ASTAR: Automated induction of STAndardized radiology Reporting templates from large-scale clinical free-text corpora

Xinfeng Zhang[1†], Mingxuan Liu[1†], Yifei Chen[1], Juncheng Zhu[2], Kasidit Anmahapong[1], Yiming Huang[3], Yuan Zhang[4], Hongjia Yang[1], Yi Liao[2], Gang Ning[2], Haibo Qu[2], and Qiyuan Tian[1*]

[1] Tsinghua University, Beijing, China
[2] Sichuan University, Chengdu, China
[3] University of California San Diego, San Diego, USA
[4] Southeast University, Nanjing, China
qiyuantian@tsinghua.edu.cn

Abstract. Structured reporting converts free-text radiology narratives into queryable data keys, facilitating cohort assembly, longitudinal tracking, and training label generation for medical AI. The prevailing paradigm follows a two-stage pipeline: (1) constructing a reporting template, (2) extracting information to populate it. While the extraction stage has benefited from advances in large language models (LLMs), template construction remains a manual bottleneck relying on labor-intensive expert consensus that is static, difficult to scale, and may fail to capture real-world reporting diversity. We address this limitation with `ASTAR`, an LLM-based framework for Automated induction of STAndardized radiology Reporting templates from large-scale clinical free-text corpora. Extensive experiments on 4,215 fetal brain MRI reports from multiple centers demonstrate that, in this reporting scenario, the `ASTAR`-induced template surpasses two expert-curated templates across template coverage, information fidelity, diagnostic fidelity, and expert-rated usability, reducing template development from weeks of committee deliberation to hours of automated processing. Code: `https://github.com/birthlab/ASTAR`

Keywords: Radiology report · LLM · Structure information extraction.

1 Introduction

Radiology departments worldwide generate clinical reports at an ever-growing scale. In the United States alone, over 16.1 million CT examinations were reported during a nine-month period in 2020 [11], and imaging utilization across all modalities is projected to rise 17–27% by 2055 [7]. Each examination produces a detailed free-text narrative synthesizing categorical descriptors, quantitative measurements, and diagnostic impressions, thereby forming one of the richest

[†] Equal contribution. [*] Corresponding author.

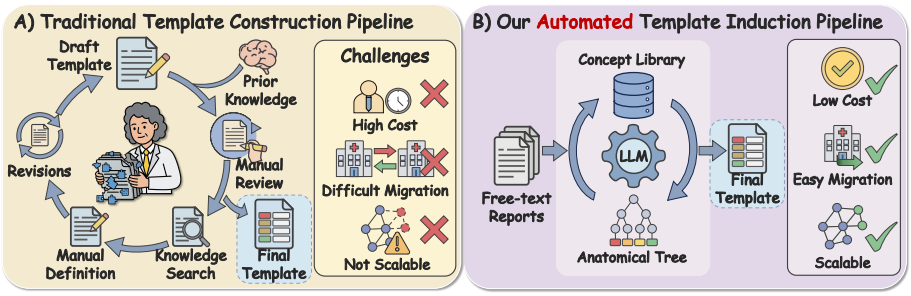

Fig. 1. (A) Traditional template construction pipeline. (B) Our automated template induction pipeline (`ASTAR`).

data repositories in modern healthcare [17,5]. However, the utility of these reports is constrained because findings remain archived as unstructured text with wide stylistic, terminological, and structural variations across radiologists, institutions, and languages [15,16]. Such heterogeneity hinders systematic case retrieval, cohort assembly, longitudinal tracking, and training-label generation for medical AI [6,17]. Radiology societies therefore increasingly advocate for structured reporting [20,4], converting narratives into queryable keys that support rapid cohort identification and high-quality ground-truth generation [1,2].

Nevertheless, achieving these benefits at scale requires automated tools, since manual structuring of the massive and continuously growing volume of narrative reports is highly impractical. Therefore, recent studies increasingly employ natural language processing (NLP) and large language models (LLMs) for automated information extraction [12,23]. Specifically, the prevailing paradigm follows a two-stage pipeline: (1) a reporting template (i.e., a target template defining the set of keys and their permissible values) is constructed; and (2) relevant information is extracted from free-text reports to populate this template. For example, BURExtract-Llama [6] extracts clinical concepts from breast ultrasound reports with an average F1 score of 84.6%.

However, while the extraction stage has benefited substantially from recent advances in NLP and LLMs, the template construction stage, as an upstream prerequisite for all extraction methods, remains a manual bottleneck, since existing templates are typically produced through labor-intensive expert consensus (Fig. 1A). For example, the ESPR template [20] was derived from multidisciplinary panel discussions, and the JSON template used by FetalExtract-LLM [15] was compiled from ESPR guidelines [20], MeSH [14], and a medical dictionary. Beyond this scalability limitation, expert-designed templates are inherently static and may fail to capture the full spectrum of linguistic diversity and evolving clinical descriptions present in massive historical report corpora [4]. As a result, rare but clinically meaningful findings risk being absent from the predefined template, while institution-specific reporting conventions cannot be accommodated by a one-size-fits-all structure.

To the best of our knowledge, no automated method exists for standardized radiology reporting templates construction. To fill this gap, we propose `ASTAR` (Automated induction of STAndardized radiology Reporting templates), an LLM-based framework (Fig. 1B) that automatically mines unified reporting structures from large-scale radiology report corpora. We demonstrate its utility on 4,215 fetal brain MRI reports from three centers, where the `ASTAR`-induced template achieved stable information retention with progressive consolidation of long-tail keys as corpus size increases, offering a scalable, data-driven complement to labor-intensive expert consensus.

2 Method

2.1 Overview of `ASTAR`

Given a corpus of N de-identified free-text radiology reports $\mathcal{D} = \{r_1, \ldots, r_N\}$, the goal of `ASTAR` is to automatically induce a hierarchical reporting template $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \{f_v\}_{v \in \mathcal{V}_{\text{leaf}}})$, where internal nodes represent anatomical or semantic groupings and each leaf node $v \in \mathcal{V}_{\text{leaf}}$ carries a field descriptor $f_v = (k_v, \tau_v, \Omega_v)$ specifying a standardized field name k_v, a value type $\tau_v \in \{\texttt{categorical}, \texttt{numerical}, \texttt{free-text}\}$, and a type-dependent value specification Ω_v. `ASTAR` obtains $\mathcal{T}$ through three stages (Fig. 2): $\mathcal{D} \xrightarrow{\textbf{I}} \Phi \xrightarrow{\textbf{II}} T_1 \xrightarrow{\textbf{III}} \mathcal{T}^*$, where Stage I builds a de-duplicated concept slots library $\Phi = \{\phi_1, \ldots, \phi_L\}$ with each $\phi_j = (k_j, \tau_j, \Omega_j)$ sharing the same structure as f_v; Stage II organizes Φ into a hierarchical tree, mapping each surviving slot to exactly one leaf ($f_v \leftarrow \phi_j$); and Stage III refines the tree into the final template $\mathcal{T}^*$.

2.2 Stage I: Concept Slots Library Construction

In stage I, raw narratives are converted into a de-duplicated library of canonical clinical concept slots in two steps. (1) Span Extraction. An LLM (Qwen-Max [21]) in JSON-mode segments each report r_i into spans $\mathcal{S}_i$, each represented as a 7-tuple (a_{struct}, a_{sub}, a_{attr}, a_{type}, a_{val}, a_{role}, a_{sec}) encoding anatomical location (structure / substructure), attribute identity (name / type), a canonicalized value (numerics abstracted as $\langle\texttt{NUM}\rangle$), clinical role (`finding` | `measurement` | `diagnosis`), and report section (`description` | `impression`). Spans from all reports are pooled and de-duplicated by the composite key $(a_{\text{val}}, a_{\text{role}})$ into a global span set $\mathcal{S}$. (2) Clustering and Slot Induction. Each span is embedded via Qwen3-Embedding-8B [21] and grouped by K-means into C clusters, with the optimal number of clusters C^* chosen by maximizing the silhouette score [19] over $C \in [50, 300]$. Because raw clusters may still mix heterogeneous concepts (e.g., *ventricular width* and *ventricular morphology*), the LLM subpartitions each cluster into semantically coherent subgroups and induces a canonical slot $\phi = (k, \tau, \Omega)$ per subgroup, where k is a standardized key name, $\tau \in \{\texttt{categorical}, \texttt{numerical}, \texttt{free-text}\}$ the value type, and Ω the permissible value set. Slots sharing identical standardized key names (k) are then globally merged by the LLM, yielding the unified concept slots library $\Phi = \{\phi_1, \ldots, \phi_L\}$.

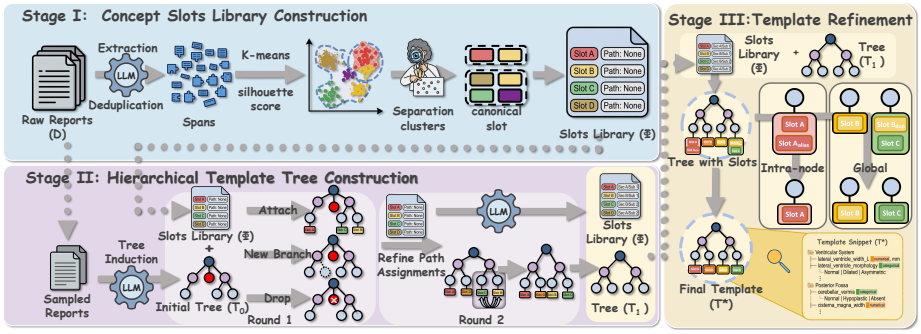

Fig. 2. Overview of ASTAR. (1) Stage I: Concept Slots Library Construction. (2) Stage II: Hierarchical Template Tree Construction. (3) Stage III: Template Refinement.

2.3 Stage II: Hierarchical Template Tree Construction

In Stage II, given the slots library Φ, the flat set of slots is organized into a clinically meaningful hierarchy through a two-round LLM-driven process.
Round 1: Template Construction and Initial Assignment. An LLM (Qwen-Max) first generates an initial Template $\mathcal{T}_0$ from a representative subset of reports, with internal nodes organized by major anatomical regions. Each slot $\phi_j \in \Phi$ is then presented to the LLM together with its descriptive statistics and the current template. The LLM decides to `attach` it under an existing node, `propose` a new branch, or `drop` it. Proposed branches are incorporated into the template incrementally. **Round 2: Global Path Refinement.** Because Round 1 slots are assigned against an evolving template, early slots never see branches proposed later. A second pass therefore re-evaluates all surviving slots in batches on the now-complete tree, taking each slot's Round 1 path as reference, and outputs a revised path, yielding $\mathcal{T}_1$.

2.4 Stage III: Template Refinement

The final stage performs two refinement operations on $\mathcal{T}_1$ to ensure global consistency and output the final template $\mathcal{T}^*$. (1) Intra-node Semantic Reorganization. Within each node, the LLM consolidates synonymous permissible values (e.g., [*"normal"*, *"no abnormality"*, *"unremarkable"*] $\rightarrow$ [*"Normal"*]) and harmonizes value type assignments across sibling keys, reducing value-level redundancy while preserving clinical granularity. (2) Global Concept Harmonization. A tree-wide pass identifies cross-node redundancies, i.e., keys whose names differ across subtrees but refer to the same concept, and standardizes terminology throughout the tree (e.g., unifying *"cerebral hemisphere"* vs. *"brain hemisphere"* across subtrees).

2.5 Evaluation Protocol for Radiology Reporting Templates

Because there is no widely used protocol for evaluating radiology reporting templates, we introduce a three-level protocol to fill this gap: (1) Template

quality assesses whether the template is comprehensive and well-structured; (2) structuring fidelity measures whether populating the template preserves the information in the original report; and (3) radiologist judgment captures whether clinicians find the template usable and clinically valuable. Specifically,

(1) Template quality, computed on a held-out test dataset, quantifies how completely $\mathcal{T}^*$ captures the clinical concepts present in unseen reports, measured by *case-level* (macro-averaged over reports) and *key-level* (micro-averaged over all keys) coverage. For each test report r_i, concept keys K_i are extracted by concatenating entity and attribute keys of each span. An LLM (Qwen-Max) judges whether each key has a semantic match in the template key set $\mathcal{F}(\mathcal{T}^*)$. The two metrics are defined as:

$$\mathrm{Cov}_{\mathrm{case}} = \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_i \frac{\sum_{k \in K_i} \mathrm{match}(k, \mathcal{F})}{|K_i|}, \quad \mathrm{Cov}_{\mathrm{key}} = \frac{\sum_i \sum_{k \in K_i} \mathrm{match}(k, \mathcal{F})}{\sum_i |K_i|}. \tag{1}$$

(2) Structuring fidelity, also computed on the held-out test dataset, measures whether populating the template preserves the information in the original report, assessed via *information fidelity* and *diagnostic fidelity*. Specifically, (a) *information fidelity*, evaluated through a four-step extract–reconstruct round trip, quantifies textual information preservation. *(i)* An LLM (Gemini-2.5-Pro-Preview [8]), chosen independently of the template-construction LLM to prevent extraction bias, populates the template from each report's `imaging description`; *(ii)* extracted key–value pairs are mapped into the full hierarchy and re-flattened for cross-template consistency; *(iii)* another LLM (Qwen-Max) reconstructs a free-text report solely from the structured report, guided by a fixed style-reference text; *(iv)* the reconstruction $\hat{r}_i$ is compared with the original r_i using ROUGE-1/2/L (F1) [13], chrF [18], $\mathrm{BERTScore}_R$ (F1) (with RoBERTa-wwm-ext [10]), and $\mathrm{BERTScore}_M$ (F1) [22] (with MacBERT [9]). As a sensitivity check for model-dependent evaluation, we repeated the round-trip reconstruction with GPT and observed the same relative ranking among the three templates. (b) *Diagnostic Fidelity*, evaluated through an infer-then-judge protocol, tests whether the structured report retains sufficient clinical information for diagnosis. Qwen-Max receives only the structured output, infers a diagnosis d_i^{struct} under strict information closure, and an LLM-as-Judge protocol (with Qwen-Max) is used to compare it against the ground-truth impression d_i^{gt} (i.e., the `diagnostic impression` in the original free-text report) using three metrics (all normalized to $[0, 1]$): *Primary Diagnosis Accuracy* (PDA; 0–5 integer, macro correctness), *Key Finding Preservation* (KFP; 0–1 continuous, abnormal-finding retention), and *Clinical Actionability* (CA; 0–5 integer, decision-support sufficiency).

(3) Radiologist judgment directly evaluates the template via structured clinician ratings. Two radiologists from different countries, ensuring cross-national diversity in clinical conventions and mitigating institution-specific bias, rated all three templates on 28 Likert-scale items (1–5) across three dimensions. (a) *Template Structure* (12 items) evaluates the schema design across six sub-dimensions: coverage & completeness (COMP, ABNO), structure & granularity (LOGI, GRAN)

terminology & clarity (NAME, ENUM), safety & error prevention (NOMS, NOFN, UNCE), adaptability (ADAP), and deployment readiness (ADOP, SIGN); (b) *System Usability* (10 items) follows the standard scale [3], comprising FREQ (willingness to use), COMP (complexity†), EASE (ease of use), TECH (technical support needed†), INTE (integration), INCO (inconsistency†), LEAR (learnability), CUMB (cumbersomeness†), CONF (confidence), and PREL (prior learning needed†), where † marks negatively worded items reverse-scored as $6-x_{\text{raw}}$; (c) *Clinical Impact* (6 items) assesses downstream utility: NOMS (omission reduction), CONS (consistency), ORDE (workflow alignment), NAEX (not-assessed expressibility), COMM (communication efficiency), and OVR (overall satisfaction).

Table 1. Template quality (coverage) and structuring fidelity ((information fidelity & diagnostic fidelity)) across templates evaluated on ID (n=100) and OoD (n=115) test sets. Bold indicates best results. The three templates differ in key count (ESPR: 159, FetalExtract: 51, `ASTAR`: 88), and the results suggest that key relevance and organization, rather than key count alone, contribute to the observed differences.

Category	Metric	ESPR [20]	FetalExtract [15]	`ASTAR`
Structure	Number of keys	159	51	**88**
ID test dataset (n=100)				
Coverage	Case-level	45.41%	36.34%	**81.34%**
	Key-level	45.76%	37.82%	**82.42%**
Information Fidelity	ROUGE-1 (F1)	0.651±0.139	0.409±0.107	**0.749±0.079**
	ROUGE-2 (F1)	0.513±0.167	0.275±0.102	**0.689±0.104**
	ROUGE-L (F1)	0.562±0.124	0.397±0.103	**0.694±0.100**
	chrF	0.576±0.076	0.539±0.093	**0.578±0.075**
	BERTScore$_M$ (F1)	0.873±0.016	0.865±0.020	**0.877±0.017**
	BERTScore$_R$ (F1)	0.892±0.016	0.882±0.021	**0.895±0.016**
Diagnostic Fidelity	PDA	0.879±0.119	0.858±0.158	**0.953±0.090**
	KFP	0.882±0.134	0.844±0.226	**0.958±0.093**
	CA	0.876±0.104	0.858±0.147	**0.953±0.085**
OoD test dataset (n=115)				
Coverage	Case-level	54.32%	37.22%	**66.12%**
	Key-level	55.48%	37.09%	**66.40%**
Information Fidelity	ROUGE-1 (F1)	0.482±0.228	0.316±0.215	**0.717±0.264**
	ROUGE-2 (F1)	0.347±0.233	0.219±0.227	**0.590±0.318**
	ROUGE-L (F1)	0.394±0.175	0.305±0.219	**0.503±0.192**
	chrF	0.322±0.094	0.269±0.123	**0.376±0.073**
	BERTScore$_M$ (F1)	0.843±0.024	0.843±0.031	**0.850±0.020**
	BERTScore$_R$ (F1)	0.853±0.029	0.851±0.033	**0.862±0.023**
Diagnostic Fidelity	PDA	0.790±0.189	0.802±0.143	**0.847±0.125**
	KFP	0.733±0.206	0.739±0.182	**0.811±0.154**
	CA	0.751±0.137	0.757±0.134	**0.847±0.107**

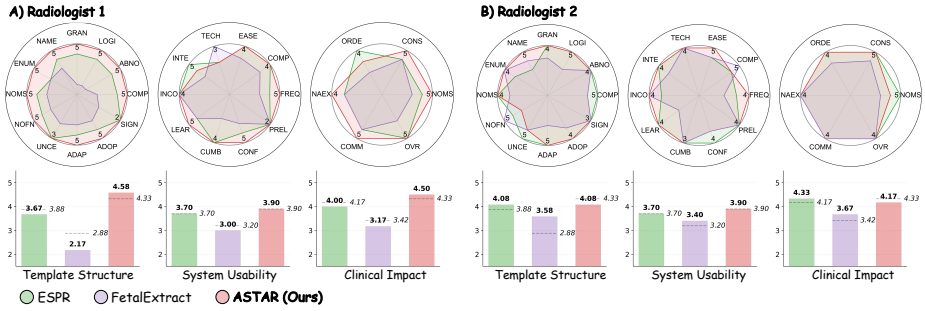

Fig. 3. Two radiologists from different countries rated all templates on 28 Likert-scale items (1–5, higher is better). **(A)** Radiologist 1; **(B)** Radiologist 2. **Top row:** per-item radar charts for Template Structure (12 items), System Usability (10 items), and Clinical Impact (6 items); numbers on each axis denote the maximum score among the three templates for that item. **Bottom row:** dimension-level means; dashed lines show cross-radiologist averages.

3 Experiments and Results

3.1 Data Collection

A total of 4,100 de-identified fetal brain MRI free-text reports were retrospectively collected from West China Second University Hospital, covering examinations performed between October 2015 and December 2024 on one 3.0 T system (Siemens MAGNETOM Skyra) and two 1.5 T systems (Philips Achieva and United Imaging uMR 570). Each report comprises a narrative *imaging description* section and a *diagnostic impression* section. Of these, 4,000 reports were used for template construction with `ASTAR` and 100 were held out as an in-distribution (ID) test dataset. To assess cross-institutional generalizability, an additional out-of-distribution (OoD) test dataset of 115 de-identified fetal brain MRI reports was collected from Sichuan Provincial Woman's and Children's Hospital and Chengdu Women's and Children's Central Hospital.

3.2 Comparison with Expert-Designed Templates

The ASTAR-induced template was compared against two expert-designed templates. (1) ESPR Template [20]: a standardized fetal MRI reporting template released in 2024 by the Fetal Task Force of the European Society of Paediatric Radiology. (2) FetalExtract Template [15]: a JSON template compiled from ESPR guidelines, MeSH, and a medical dictionary, used for structured information extraction with FetalExtract-LLM.

Across all metrics, `ASTAR` consistently outperforms both expert-designed templates on both ID and OoD datasets, achieving higher coverage with fewer keys,

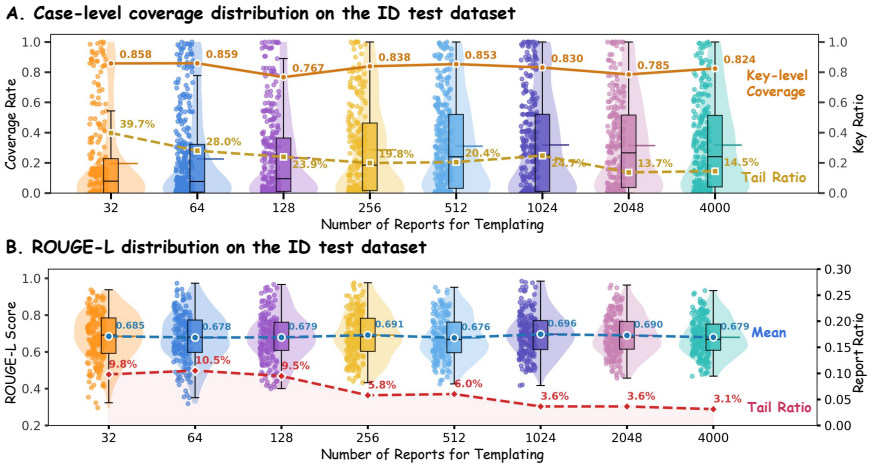

Fig. 4. Scaling behavior of ASTAR-induced templates on the ID test set (n=100). **(A)** Case-level coverage (left axis), key-level coverage (orange, left axis) and key coverage tail ratio (gold, right axis). Coverage values are ratios bounded in [0,1]. **(B)** ROUGE-L distributions (left axis), mean ROUGE-L (blue, left axis) and report-level tail ratio (red dashed, right axis).

superior information and diagnostic fidelity, and the highest expert-rated usability (Table 1, Fig. 3). (1) Template Quality: On the ID test dataset, `ASTAR` achieves case-/key-level coverage of 81.34%/82.42%, outperforming ESPR and FetalExtract. `ASTAR` is data-induced, higher ID coverage is expected and should be interpreted together with OoD and radiologist evaluations. This advantage persists on the OoD set, though the gap narrows due to expected domain shift. (2) Structuring Fidelity: `ASTAR` consistently achieves the highest scores across all six metrics for *information fidelity* on both datasets. Specifically, on the ID dataset, `ASTAR` attained ROUGE-L 0.694 ± 0.100, substantially outperforming ESPR and FetalExtract; $\text{BERTScore}_\text{R}$ reached 0.895 ± 0.016 vs. 0.892 ± 0.016 and 0.882 ± 0.021. On the OoD test dataset, `ASTAR` maintains a clear lead (ROUGE-L 0.503 ± 0.192 vs. 0.394 ± 0.175 and 0.305 ± 0.219), with larger standard deviations reflecting greater cross-institutional stylistic heterogeneity. For *diagnostic fidelity*, `ASTAR` achieves the highest scores across all three dimensions on the ID set: PDA 0.953 ± 0.090, KFP 0.958 ± 0.093, and CA 0.953 ± 0.085, outperforming ESPR (PDA/KFP/CA: 0.879/0.882/0.876) and FetalExtract (0.858/0.844/0.858). On the OoD set, `ASTAR` retains its advantage (PDA 0.847, KFP 0.811, CA 0.847), with consistently high KFP (>0.81) confirming that the automatically induced template preserves critical abnormal findings (e.g., ventriculomegaly, posterior fossa anomalies) essential for clinical decision-making. (3) Radiologist Judgment: `ASTAR` receives the highest scores across *Template Structure* 4.33/5.00 (vs. ESPR 3.88, FetalExtract 2.88), *System Usability* 3.90 (vs. 3.70, 3.20), and *Clinical Impact* 4.33 (vs. 4.17, 3.42).

3.3 Scaling Behavior of ASTAR

To investigate how corpus size affects template quality, we constructed ASTAR templates from subsets of n=32 to n=4000 reports in the training dataset and evaluated each on the ID test dataset (Fig. 4). Key-level coverage remains consistently high (0.767–0.859), while the tail ratio (case-level coverage < 0.01) decreases from 39.7% at n=32 to 14.5% at n=2048 before stabilizing, confirming that ASTAR progressively consolidates rare-but-valid concepts into coherent keys (Fig. 4A). Additionally, mean ROUGE-L remains stable across all corpus sizes (0.676–0.696), while the report-level tail ratio (case-level ROUGE-L < 0.5) drops from 9.8–10.5% at $n \leq 64$ to 3.1% at n=4000, indicating that additional reports primarily improve coverage of structurally atypical or linguistically rare cases rather than boosting average fidelity (Fig. 4B).

4 Conclusion

We presented ASTAR, the LLM-based framework that automatically induces standardized radiology reporting templates from large-scale free-text corpora. Evaluated on 4,215 multi-center fetal brain MRI reports, the ASTAR-induced template outperforms two expert-designed baseline templates in template quality, structuring fidelity, and radiologist-rated usability, reducing template development from weeks of expert consensus to hours of automated processing. Future work will extend the evaluation beyond fetal brain MRI to broader radiology reporting scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (grant number 82572198), the Natural Science Foundation of Sichuan Province (grant number 2025ZNSFSC1768), the Scientific Research Project of Sichuan Medical Association (grant number 2024HR130), the Science and Technology Department of Sichuan Province, China (grant number 25SYSX0255), the Tsinghua University Startup Fund, and the Tsinghua University Dushi Program (grant numbers 20241080026, 20251080056).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. 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
2. Bai, X., Liu, M., Song, T., Chen, Y., Yang, H., Anmahapong, K., Li, Z., Zhou, Y., Tian, Q.: Exact: an explainable anomaly-aware vision foundation model for analysis of 3d chest ct (2026), https://arxiv.org/abs/2604.24146
3. Brooke, J.: SUS: A 'quick and dirty' usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)

4. Busch, F., Hoffmann, L., Dos Santos, D.P., Makowski, M.R., Saba, L., Prucker, P., Hadamitzky, M., Navab, N., Kather, J.N., Truhn, D., Cuocolo, R., Adams, L.C., Bressem, K.K.: Large language models for structured reporting in radiology: past, present, and future. European Radiology **35**(5), 2589–2602 (Oct 2024). https://doi.org/10.1007/s00330-024-11107-6, https://link.springer.com/10.1007/s00330-024-11107-6
5. Chen, H., Chen, Y., Yan, Z., Ding, M., Li, C., Zhu, Z., Qin, F.: Mmlnb: Multimodal learning for neuroblastoma subtyping classification assisted with textual description generation (2025), https://arxiv.org/abs/2503.12927
6. 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. pp. 53–58. ACM, Melbourne VIC Australia (Oct 2024). https://doi.org/10.1145/3688868.3689200, https://dl.acm.org/doi/10.1145/3688868.3689200
7. Christensen, E.W., Drake, A.R., Parikh, J.R., Rubin, E.M., Rula, E.Y.: Projected US imaging utilization, 2025 to 2055. Journal of the American College of Radiology **22**(2), 151–158 (2025). https://doi.org/10.1016/j.jacr.2024.10.017
8. Comanici, G., Bieber, E., Schaekermann, M., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025)
9. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for chinese natural language processing. In: Findings of the Association for Computational Linguistics: EMNLP 2020. p. 657–668. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.58, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.58
10. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for chinese bert. IEEE Transactions on Audio, Speech and Language Processing (2021). https://doi.org/10.1109/TASLP.2021.3124365, https://ieeexplore.ieee.org/document/9599397
11. Davenport, M.S., Fruscello, T., Chatfield, M., Weinstein, S., Sensakovic, W.F., Larson, D.B.: Ct volumes from 2,398 radiology practices in the united states: A real-time indicator of the effect of COVID-19 on routine care, january to september 2020. Journal of the American College of Radiology **18**(3), 380–387 (2021). https://doi.org/10.1016/j.jacr.2020.10.010
12. Fytas, P., Breger, A., Selby, I., Baker, S., Shahipasand, S., Korhonen, A.: Can rule-based insights enhance LLMs for radiology report classification? introducing the RadPrompt methodology. In: Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., Tsujii, J. (eds.) Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. pp. 212–235. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.bionlp-1.17
13. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
14. Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association **88**(3), 265 (2000)
15. Liu, M., Li, Y., Zhu, J., Yang, H., Huang, Y., Li, H., Chen, Y., Bai, X., Liao, Y., Qu, H., Tian, Q.: FetalExtract-LLM: Structured information extraction from free-text fetal MRI reports based on privacy-ensuring open-weights large language models.

In: Link-Sourani, D., Abaci Turk, E., Bastiaansen, W., Hutter, J., Melbourne, A., Licandro, R. (eds.) Perinatal, Preterm and Paediatric Image Analysis. pp. 119–129. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-05997-0_11

16. Nowak, S., Biesner, D., Layer, Y.C., Theis, M., Schneider, H., Block, W., Wulff, B., Attenberger, U.I., Sifa, R., Sprinkart, A.M.: Transformer-based structuring of free-text radiology report databases. European Radiology **33**(6), 4228–4236 (Mar 2023). https://doi.org/10.1007/s00330-023-09526-y
17. 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
18. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (2015)
19. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987). https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7
20. 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 (Aug 2024). https://doi.org/10.1007/s00247-024-06010-7
21. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
22. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert (2020), https://arxiv.org/abs/1904.09675
23. Zhang, Y., Zhang, X., Qi, X., Wu, X., Chen, F., Yang, G., Fu, H.: Content generation models in computational pathology: A comprehensive survey on methods, applications, and challenges. IEEE Reviews in Biomedical Engineering **19**, 374–395 (2026). https://doi.org/10.1109/RBME.2025.3619086