullaris Position	Normal/High/Low
	Thickening of Filum Terminale	Yes/No
Thorax & Abdomen	Heart	Normal/Abnormal
	Lungs	Normal/Abnormal
	Abdomen	Normal/Abnormal
	Liver	Normal/Abnormal
	Kidneys	Normal/Abnormal
	Spleen	Normal/Abnormal
	Stomach	Normal/Abnormal
	Intestines	Normal/Abnormal
	Rectum	Normal/Abnormal
	Bladder	Normal/Abnormal

2 Materials and Methods

2.1 Problem Formulation

Formally, let x denote a fetal MRI report containing textual descriptions of fetal anatomy and study metadata. Our goal is to transform each report x into a structured JSON object organized into five predefined categories:

$$y = \left\{ \begin{array}{c} \text{Basic Information} \quad : \mathbf{B} \\ \text{Placental \& Maternal} : \mathbf{P} \\ \text{Brain Structure} \quad : \mathbf{C} \\ \text{Spine \& Spinal Cord} \ : \mathbf{S} \\ \text{Thorax \& Abdomen} \ : \mathbf{T} \end{array} \right\} \tag{1}$$

Each category $\mathbf{Z} \in \{\mathbf{B}, \mathbf{P}, \mathbf{C}, \mathbf{S}, \mathbf{T}\}$ is defined as a dictionary:

$$\mathbf{Z} = \{(k_1, v_1), (k_2, v_2), \ldots, (k_n, v_n)\}$$

where k_i represents standardized anatomical attributes within a category, and v_i are the corresponding values extracted from the report.

As summarized in Table 1, our proposed JSON schema was derived from three authoritative resources: (i) the standardized, structured fetal MRI reporting template released in 2024 by the Fetal Task Force of the European Society of Paediatric Radiology [25]; (ii) the U.S. National Library of Medicine's Medical Subject Headings (MeSH) [13]; and (iii) the Merriam-Webster Medical Dictionary [16]. The final schema consists of 50 keys organized into five thematic categories.

2.2 FetalExtract-LLM Development Pipeline

Free-text Report Data Collection Between October 2015 and December 2024, the radiology department of the West China Second University Hospital generated 6,032 fetal MRI reports for pregnant women scanned on one 3.0 T system (Siemens MAGNETOM Skyra) and two 1.5 T systems (Philips Achieva and United Imaging uMR 570). For model development, the corpus was randomly split into three mutually exclusive subsets: 5,437 reports for training, 600 for validation, and 265 for testing. The training and validation sets were annotated with Qwen-3-235B-A22B, whereas the test set was independently labeled by experienced clinicians. Since the evaluation relies on proprietary LLMs accessed through the external API, all reports were fully de-identified to safeguard patient privacy.

Text Data Preprocessing A typical fetal MRI report includes five sections: 1) patient and examination details; 2) scanning sequences and methodology; 3) observation, detailing the observed features of fetal anatomy and maternal structures; 4) impression, providing a concise interpretation of findings and abnormalities; and 5) clinical recommendations. The observation and impression sections contain critical clinical information in an unstructured narrative format. Therefore, we focus on extracting and standardizing information from these two sections. As shown in Fig. 1(A), we employed the Qwen3-8B [26] to process and restructure these sections, ensuring consistency and removing redundant information. The specific prompt used for this extraction is as follows:

Prompt for Fetal MRI Report Processing:
Please organize the following fetal MRI report. The report is divided into "Observation" and "Impression" sections. Organization requirements are as follows:

1. Remove redundant modifiers and scanning sequence-related descriptions.
2. Remove all clinical recommendation content.
3. Standardize gestational age format: prioritize using the corrected gestational age, formatted as "Gestational age: xx+xx weeks"; if no corrected gestational age, use the last menstrual period gestational age.
4. The organized content should be divided into "Observation" and "Impression" sections:
   - Observation should include in order (skip if no content): gestational age, intrauterine pregnancy, fetal position, placental location, cervical canal length, fetal brain-related descriptions, fetal body-related descriptions. Separate items with semicolons.
   - Impression should only retain pathology-related descriptions, removing gestational age and non-pathological content that repeats the Observation, with content listed in points; if no abnormalities, uniformly write "No abnormalities detected".
5. For multiple pregnancies, output in the following format: Fetus 1: Observation: ..., Impression: ...; Fetus 2: Observation: ..., Impression: ....

Original report text follows: {Original reports}

Generating Training Labels We used the open-weights LLM Qwen3-235B-A22B [26] to generate training labels. Leveraging its in-context learning (ICL) capabilities [6], we designed a prompt following previous research [3], incorporating 10 curated examples to demonstrate the expected JSON format. The specific prompt for generating training labels is:

Prompt for Generating Training Labels:
Role Description
You are a medical assistant specializing in the field of obstetrics and gynecology. Your task is to extract structured information from provided fetal MRI reports and output JSON objects in the specified format.
Task Requirements

1. Based on the input text report content and JSON structuring requirements, output JSON objects in the specified format.
2. Only return the required keys and their corresponding values, without any additional text explanations.
3. All JSON punctuation must use English half-width format.
4. For keys that are specified to only select from a designated range, values must be strictly chosen from the specified options, and no other content is allowed.
5. If a key is not mentioned in the report, uniformly fill its value as "N/A", unless specific requirements are given in the "JSON Output Requirements" section.
6. For multiple pregnancies, supplement corresponding structural content according to "Fetus 1", "Fetus 2", "Fetus 3" sequentially.

Input Format
You will receive the following three parts of input:

1. Detailed requirements for structured JSON output and optional values for each key: {JSON template}
2. 10 examples provided by professional obstetricians and gynecologists: {Examples of input and output}
3. Fetal MRI report: {Observation + Impression}

Instruction Tuning FetalExtract-LLM We fine-tuned DeepSeek-R1-0528-Qwen3-8B [5] using LLaMA-Factory [28] with LoRA+ adapters [10] for instruction tuning on our fetal MRI report corpus. DeepSeek-R1-0528-Qwen3-8B is produced by distilling the chain-of-thought from DeepSeek-R1-0528 into the Qwen3-8B base model; on the AIME-2024 benchmark it sets the open-source state of the art, exceeding Qwen3-8B by 10.0 pp [5]. LoRA+ extends LoRA by assigning component-specific learning rates, which accelerates convergence and improves final accuracy. All linear layers were adapted with rank = 8, $\alpha = 16$, dropout = 0, and a LoRA+ learning-rate ratio of 16. Training was carried out in bf16 with a 2,048-token context window for 4 epochs, using a micro-batch size of 4 and no gradient accumulation. The model was optimized using the AdamW optimizer with an initial learning rate of 2×10^{-4}, cosine learning rate decay, no warm-up phase, and global gradient-norm clipping at 0.3. Instruction tuning was conducted on eight NVIDIA A800 80GB GPUs.

3 Experiments and Results

3.1 Evaluation Metrics

Following a previous study [4], we assessed model performance with two metric suites: per-key and per-report matching. For per-key matching, we tested the average F1 score over all 50 keys. For per-report matching, we used the following four metrics: (1) JSONable Accuracy (JA), proportion of LLM's outputs that can be parsed into a valid list of dictionaries. (2) Domain-Compliance Accuracy (DCA), proportion of cases in which every key's value lies within its predefined domain. (3) Positive-Finding Accuracy (PFA), for positive cases, the fraction of positive-related keys predicted correctly. (4) Exact-Match Accuracy (EMA): proportion of outputs which, once converted to JSON, match the ground truth exactly across all 50 keys.

Table 2. Comparative Performance Evaluation of Proprietary LLMs, Open-Weight LLMs, and FetalExtract-LLM.

Models	Average F1-score	JA	DCA	PFA	EMA
Proprietary LLMs					
GPT-4.1	0.803	1.000	0.989	0.800	0.098
Claude Sonnet 4	0.965	1.000	0.996	0.900	0.702
Gemini 2.5 Pro Preview	0.949	1.000	0.996	0.936	0.502
Open-weights LLMs					
Qwen3-235B-A22B	0.954	1.000	0.955	0.836	0.732
NuExtract-2-8B	0.580	1.000	0.272	0.757	0.000
Qwen3-8B	0.672	0.898	0.049	0.771	0.004
DeepSeek-R1-0528-Qwen3-8B	0.927	0.996	0.962	0.786	0.498
Open-weights LLMs + Instruction Tuning with Simulation Data					
Qwen3-8B	0.887	1.000	0.722	0.719	0.120
DeepSeek-R1-0528-Qwen3-8B	0.945	1.000	0.693	0.655	0.121
Open-weights LLMs + Instruction Tuning with Clinical Data					
Qwen3-8B	0.987	1.000	0.966	0.806	0.754
FetalExtract-LLM	0.987	1.000	0.966	0.829	0.770

3.2 Comparative Models

We compared four LLM categories: (1) Proprietary LLMs: OpenAI's GPT-4.1 (April 14, 2025) [21], Anthropic's Claude Sonnet 4 (May 22, 2025) [1], and Google's Gemini 2.5 Pro Preview (June 5, 2025) [8]. (2) Open-Source LLMs:

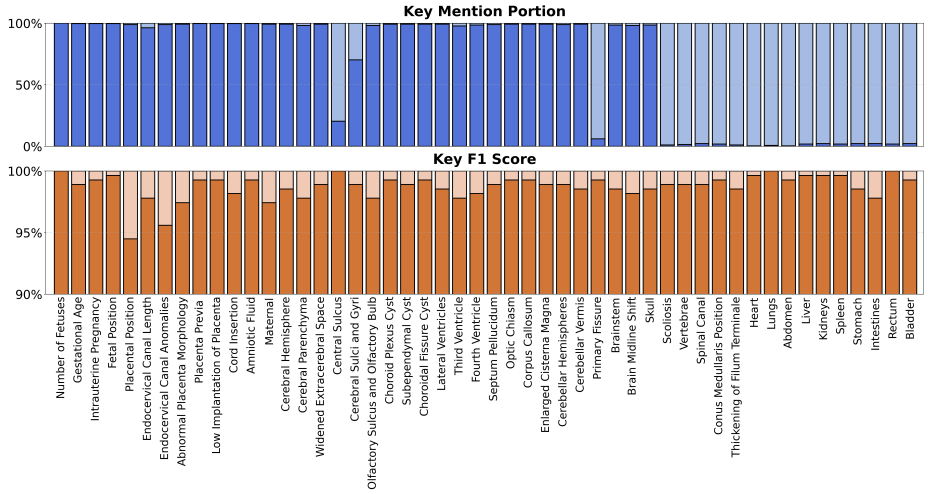

Fig. 2. The proportion of reports in which each key is mentioned (top), and the F1-score of FetalExtract-LLM for each key (bottom).

Qwen3-235B-A22B [26], Qwen3-8B [26], DeepSeek-R1-0528-Qwen3-8B [5], and NuExtract-2-8B [20], a model designed for structured information extraction. (3) LLMs Fine-tuned with Simulated Data: We used Qwen3-235B-A22B to generate 6,000 simulated pairs from 10 examples, then instruction-tuned Qwen3-8B and DeepSeek-R1-0528-Qwen3-8B for comparison. (4) LLMs Fine-tuned with Clinical Data: DeepSeek-R1-0528-Qwen3-8B fine-tuned with clinical data (FetalExtract-LLM), and Qwen3-8B fine-tuned under identical conditions to assess base model impact.

3.3 Results

The performance comparison between FetalExtract-LLM and the baseline models is presented in Table 2. The experimental results yield the following findings: (1) In terms of both average F1-score and EMA, which best capture overall model performance, FetalExtract-LLM achieves the highest scores (0.987 and 0.770), outperforming all proprietary LLMs. NuExtract-2-8B, which is specifically designed for structured information extraction, records the lowest F1-score, likely due to the lack of domain-specific pretraining on fetal medical corpus. (2) Except for Qwen3-8B and DeepSeek-R1-0528-Qwen3-8B without fine-tuning, all models reach 1.000 JA. (3) Regarding the DCA metric, proprietary LLMs Claude Sonnet 4 and Gemini 2.5 Pro Preview exhibit the best performance (0.996), with FetalExtract-LLM closely following (0.966). This demonstrates their strong ability to follow prompt instructions and generate the desired output formats. (4) For the PFA metric, there is a substantial gap between FetalExtract-LLM and the top-performing Gemini 2.5 Pro Preview (0.829 vs. 0.936). This may be attributed

to the greater difficulty in extracting relevant keys in positive cases, which often requires extensive contextual understanding. (5) Fine-tuning DeepSeek-R1-0528-Qwen3-8B with simulated data leads to improvements in F1-score and JA. However, declines are observed in DCA, PFA, and EMA, indicating that simulated text generated by Qwen3-235B-A22B with only 10 examples lacks representativeness and cannot effectively substitute for real clinical data. (6) Compared to Qwen3-8B fine-tuned with clinical data, FetalExtract-LLM demonstrates superior PFA (0.829 vs. 0.806) and EMA (0.770 vs. 0.754), suggesting that using a stronger base model for fine-tuning yields better performance.

Figure 2 further illustrates the frequency with which each key appears in the reports, as well as the F1-score for each key predicted by FetalExtract-LLM. Notably, only the F1-score for Placental Position falls below 95%, likely because the placenta may occur in multiple locations, resulting in a multi-label classification problem and greater classification complexity. But above discrepancy only affect anatomical descriptions, as high-risk conditions such as placenta previa are accurately determined through separate fields, preventing clinical decision-making bias. Additionally, the F1-score for Endocervical Canal Anomalies is relatively lower. This is due to the diverse manifestations of anomalies —such as elongation, shortening, or dilation —and the fact that the term "anomaly" may not be explicitly mentioned in the text. Instead, the LLM must infer abnormalities based on the length of the endocervical canal or contextual information. Overall, pretraining the LLM on a richer corpus within obstetrics and gynecology, or integrating it with medical knowledge bases, may further enhance its performance.

4 Conclusion

We present FetalExtract-LLM, the first privacy-preserving open-weight large language model for structured information extraction from fetal MRI reports. Through instruction tuning of DeepSeek-R1-0528-Qwen3-8B on 6,032 clinical reports, our model achieves competitive performance with proprietary LLMs while enabling secure in-house deployment. Future work will incorporate multi-center data and rare cases, enhancing performance via pretraining on richer obstetrics and gynecology corpora or medical knowledge base integration. Additionally, introducing rule-based systems and expert evaluations will quantify LLM's improvement over traditional methods and validate clinical decision reliability.

Acknowledgments This work was supported by the Scientific Research Project of Sichuan Medical Association (Grant No. 2024HR130), Science and Technology Department of Sichuan Province (Grant No. 25SYSX0255), Tsinghua University Startup Fund, and Tsinghua University Dushi Program (grant numbers 20241080026, 20251080056).

References

1. Anthropic: Introducing claude 4. `https://www.anthropic.com/news/introducing-claude-4` (2025), accessed: 2025-06-17

2. Bai, X., Liu, M., Chen, Y., Yang, H., Tian, Q.: Chest-OMDL: Organ-specific multidisease detection and localization in chest computed tomography using weakly supervised deep learning from free-text radiology report. In: Medical Imaging with Deep Learning (2025), https://openreview.net/forum?id=ns6nq592HX
3. Balasubramanian, J.B., Adams, D., Roxanis, I., de Gonzalez, A.B., Coulson, P., Almeida, J.S., García-Closas, M.: Leveraging large language models for structured information extraction from pathology reports (2025), https://arxiv.org/abs/2502.12183
4. Chen, Y., Yang, H., Pan, H., Siddiqui, F., Verdone, A., Zhang, Q., Chopra, S., Zhao, C., Shen, Y.: Burextract-llama: An llm for clinical concept extraction in breast ultrasound reports. In: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. p. 5358. MCHM'24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3688868.3689200
5. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
6. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., et al.: A survey on in-context learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1107–1128 (2024)
7. Flanders, A.E., Wang, X., Wu, C.C., Kitamura, F.C., Shih, G., Mongan, J., Peng, Y.: The evolution of radiology image annotation in the era of large language models. Radiology. Artificial Intelligence **7**(4), e240631 (2025). https://doi.org/10.1148/ryai.240631, https://doi.org/10.1148/ryai.240631
8. Google: Gemini 2.5 pro: Access google's latest preview ai model. https://blog.google/products/gemini/gemini-2-5-pro-updates/ (2025), accessed: 2025-06-17
9. Grothey, B., Odenkirchen, J., Brkic, A., Schömig-Markiefka, B., Quaas, A., Buettner, R., Tolkach, Y.: Comprehensive testing of large language models for extraction of structured data in pathology. Communications Medicine **5** (2025), https://api.semanticscholar.org/CorpusID:277464469
10. Hayou, S., Ghosh, N., Yu, B.: LoRA+: Efficient low rank adaptation of large models. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 17783–17806. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/hayou24a.html
11. Howell, L., Zarei, A., Wah, T.M., Chandler, J.H., Karthik, S., Court, Z., Ng, H., McLaughlan, J.R.: Radex: a rule-based clinical and radiology data extraction tool demonstrated on thyroid ultrasound reports. European Radiology pp. 1–12 (2025)
12. Li, Y., Liu, M., Yang, H., Li, H., Bai, X., Liao, Y., Qu, H., Tian, Q.: Anatomy-guided test-time adaptation for automated fetal brain MRI morphometry. In: Medical Imaging with Deep Learning - Short Papers (2025), https://openreview.net/forum?id=iLBipDelQu
13. Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association **88**(3), 265 (2000)
14. Liu, M., Li, H., Li, Z., Yang, H., Zheng, J., Qu, H., Tian, Q.: Unsupervised Fetal Brain MRI Quality Assessment based on Orientation Prediction Uncertainty. In: 2025 OHBM Annual Meeting. Brisbane, Australia (Jun 2025), https://hal.science/hal-04974115

15. Liu, M., Li, H., Li, Z., Yang, H., Zheng, J., Zhang, X., Tian, Q.: Image Quality Assessment using an Orientation Recognition Network for Fetal MRI. In: 2024 ISMRM & ISMRT Annual Meeting & Exhibition. Singapore, Singapore (May 2024), https://hal.science/hal-05039081
16. Merriam-Webster, I.: Merriam-webster's medical dictionary. Merriam-Webster (1995)
17. Nowak, S., Biesner, D., Layer, Y., Theis, M., Schneider, H., Block, W., Wulff, B., Attenberger, U., Sifa, R., Sprinkart, A.: Transformer-based structuring of free-text radiology report databases. European radiology **33**(6), 4228–4236 (2023)
18. Nowak, S., Schneider, H., Layer, Y.C., Theis, M., Biesner, D., Block, W., Wulff, B., Attenberger, U.I., Sifa, R., Sprinkart, A.M.: Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. European radiology **34**(5), 2895–2904 (2024)
19. Nowak, S., Wulff, B., Layer, Y.C., Theis, M., Isaak, A., Salam, B., Block, W., Kuetting, D., Pieper, C.C., Luetkens, J.A., Attenberger, U., Sprinkart, A.M.: Privacy-ensuring open-weights large language models are competitive with closed-weights gpt-4o in extracting chest radiography findings from free-text reports. Radiology **314**(1), e240895 (2025). https://doi.org/10.1148/radiol.240895, https://doi.org/10.1148/radiol.240895, pMID: 39807977
20. NuMind AI: Nuextract-2.0-8b. https://huggingface.co/numind/NuExtract-2.0-8B (2024), accessed: 2025-06-17
21. OpenAI: Introducing gpt-4.1 in the api. https://openai.com/index/introducing-gpt-4-1/ (1 2025), accessed: 2025-06-17
22. Papaioannou, G., Caro-Domínguez, P., Klein, W.M., Garel, C., Cassart, M.: Indications for magnetic resonance imaging of the fetal body (extra-central nervous system): recommendations from the european society of paediatric radiology fetal task force. Pediatric Radiology **53**(2), 297–312 (2023)
23. Papaioannou, G., Klein, W., Cassart, M., Garel, C.: Indications for magnetic resonance imaging of the fetal central nervous system: recommendations from the european society of paediatric radiology fetal task force. Pediatric Radiology **51**(11), 2105–2114 (2021)
24. Sato, J., Sugimoto, K., Suzuki, Y., Wataya, T., Kita, K., Nishigaki, D., Tomiyama, M., Hiraoka, Y., Hori, M., Takeda, T., et al.: Annotation-free multi-organ anomaly detection in abdominal ct using free-text radiology reports: a multi-centre retrospective study. EBioMedicine **110** (2024)
25. Sofia, C., Aertsen, M., Garel, C., Cassart, M.: Standardised and structured reporting in fetal magnetic resonance imaging: recommendations from the fetal task force of the european society of paediatric radiology. Pediatric Radiology **54**(10), 1566–1578 (2024)
26. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
27. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., Wang, G.: Instruction tuning for large language models: A survey (2024), https://arxiv.org/abs/2308.10792
28. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z.: LlamaFactory: Unified efficient fine-tuning of 100+ language models. In: Cao, Y., Feng, Y., Xiong, D. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. pp. 400–410. Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-demos.38