An Agentic LLM Pipeline for Structured Adaptation of Clinical AI Literature: Construction and Validation of HOps

Anonymous Authors[1]

Abstract

Actionable clinical AI guidance is dispersed across hundreds of heterogeneous, unstructured PDF documents spanning reporting guidelines, implementation science frameworks, and digital health evaluation tools. We present a six-agent sequential LLM pipeline that converts this unstructured medical literature into a structured, machine-readable tabular dataset at scale. Applied to 55 clinical AI framework documents, the pipeline extracted 2,386 classified items mapped to eight predefined lifecycle stages, revealing systematic coverage gaps across the literature. This structured synthesis directly informed the development of **HOps**, a 20-question guiding framework for hospital digital product development, which achieved 87% endorsement and 67% question-level agreement across five domain experts. Our work demonstrates that agentic structured adaptation of clinical literature is feasible, produces reproducible outputs, and yields structured datasets with direct utility for clinical AI governance and deployment research.

1. Introduction

The development and deployment of clinical AI systems requires navigating a complex landscape of guidance documents: AI/ML reporting guidelines (e.g., TRIPOD+AI, DECIDE-AI, SPIRIT-AI), implementation science frameworks, and digital health evaluation standards. This guidance exists almost exclusively as unstructured free text in PDF documents, making systematic comparison, gap analysis, and evidence synthesis a time-intensive manual process.

Structured data is the foundation of reproducible medical research. Yet the meta-level question of *what the field recommends* about developing and deploying clinical AI remains locked in prose. This creates a bottleneck: practitioners cannot efficiently identify which aspects of a development lifecycle are well-covered by existing guidance and which are neglected; researchers cannot systematically compare frameworks at scale; and developers cannot programmatically query the state of the field.

We address this gap with a **six-agent LLM pipeline** that performs structured adaptation of unstructured clinical documents, transforming free-text framework PDFs into validated, structured tabular data. This structured dataset then serves as the empirical foundation for **HOps** (Hospital Operations framework), a 20-question guiding framework for hospital digital product development, expert-validated via structured content validity assessment.

Our contributions are:

1. A reproducible six-agent pipeline for structured adaptation of clinical AI literature, including a self-correcting reconciliation loop;
2. A structured dataset of 2,386 items extracted from 55 clinical AI framework documents, classified across eight lifecycle stages;
3. HOps — a validated 20-question framework derived from the structured synthesis — with expert-assessed content validity.

2. Related Work

Structured information extraction from medical text. LLMs have demonstrated strong performance on named entity recognition, relation extraction, and structured output generation from clinical notes and biomedical literature (Agrawal et al., 2022; Guo et al., 2023). Forced JSON-mode generation and schema-constrained decoding have improved the reliability of structured outputs (Tam et al., 2024). Our work extends this to a novel domain: meta-level clinical governance documents, which are longer, less structured, and require hierarchical schema extraction.

Agentic pipelines for complex tasks. Sequential multi-agent architectures have demonstrated effectiveness

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

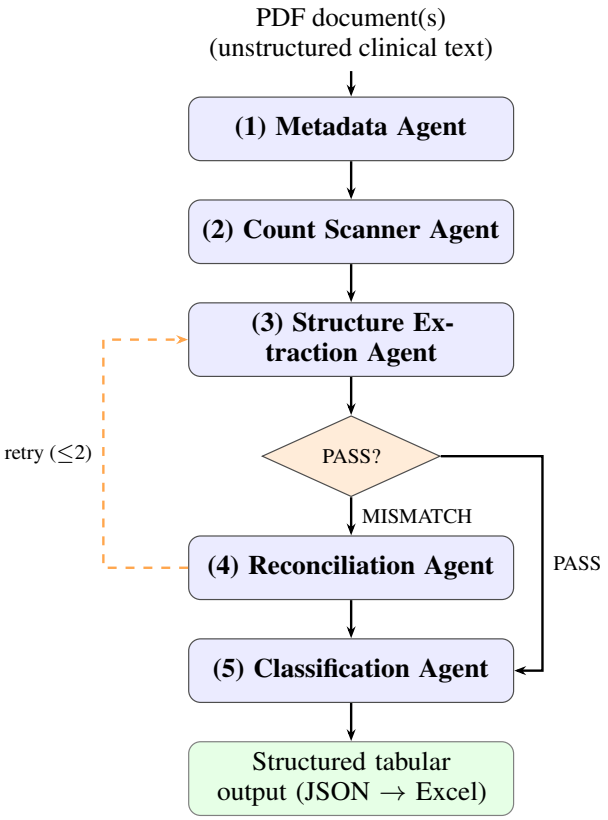

Figure 1. Six-agent pipeline for structured adaptation of clinical AI literature. Dashed arrow indicates the self-correcting reconciliation loop (up to 2 retries).

for tasks requiring decomposition, validation, and self-correction (Wu et al., 2023; Shinn et al., 2023). Our pipeline applies the reflexion pattern specifically to the problem of count-constrained structured extraction, where documents self-report item counts that can serve as verifiable targets.

Clinical AI framework analysis. Manual scoping reviews of clinical AI frameworks (Clusmann et al., 2023; Vasey et al., 2022) are limited in scale and reproducibility. Automated structured extraction enables larger-scale analysis and systematic coverage mapping not previously possible.

3. Method: The Six-Agent Pipeline

3.1. Overview

Figure 1 illustrates the complete pipeline. All agents use DeepSeek-V3 (128k-token context window) via the OpenAI-compatible API. Long documents are handled by a token-aware chunker using `tiktoken` with the `cl100k_base` encoding, splitting at paragraph boundaries to a maximum of 100,000 tokens per chunk, with chunk-level outputs merged by category name deduplication.

3.2. Agent Descriptions

(1) Metadata Agent. Operates on the first 3,000 tokens of the document at `temperature=0`. Extracts framework name, publication year, DOI, and primary focus area via forced JSON-mode generation. Output feeds document provenance fields in the structured dataset.

(2) Count Scanner Agent. Searches the document abstract and conclusion sections (first 8,000 + last 3,000 tokens) for self-reported item counts — sentences of the form "comprises 17 items and 28 sub-items" or "42 reporting items across 6 domains." Returns expected total leaf-item count and calculation logic (e.g., "10 standalone + 28 sub-items = 38"). When no counts are found, flags the document for blind extraction.

(3) Structure Extraction Agent. The core extraction agent. Applies forced JSON-mode generation with `temperature=0` to produce a hierarchical schema:

$$\texttt{categories} \rightarrow \texttt{items} \rightarrow \texttt{subitems}$$

Each item carries: `item_id`, `item_label`, `item_text`, `item_type` (`explicit` | `inferred` | `header`), and verbatim `source_text`. Items typed as `header` are excluded from leaf-item counts (they serve as parent containers). Narrative paragraphs without explicit enumeration are decomposed into `inferred` items. A system prompt enforces: *"CATEGORIES = top-level phases; ITEMS = numbered or bulleted actionable entries; SUB-ITEMS = sub-entries. If an item HAS sub-items → item_type='header'. Capture EVERY item — do not skip or merge any."* When documents exceed the context limit, the chunker splits at paragraph boundaries and merges extracted categories by name.

(4) Validation Agent. Computes leaf-item count from the extracted structure (standalone items + all sub-items; headers excluded). Compares against Count Scanner output. Returns a three-state verdict: `PASS` (counts match), `MISMATCH` (counts differ), or `NO_EXPECTED_COUNTS` (document did not self-report counts). Difference direction (over- vs. under-extraction) is recorded for audit.

(5) Reconciliation Agent. Triggered only on `MISMATCH`. Reinjects the expected count and the Count Scanner's source sentence as a hard constraint into the Structure Extraction Agent system prompt: *"Your extraction MUST match these counts exactly: [count] leaf items = [logic]."* Retries up to `MAX_RECONCILE_ATTEMPTS=2`. The extraction with the smallest absolute count difference is retained as the best candidate.

(6) Classification Agent. Assigns each leaf item to one of eight predefined Clinical AI Lifecycle (CAL) stage codes (Table 2) using a structured prompt at `temperature=0`.

Batches of items are classified per document to minimise API calls.

3.3. Output Schema

The pipeline outputs two artefacts: (i) a hierarchical JSON document preserving the full extraction with provenance, and (ii) a flat tabular dataset where each row represents one leaf item with fields: framework ID, framework name, year, category/phase, item label, item text, level (ITEM/SUB-ITEM), item type, CAL stage code, CAL stage name, verbatim source text, and DOI.

4. Corpus and Structured Dataset

4.1. Corpus

We applied the pipeline to a corpus of 55 clinical AI framework documents assembled via systematic scoping review across three categories:

- **G1 — AI/ML Reporting Guidelines** (n=15): Documents specifying reporting standards for AI systems in clinical research, including CONSORT-EHEALTH (Eysenbach & CONSORT-EHEALTH Group, 2011), SPIRIT-AI, TRIPOD+AI (Collins et al., 2021), DECIDE-AI (Vasey et al., 2022), CLAIM 2024, and related guidelines.

- **G2 — Implementation Science** (n=10): Frameworks governing adoption and scale-up of health interventions, including implementation science models and quality improvement frameworks.

- **G3 — Product Development, Digital Health, and Evaluation** (n=30): Frameworks for digital health product design, hospital operations, and clinical AI evaluation, including Stanford Biodesign (Zenios et al., 2015) and others.

Documents ranged from highly structured checklists (e.g., itemised reporting guidelines with explicit numbered items) to unstructured editorial prose (e.g., CONSORT-EHEALTH, a narrative editorial describing guideline development with no pre-built checklist table). For the latter, the Structure Extraction Agent applied an `inferred` decomposition strategy, extracting implied items from narrative sections.

4.2. Structured Dataset

The pipeline extracted **2,386 leaf items** across 55 documents. Table 1 summarises extraction by corpus group and lifecycle stage.

Table 1. Extracted items by corpus group and Clinical AI Lifecycle (CAL) stage. PF=Problem Framing, UN=User Needs, DP=Design & Prototyping, DI=Development & Integration, VI=Validation & Testing, IA=Implementation & Adoption, SS=Scale-Up & Sustainability, PE=Post-Deployment Evaluation.

Stage	G1	G2	G3	Total
PF	68	32	238	338
UN	27	24	151	202
DP	38	13	127	**178**
DI	120	2	138	260
VI	151	28	187	366
IA	30	191	265	**486**
SS	33	76	197	306
PE	42	61	147	250
Total	**509**	**427**	**1450**	**2386**

Table 2. Eight Clinical AI Lifecycle (CAL) stage codes used for classification.

Code	Description
PF	Problem framing, clinical rationale, target population
UN	User and patient needs, workflow context
DP	Model architecture, algorithm selection, output design
DI	Data, annotation, training, reproducibility
VI	External validation, bias testing, calibration
IA	Deployment, user training, rollout governance
SS	Regulatory status, funding, long-term planning
PE	Post-deployment surveillance, outcome tracking

4.3. Coverage Gap Analysis

The structured dataset reveals systematic imbalances in how existing clinical AI literature covers the development lifecycle. Implementation & Adoption (IA, n=486) and Validation & Testing (VI, n=366) are the most represented stages, driven by the large G2 implementation science corpus and the G1 reporting guidelines' emphasis on statistical validation. Design & Prototyping (DP, n=178) and User Needs (UN, n=202) are critically underrepresented — DP is the least-covered stage in G1 and G2 combined, and UN receives fewer than 60% of the coverage that IA receives.

This structured analysis would not be possible at this scale through manual review. The 2,386-item dataset enables queries such as: *Which frameworks cover post-deployment evaluation? Which reporting guidelines address algorithm selection?*

Table 3. HOps expert content validity results (n=5 experts).

Measure	n	Agreement
Question-level (≥4/5)	67/100 ratings	67%
Question range (min–max)	—	20–100%
Framework endorsement	13/15 statements	87%
Endorsement breakdown		
Addresses essential domains	5/5 experts	100%
Logically organised	4/5 experts	80%
Would strengthen practice	4/5 experts	80%

5. HOps: Framework Construction and Validation

5.1. Framework Construction

HOps (Hospital Operations framework) was derived through structured synthesis of the 2,386-item dataset. Gaps identified in the coverage analysis directly informed item generation: the underrepresentation of DP and UN in existing literature motivated inclusion of dedicated questions on user needs elicitation (Q6, Q7) and the "Mandatory Handshake" design gate (Q8–Q10), which explicitly bridges clinical requirements to technical specifications — a transition absent from most existing guidelines. The final framework comprises **20 key guiding questions** organised across 8 stages aligned with the CAL codes, divided into two phases: Phase 1 (Framing & Foundation: Steps 1–4) and Phase 2 (Execution & Sustainability: Steps 5–8).

5.2. Expert Content Validity Assessment

Five domain experts (clinical informaticians and digital health researchers) assessed each of the 20 questions across three dimensions: **clarity** (is the question unambiguous?), **completeness** (does it capture the relevant construct?), and **applicability** (is it actionable in hospital settings?). Each dimension was rated on a 5-point scale; agreement was defined as a rating of ≥4/5.

Three summary endorsement statements were also rated: (i) HOps addresses essential digital product development domains; (ii) HOps is logically organised; (iii) HOps would strengthen hospital digital product development practice.

Results are summarised in Table 3. Overall question-level agreement was **67%** (67/100 ratings at ≥4/5), with individual question agreement ranging from 20% to 100%. Framework-level endorsement across the three summary statements was **87%** (13/15). Questions with lower agreement (notably early governance questions Q1–Q2) were revised following expert feedback, with final wording reflecting specific accountability and decision-authority language.

6. Discussion

Structured adaptation as a research method. Our pipeline demonstrates that systematic structured extraction from unstructured medical literature is achievable with LLM-based agents, producing auditable, reproducible outputs. The self-correcting reconciliation loop — which uses documents' own self-reported counts as verifiable extraction targets — is a practically useful quality mechanism unique to this domain: few other corpora contain such explicit ground-truth count statements.

Coverage gap findings. The consistent underrepresentation of Design & Prototyping (DP) across all three corpus groups is clinically significant. Most reporting guidelines focus on what to *report after* a system is built, not how to design it. This bias in the literature translates directly into gaps in practitioner guidance — a finding that only emerges from structured cross-framework analysis.

HOps validity. The 67% question-level agreement, while positive, also indicates room for refinement. Questions with low agreement clustered around governance and accountability (Q1–Q2), suggesting these constructs require more precise operationalisation in the hospital context. The 87% framework endorsement is a stronger signal that the overall structure and coverage of HOps are appropriate.

Limitations. The expert validation cohort (n=5) is small; a larger Delphi study is planned. The corpus is English-only. The pipeline's reconciliation mechanism does not guarantee perfect extraction — best-candidate selection may still yield mismatches. All extraction outputs should be treated as machine-assisted drafts requiring domain review.

Future work. We plan to (i) release the structured 2,386-item dataset as an open resource, (ii) extend the corpus to non-English frameworks, and (iii) apply the classification schema to individual clinical AI papers to enable automated coverage assessment at the literature level.

7. Conclusion

We presented a six-agent LLM pipeline for structured adaptation of unstructured clinical AI literature, producing a 2,386-item structured dataset from 55 framework documents. The pipeline's self-correcting reconciliation loop, forced JSON-mode generation, and zero-temperature classification agents collectively enable reproducible structured extraction at a scale not achievable through manual review. The resulting structured dataset revealed systematic coverage gaps — particularly in Design & Prototyping and User Needs — that directly informed the construction and expert validation of HOps, a 20-question hospital digital product development framework. This work establishes agentic structured adaptation as a viable and valuable method for

synthesising clinical AI guidance literature into machine-readable structured data.

Impact Statement

This paper presents work aimed at advancing structured data methods in clinical AI governance. The structured dataset and HOps framework are intended to support safe, evidence-based deployment of AI systems in hospital settings. Potential societal benefits include improved governance of clinical AI, more systematic gap identification in development guidance, and a reproducible infrastructure for meta-research on clinical AI frameworks. We are not aware of specific harms arising from this work, though we note that automated extraction outputs require human expert review before use in safety-critical governance decisions.

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. Large language models are few-shot clinical information extractors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, 2022.

Clusmann, J., Kolbinger, F. R., Muti, H. S., et al. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141, 2023.

Collins, G. S., Dhiman, P., Andaur Navarro, C. L., et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7):e048008, 2021.

Eysenbach, G. and CONSORT-EHEALTH Group. CONSORT-EHEALTH: Improving and standardizing evaluation reports of web-based and mobile health interventions. *Journal of Medical Internet Research*, 13 (4):e126, 2011.

Guo, L., Hu, M., Zhao, Y., et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023.

Shinn, N., Cassano, F., Labash, B., et al. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Tam, J. et al. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

Vasey, B., Nagendran, M., Campbell, B., et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5):924–933, 2022.

Wu, Q., Bansal, G., Zhang, J., et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *arXiv preprint arXiv:2308.08155*, 2023.

Zenios, S., Makower, J., Yock, P., et al. *Biodesign: The Process of Innovating Medical Technologies*. Cambridge University Press, 2nd edition, 2015.