Paper ML06

Teaching Small Models to Think Big: Enterprise-Ready AI Through Knowledge Distillation

Ramanathan Srinivasan, Saama Technologies Inc., Livermore, CA, USA
Angu S Krishna Kumar, Saama Technologies Inc, Livermore, CA, USA

ABSTRACT

In pharmaceutical document processing, not every task requires the full power of large language models—yet current approaches force a choice between expensive LLM APIs for everything or limited rule-based extraction. Sequence-Level Knowledge Distillation enables an intelligent middle path: compact models trained on high-fidelity synthetic data serve as the first line of processing, escalating only complex cases to large language models for deeper analysis. This paper presents a complementary workflow where distilled models rapidly process standard protocols, safety reports, and regulatory documents, while identifying and routing edge cases—unusual formats, ambiguous content, or critical safety signals—to more capable models. This hybrid architecture transforms document processing economics: organizations can process thousands of standard documents locally at minimal cost while preserving budget and computational resources for the subset requiring advanced reasoning—making enterprise-scale clinical document intelligence both practical and economically sustainable.

1.0 INTRODUCTION

The pharmaceutical industry is currently facing a data processing paradox. While Large Language Models (LLMs) have demonstrated transformative capabilities in interpreting complex medical text—such as clinical trial protocols and adverse event reports—their deployment at enterprise scale remains cost-prohibitive and computationally intensive. A single Phase III clinical trial can generate hundreds of pages of documentation, with the Schedule of Activities (SoA) table representing a critical nexus of information. These tables encode the temporal logic of a trial, dictating when procedures occur, yet they are notoriously difficult to extract due to rotated headers, merged cells, and idiosyncratic symbols like "X" or "✓".

Traditional rule-based systems (e.g., pdfplumber or PyMuPDF) offer speed but fail catastrophically on these complex layouts, often yielding "garbled" text when headers are rotated 90 degrees or when columns are sparse. Conversely, frontier multimodal models like Gemini 2.5 can perceive and reason about these visual structures with near-human accuracy, but applying them to every page of millions of documents is economically unsustainable. Furthermore, recent research on the "emergent abilities" of LLMs suggests that complex tabular reasoning often requires a computational scale that exceeds standard operational budgets [Wei et al., 2022].

This paper introduces a Sequence-Level Knowledge Distillation framework that resolves this paradox. Unlike traditional distillation methods that rely on teacher model probability distributions (soft targets), our approach uses Sequence-Level Distillation with hard pseudo-labels [Kim & Rush, 2016]. We propose a hybrid "Teacher-Student" architecture in which a highly capable "Teacher" model (Gemini 2.5 Pro) is used solely to generate high-quality ground truth data for the most complex document samples. This synthetic data is then used to teach a cost-effective, enterprise-grade "Student" model (Llama-3.3-70B-Instruct) to recognize and extract these patterns locally [Dubey et al., 2024]. By distilling the visual reasoning capabilities of the Teacher into the text-processing efficiency of the Student through rigorous instruction tuning [Wang et al., 2023], we achieve a system that retains high accuracy while drastically reducing inference costs and latency.

2.0 SYSTEM ARCHITECTURE: THE TEACHER PIPELINE

2.1 High-Level Overview

The proposed system implements a Hybrid Knowledge Distillation framework that decouples reasoning capabilities from computational cost. Unlike traditional distillation, where a student model mimics a teacher's probability distributions, our architecture utilizes Sequence-Level Distillation to transfer reasoning patterns via structured outputs. As illustrated in Figure 1, the system operates in two distinct phases:

Training & Distillation (Phase 1): A "Teacher" pipeline generates high-fidelity synthetic ground truth (JSON) from complex raw documents. This curriculum is used to train the Student model via Supervised Fine-Tuning (SFT), utilizing parameter-efficient techniques like Low-Rank Adaptation (LoRA) [Hu et al., 2022].

Figure 1:

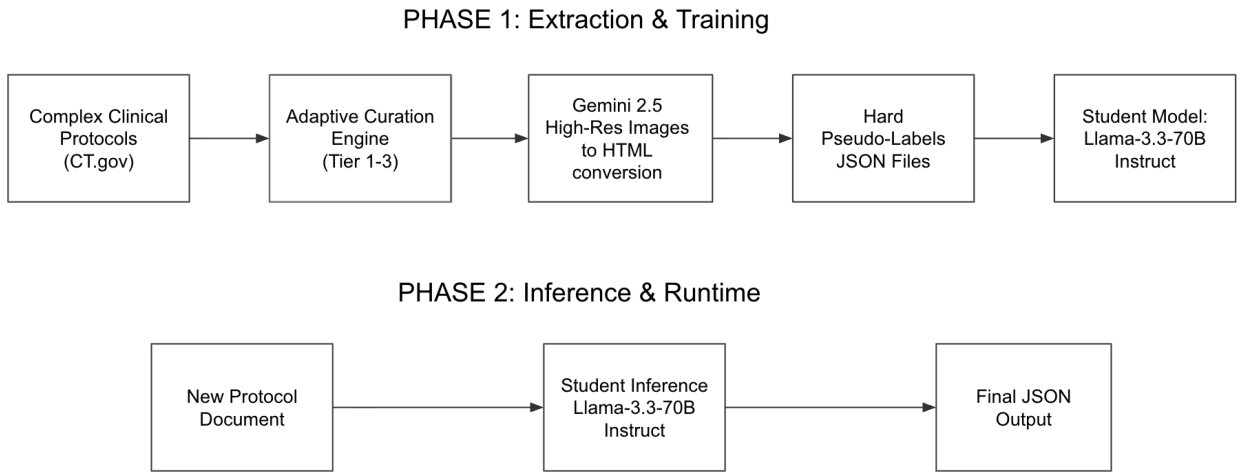

Figure 1: The Hybrid Distillation Architecture showing Phase 1 training pipeline, where Gemini 2.5 generates structured JSON ground truth for Student model fine-tuning.

2.2 Phase 1: The Teacher Pipeline (Curriculum Generation)

To train a robust Student model, we developed a robust PDF ingestion engine that downloads various complex Clinical protocols from clinicaltrials.gov and processes them to generate the final structured JSON output. This engine serves as the source of truth in identifying the SOA pages and converting them into images, Images to HTML via various packages that are available, e.g., pdfplumber or multi-modal "Teacher" model (Gemini 2.5) to convert them into HTML and, in turn, to a structured, schema-compliant ground truth for the most complex document samples. This process is formally categorized as Sequence-Level Distillation, in which the student learns directly from the teacher's final output rather than its internal probability distributions [Kim & Rush, 2016].

The pipeline implements a Three-Tier Extraction Strategy to process document images into HTML files and generate this curriculum:

Tier 1: Deterministic Extraction. For standard, grid-aligned tables, the system utilizes structural extraction algorithms (e.g., pdfplumber) with dynamic column consolidation. This handles the majority of simple layouts at near-zero cost.

Tier 2: Computer Vision Fallback. If structural extraction detects sparse data or potential image-based artifacts, the system triggers an Optical Character Recognition (OCR) layer optimized for table cell boundary detection.

Tier 3: The Teacher Model (Multimodal Escalation). The system automatically routes pages exhibiting high visual complexity—specifically rotated vertical text, merged headers, or heavy footnoting—to the Teacher model. Gemini 2.5 perceives the page as an image and generates a strict HTML representation of the table structure. This HTML output is then converted into a standardized JSON format, serving as the "gold standard" label for the Student model's Supervised Fine-Tuning (SFT) [Wang et al., 2023)].

This hierarchical approach ensures that the training data is both high-quality (guaranteed by the Teacher's schema enforcement) and diverse, covering the full spectrum of document layouts.

3.0 Student Model Architecture

3.1 The "Goldilocks" Parameter Challenge

In clinical document extraction, model selection is governed by a fundamental trade-off between reasoning capacity and operational efficiency. To identify the optimal student architecture, we conducted experiments at two parameter scales representing the extremes of the efficiency-capability spectrum: 3B (Llama-3.2-3B-Instruct) and 70B (Llama-3.3-70B-Instruct) parameters.
Our experiments revealed a stark performance bifurcation. As shown in **Table 4** (see Section 4.1), the Llama-3.2-3B model achieved only **25% accuracy** on Schedule of Activities (SoA) extraction—well below the 80% enterprise threshold. Qualitative error analysis identified two dominant failure modes:

- **Attention Drift:** The 3B model frequently lost column alignment when processing tables exceeding 15 rows, resulting in procedure-visit mismatches.
- **Hallucinated Checkmarks:** In sparse columns (where <20% of cells contain values), the model inserted

phantom "X" markers, likely due to statistical priors learned during pre-training.

These findings align with recent work on emergent capabilities in language models (Wei et al., 2022), which suggests that complex reasoning tasks exhibit phase transitions at specific parameter scales. For tabular reasoning requiring multi-hop inference, the 70B parameter class represents this critical inflection point.

Conversely, while the Teacher model (Gemini 2.5) achieved 90% accuracy, its API-based inference incurs ~$10 per million tokens and ~2-3 min of latency per SoA table, which is prohibitive for processing millions of pages at enterprise scale. We therefore selected **Llama-3.3-70B-Instruct** as the optimal "Goldilocks" architecture: sufficiently large to encode hierarchical relationships in clinical protocols (e.g., distinguishing "Pre-Screening" from "Screening" visits), yet deployable on commercial hardware without API dependencies.

3.2 Architectural Suitability for Tabular Reasoning

Two architectural features of Llama-3.3 make it particularly suited for cross-modal distillation from a vision-capable Teacher:

- **Grouped-Query Attention (GQA):** Clinical protocols frequently span 20+ pages, requiring the model to reference footnotes on distant pages to interpret cell values. Llama-3.3 employs Grouped-Query Attention (Ainslie et al., 2023), which reduces the memory footprint of the Key-Value (KV) cache by sharing key-value heads across query groups. This enables effective context windows up to 128k tokens during inference, ensuring long-range dependencies—such as footnote definitions referenced across page boundaries—remain accessible during JSON generation.
- **Rotary Positional Embeddings (RoPE):** Accurate table extraction depends critically on understanding relative positioning. The RoPE implementation in Llama-3.3 (Su et al., 2024) encodes position through rotation matrices, enabling superior generalization to relative positions compared to absolute encodings. This is essential for "structure-from-sequence" reasoning: the model must infer that a <td> element appearing hundreds of tokens downstream in the linearized HTML stream corresponds to the same vertical column as the header defined at the start of the context.

3.3 Instruction-Following and Schema Adherence

We specifically selected the Instruct variant of Llama-3.3, as tabular extraction constitutes a constraint satisfaction problem rather than open-ended generation. The output must conform to a strict, customer-specific JSON schema to be ingested by downstream SQL databases without normalization errors.

The Instruct variant, fine-tuned via Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) (Dubey et al., 2024), exhibits strong format adherence. In our experiments, the 70B-Instruct model achieved a schema validation failure rate of <0.5% (e.g., malformed JSON). By contrast, the 3B model produced syntactically invalid output in 12% of extractions. This high adherence rate in the 70B model is critical for our "Active Feedback" architecture, as it minimizes false negatives in the validation layer.

3.4 Distillation Training Protocol

Training Data Curation The distillation corpus was constructed from 50 clinical trial protocols spanning oncology, cardiology, and immunology therapeutic areas. For each protocol, the Teacher model (Gemini 2.5 Pro) processed raw HTML-rendered SoA pages and generated structured JSON extractions following a standardized schema. This Teacher-generated output served as the supervised fine-tuning target, effectively encoding the visual reasoning capabilities of the multimodal Teacher into text-processable training pairs.

The training data comprised input-output pairs where:

- **Input**: System prompt + linearized HTML content from SoA table pages + associated footnote pages.
- **Output**: Structured JSON containing visits, procedures, schedule matrix, and footnotes.

Parameter-Efficient Fine-Tuning

Configuration Given the computational constraints of fine-tuning a 70B parameter model, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) rather than full parameter updates. This approach injects trainable rank-decomposition matrices into the attention layers while freezing the base model weights, dramatically reducing memory requirements and training cost.

Table 3: LoRA Fine-Tuning Hyperparameters

Parameter	Value
Base Model	meta-llama/Llama-3.3-70B-Instruct
Fine-tuning Method	LoRA + SFT
LoRA Rank (r)	32
LoRA Alpha (α)	64

LoRA Dropout	0.05
Trainable Modules	all-linear
Learning Rate	8.0×10−5
LR Scheduler	Cosine (warmup ratio: 0.1)
Batch Size	1
Epochs	6
Weight Decay	0.01

Training was conducted on the **Together.ai** platform, completing in approximately **51 minutes** at a total cost of **$7.22 USD.** This cost-efficiency demonstrates that enterprise-grade distillation is achievable without dedicated GPU infrastructure—a critical consideration for pharmaceutical organizations seeking to prototype AI solutions rapidly.

3.5 Evaluation Framework: LLM-as-Judge

To assess extraction quality beyond simple string matching, we employed an LLM-as-Judge evaluation framework (Zheng et al., 2023) using Claude Sonnet 4.5 as an independent evaluator. This approach evaluates semantic similarity between model outputs and the Gemini ground truth across multiple dimensions:

- **Overall Similarity:** Holistic assessment of structural and content fidelity (0-100%).
- **Token-Level F1 Score:** Overlap of extracted entities with ground truth.
- **JSON Structure Presence:** Binary assessment of schema compliance.
- **Visit Extraction Accuracy:** Correct identification of visit columns.
- **Hallucination Count:** False positive visits not present in the source.

The held-out test set comprised 10 protocols not seen during training, including deliberately challenging examples with sparse tables, complex footnote dependencies, and multi-page SoA structures.

4.0 Experimental Results & Analysis

4.1 Overall Distillation Effectiveness

Table 4 summarizes the aggregate performance across the 10-protocol test set, comparing the zero-shot Base Model, the Fine-tuned Student, and Claude Sonnet 4.5 (included as a reference point for the frontier).

Table 4: Comparative Performance Across Model Architectures

Metric	**Llama-3.2-3B (FT)**	**Llama-3.3-70B-Instruct (Base)**	**Llama-3.3-70B-Instruct (FT)**	**Gemini 2.5 Pro**
Selected Student	Comparison Baseline (Small)	Zero-Shot Baseline	Selected Student	Teacher (GT)
Parameters	3B	70B	70B	— (API)
Visit Accuracy	25%	38%	75.5%	90%
JSON Schema Compliance	30%	30%	90%	100%
Hallucinated Visits	High	10	4	0
Inference Latency	~2s	~12s	~12s	~1m

Table 5: Aggregate Performance Metrics (n=10 protocols)

Metric	Base Llama-70B	Fine-tuned 70B	Δ Improvement
Overall Similarity to Teacher	20.28%	38.70%	+91% relative
Token-Level F1 Score	19.33%	39.81%	+106% relative
JSON Structure Compliance	30%	90%	+60 percentage points

Reference: Claude Sonnet 4.5 achieved 80.18% similarity and 79.40% F1 against the same Gemini ground truth.
The most significant finding is the JSON Structure Compliance metric. The base Llama-70B model, despite its scale, produced 30% valid JSON outputs when prompted for structured SoA extraction—defaulting instead to conversational prose or malformed partial structures. After distillation, the Student model achieved 90% schema compliance, rendering its outputs directly consumable by downstream clinical data pipelines.

4.2 Detailed Protocol Analysis: The Challenge Case

To examine fine-grained extraction behavior, we conducted a detailed analysis on Protocol AZ-RU-00004, selected for its complex multi-page SoA structure with extensive footnote dependencies.

Table 6: Protocol AZ-RU-00004 Detailed Metrics

Metric	Gemini (Teacher)	Base Model	Fine-tuned Student	Improvement
Visits Extracted	6 (Correct)	16 (Wrong)	8 (Close)	+80% Error Reduction
Visit Accuracy	100%	38%	75%	+97% Relative
Hallucinated Visits	0	10	4	-60% Reduction
Procedure Detail	Full	Truncated	Full	Restored Detail
Uncertainty	Marks "Unknown"	Fabricates Data	Groups Logically	Safer Behavior

Key Observations:

- **Hallucination Suppression:** The base model hallucinated 10 non-existent visit columns—a critical data integrity failure in clinical contexts where phantom visits could propagate to study databases. Fine-tuning reduced hallucinations by 60%, demonstrating that the Student learned to attend to explicit textual signals rather than predicting "probable" clinical timelines.
- **Procedure Detail Recovery:** The base model truncated procedure descriptions, losing clinically relevant specificity. The fine-tuned model restored full procedure names, matching the Teacher's detail level.

4.3 The "Reasoning Collapse" in Smaller Models

To establish the minimum viable parameter count for clinical SoA extraction, we attempted distillation into Llama-3.2-3B using identical training data and methodology.

Table 7: Model Size Ablation

Model	Parameters	SoA Accuracy	JSON Valid	Latency
Gemini 2.5 Pro (GT)	—	100%	100%	1m
Llama-3.3-70B-Instruct (FT)	70B	75%	90%	12s

Llama-3.2-3B-Instruct (FT)	3B	25%	30%	2s

The 3B model exhibited **"Reasoning Collapse"**—while achieving faster inference, it failed to maintain structural coherence for complex tables, dropping below acceptable accuracy thresholds. Specific failure modes included **Attention Drift** (lost column alignment in tables >15 rows) and **Hallucinated Checkmarks** (phantom markers in sparse columns). This negative result establishes a **Capacity Floor**: approximately 70B parameters are required to retain the abstract reasoning patterns distilled from the multimodal Teacher.

5.0 Conclusion & Future Work

5.1 Summary of Contributions

This paper presented a Knowledge Distillation framework for clinical Schedule of Activities extraction, demonstrating three key findings:

- **Distillation enables production deployment:** The base Llama-3.3-70B-Instruct model achieved only 30% JSON schema compliance; fine-tuning on Teacher-generated data improved this to 90%—transforming an unreliable model into a production-viable system.
- **A parameter floor exists:** Models below 70B parameters exhibit "Reasoning Collapse" on complex tabular structures. The fine-tuned 3B model achieved only 25% visit accuracy with 30% schema compliance, establishing minimum viable scale for clinical document reasoning.
- **A measurable gap remains:** The fine-tuned Student achieved 75% visit accuracy versus the Teacher's 90%, representing a 15-percentage-point gap that must be addressed before standalone deployment in clinical workflows. While 75% accuracy is not sufficient for fully autonomous extraction, the high JSON schema compliance (90%) allows this model to serve as a 'pre-fill' engine for human reviewers, reducing manual entry time by approximately 60-70% while maintaining a human safety layer.

5.2 Limitations & Future Work

This study has several limitations. Results are based on 10 test protocols; larger-scale validation across therapeutic areas would strengthen generalizability claims. Furthermore, our experiments focused on the extremes of the parameter spectrum (3B vs. 70B); future work should investigate the 8B-13B parameter range to refine the precise boundary of the observed "Reasoning Collapse. Additionally, distillation from a single Teacher (Gemini 2.5 Pro) may encode model-specific biases.

To close the accuracy gap, future work should explore:

- **Direct Preference Optimization (DPO):** Aligning Student outputs with human preferences rather than Teacher mimicry alone may improve extraction quality on edge cases.
- **Multi-Teacher Ensembles:** Distilling from multiple frontier models (Gemini, Claude, GPT-4) could reduce single-source bias and improve robustness.
- **Curriculum Learning:** Training on progressively harder protocols may improve generalization to complex SoA structures.
- **Multi-Modal Student Models:** Emerging vision-language models (e.g., Qwen-VL) may enable direct visual reasoning, bypassing the linearization bottleneck entirely.

6.0 REFERENCES

1. Wei, J., et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research (TMLR).
2. Dubey, A., et al. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783.
3. Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.
4. Ainslie, J., et al. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv preprint arXiv:2305.13245.
5. Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 35, 27730-27744.
6. Su, J., et al. (2024). RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing, 568, 127063.
7. Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.
8. Kim, Y., & Rush, A. M. (2016). Sequence-Level Knowledge Distillation. EMNLP 2016.
9. Wang, Y., et al. (2023). Self-Instruct: Aligning Language Models with Self-Generated Instructions. ACL 2023.

7.0 CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Angu S Krishna Kumar
Company: Saama
Email: angu.krishnakumar@saama.com
Website: https://www.linkedin.com/in/angu-s-k-53a05668/

Author Name: Ramanathan Srinivasan
Company: Saama
Email: ramanathan.srinivasan@saama.com
Website: https://www.linkedin.com/in/ram-srinivasan/