

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

Presenters:

Angu S Krishnakumar
Ramanathan Srinivasan

March-2026, Phuse , Austin





Why Schedule of Activities Are Not Simple Tables

- Merged cells — row/column structure doesn't survive copy
- Rotated headers — text runs vertically, breaks extraction
- Footnotes — critical visit rules buried at the bottom

Protocol Activities	Screening	On-Treatment Period: (One Cycle =28 Days)							Post-Treatment Period		
	≤28 Days Prior to Enrollment	Cycle 1			Cycle 2		Cycles ≥3		End of Treatment ^[1]	Short-Term Follow-Up (Day After Last Dose 30±3, 60±3, 90±3) ^[2]	Long-Term Follow-Up (Every 12 Weeks ±14 days) ^[3]
		Day 1	Day 8 (Phase 1b ONLY)	Day 15	Day 1	Day 15	Day 1	Day 15			
Visit Window (Days)				±2	±2	±2	±2	±2	+7	±3	±14
Triplicate 12-Lead ECG ^[17]	X	X									
Single 12-Lead ECG ^[17]					X		X		X		
Ophthalmic Examination	X ^[18]				X ^[18]		Every 8 weeks (±7)		X	X (30 day visit only) ^[19]	

Footnotes for Safety and Efficacy Assessments (Phase 1b and Phase 2) Schedule of Activities:

1. **End of Treatment:** Perform tests/procedures if not completed during the previous 7 days.
2. **Short-Term Follow-up:** All patients will be followed for safety every 30 days (±3 days) through 90 days after the last dose of study treatment or until the start of new anti-cancer treatment, whichever occurs first. If the patient has discontinued from study treatment for a reason other than disease progression, the patient should continue to undergo tumor assessments until disease progression regardless of initiation of a subsequent anti-cancer therapy. Assessments during this period will be conducted according to the SOA or as clinically indicated. See [Section 6.4](#) for further details.

- 4-6 hours manual work — or \$5K per month in AI bills.



Why Accuracy, Cost, and Scale Don't Coexist Today

“Today's Reality”

Approach	Accurate on SoA	Scales
Rule-based extraction	✗	✓
Frontier AI (GPT-4, Gemini)	✓	✗ (cost >~\$5K per month)
Manual review	✓	✗

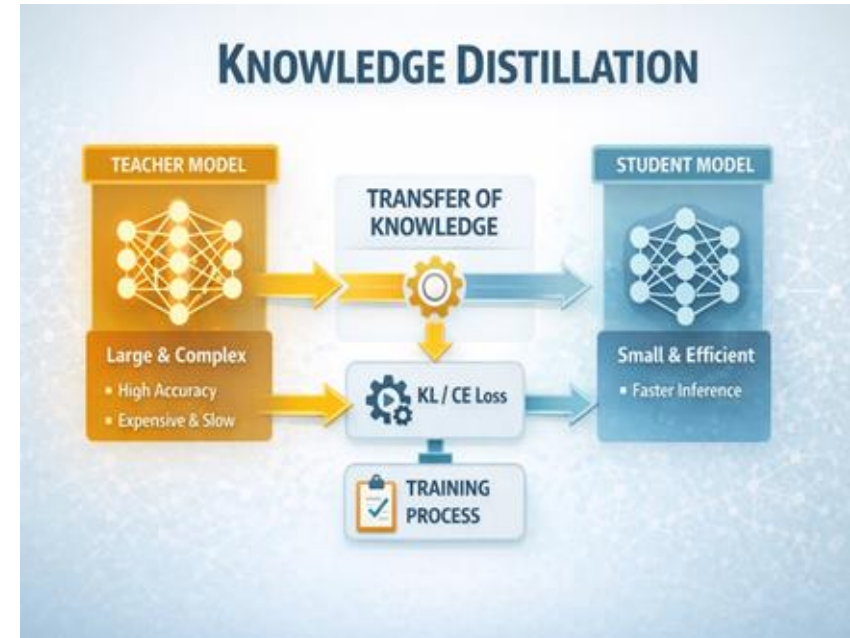
What if you could get the accuracy without the high cost?



Changing the Economics Without Losing Accuracy

- Approach: Knowledge Distillation (teacher → student)
- Teacher (Gemini, GPT-4, Claude) demonstrates how to extract complex SoA tables
- Student model learns patterns and runs locally at scale

Google ships Gemini to your phone: Nano learned from Pro and Ultra



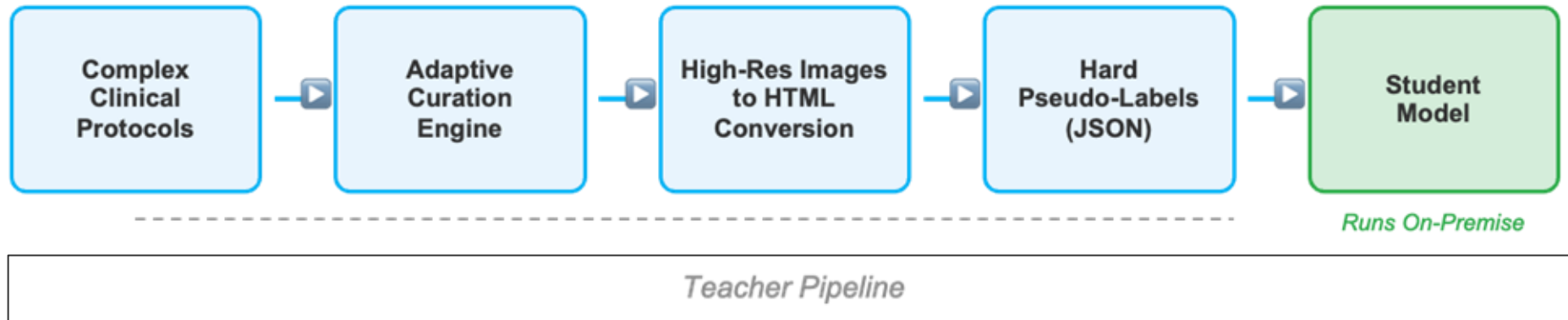


The Teacher Pipeline

Phase 1: Building the Training Data

- Choosing the Teacher: Your Student is only as good as your Teacher
- Gemini 2.5 Pro chosen for multimodal table understanding.
- Outputs validated against protocols ($\geq 90\%$ SoA accuracy)

PHASE 1: Extraction & Training

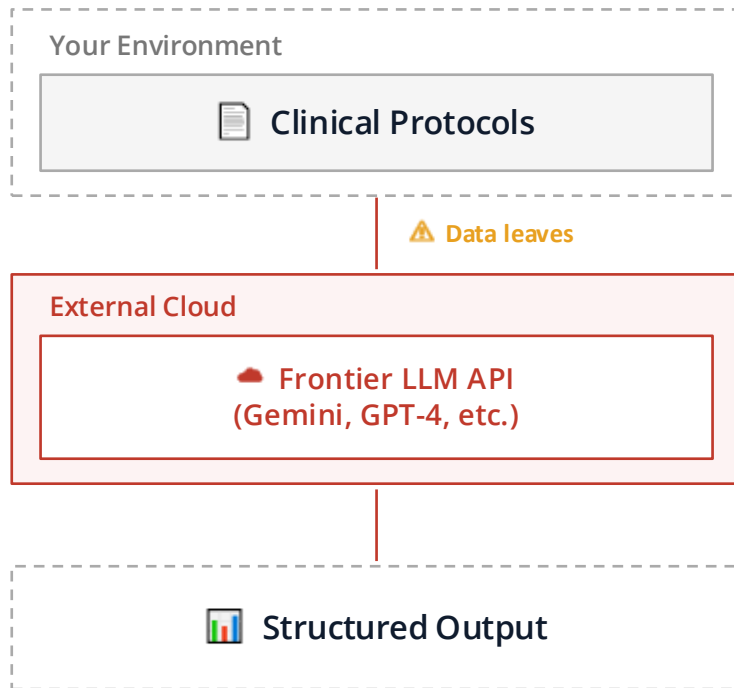




Why Privacy Matters in Clinical AI

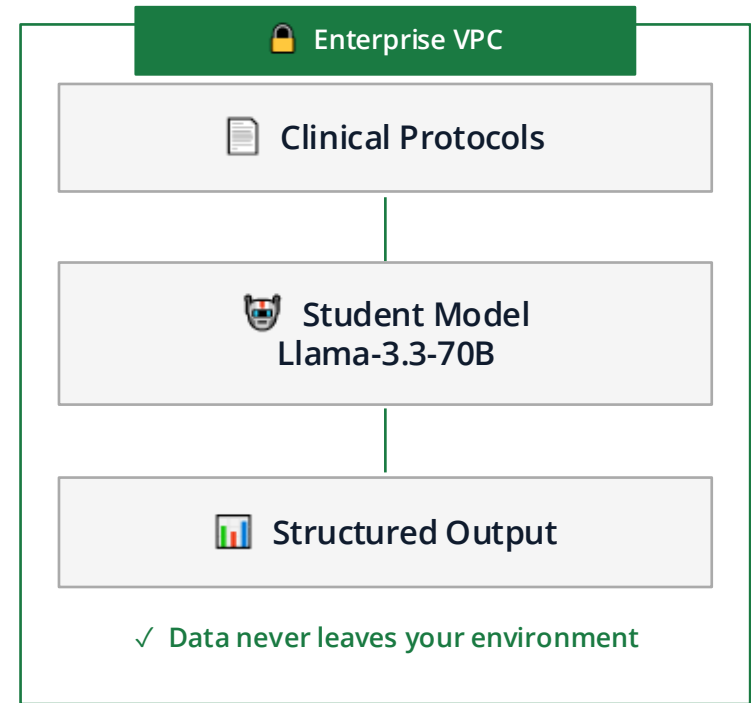
- Sensitive clinical data stays inside the enterprise boundary.

Traditional API Approach



VS

Knowledge Distillation



- ✗ Data leaves your environment
- ✗ Compliance documentation burden
- ✗ Third-party dependency

- ✓ Full data sovereignty
- ✓ Simplified audit trail
- ✓ No external dependencies



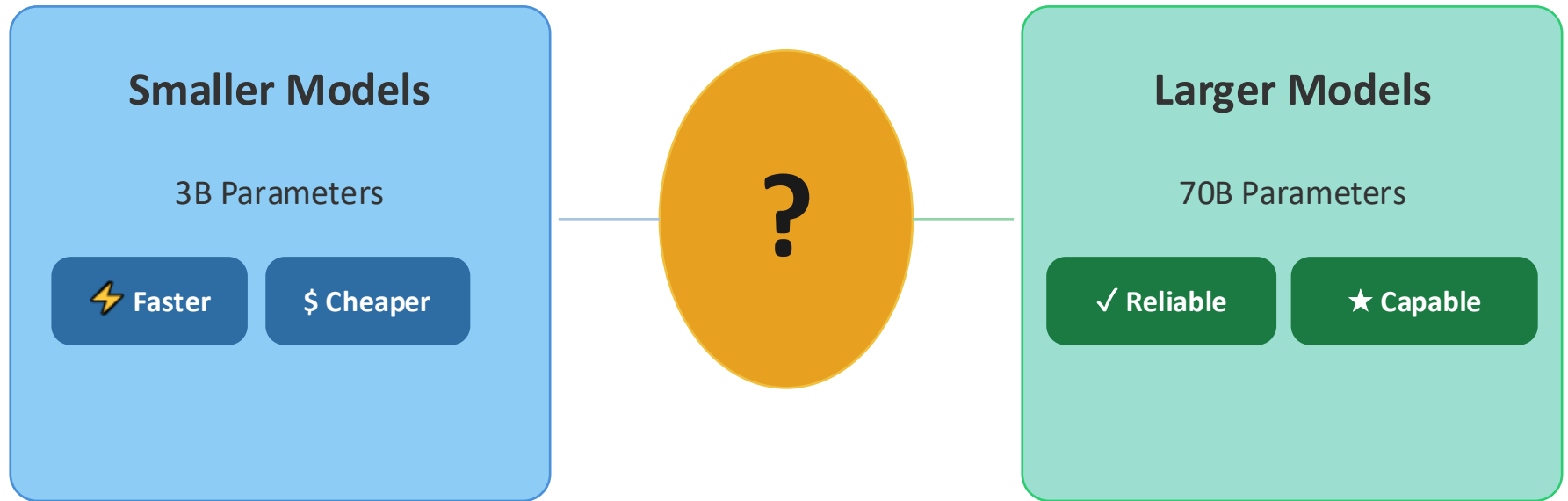
Why Llama-3.3 Was a Strong Fit

- **Designed for long documents**
 - 128k token context window
- **Strong structural understanding**
 - Maintains row/column alignment across complex tables
- **Reliable structured output**
 - Tuned to follow instructions and produce valid JSON
- **Basic understanding of clinical trial domain data**



Finding the Right Size

Llama-3.3 comes in multiple sizes: 3B, 8B, 70B



The Key Question How small can we go before reliability becomes unacceptable for clinical workflows?

Our Approach: 3B vs 70B under identical training conditions



Training: Llama-3.2-3B Instruct Experiment

Llama-3.2-3B-Instruct — Fine-tuned on identical training data

25%

Visit Extraction
Accuracy

JSON Schema Compliance: 30%

Observation:

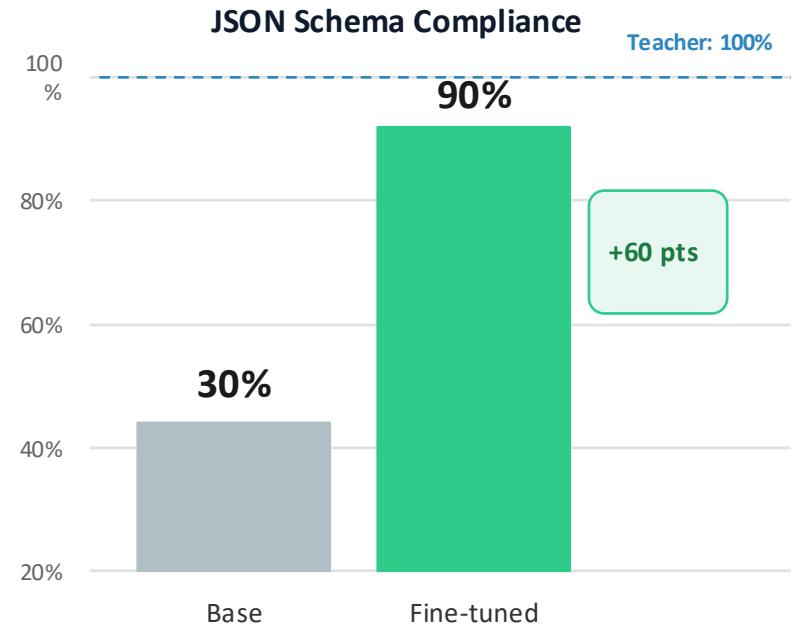
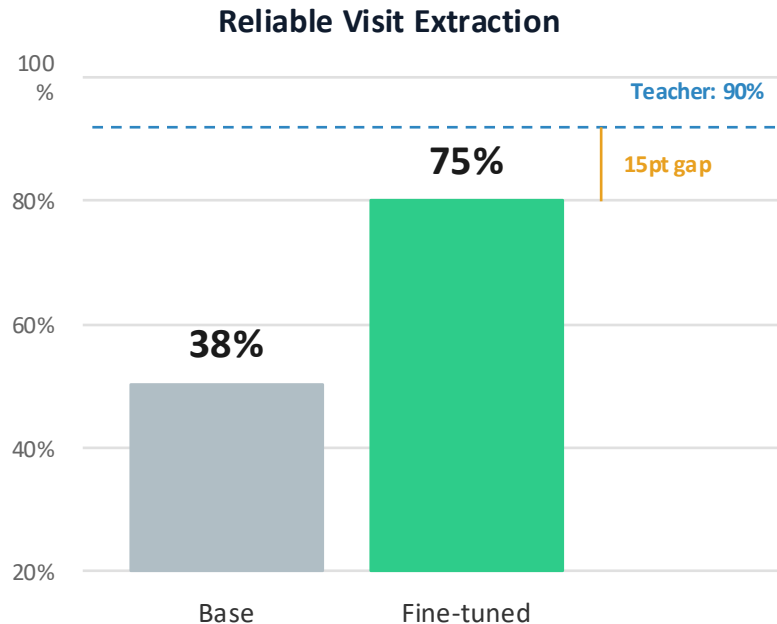
- Attention Drift
 - Lost column alignment in tables > 15 row
- Hallucinated Checkmarks
 - Inserted phantom "X" markers in empty cells

Conclusion: Below threshold for clinical workflows



Training: Llama-3.3-70B Instruct Experiment

Llama-3.3-70B-Instruct — Fine-tuned on identical training data



✓ What Improved

60% reduction in hallucinated visits

⚠ What Remains

15-point accuracy gap to Teacher

Key Finding Distillation transforms unreliable (30%) into production-viable (90% compliance). 70B is the minimum viable size for reliable clinical table extraction.

Validation Results- 70B Model



1.3 Schedule of Activities

Table 2 Schedule of Activities

Procedure	Visit number Day of study	Screening ^a (up to 14 days before Day 1)		Double-blind Treatment Phase ^a		Unscheduled Visit or PDV ^b (if applicable)	Follow-up TC (3 to 5 days after V4 or PDV)	Details in protocol section or appendix
		1	2	Period 1	Period 2			
		-14 to -8	-7 to -2	1	8±6			
Informed consent ¹		X						

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



Case Study: Protocol AZ-RU-00004 — Complex multi-page SoA with extensive footnote dependencies

Business Impact & What We Learned



- **What This Enables:**

- Pre-fill engine for human reviewers — reduces manual data entry by 60-70%
- On-premise deployment — clinical protocols never leave your environment
- Predictable processing — no API rate limits or third-party dependencies

- **What We Learned:**

- A capability floor exists — 3B models fail completely on complex clinical tables
- 70B parameters is the minimum threshold for reliable extraction
- High schema compliance (90%) matters more than perfect accuracy for downstream pipelines



Limitations & Future Work

- **Current Limitations:**
 - Validated on 10 protocols — broader testing needed
 - 15-point gap to Teacher — not ready for autonomous deployment
 - Single Teacher model — potential for encoded biases
- **Future Directions:**
 - Multi-Teacher distillation (Gemini + Claude + GPT-4) to reduce bias
 - Direct Preference Optimization (DPO) to close accuracy gap

Key Takeaways



1	Knowledge Distillation Works for Clinical Documents Small models can learn complex extraction from frontier AI — without frontier costs at inference time.	30% → 90% Schema Compliance
2	Model Size Matters — There's a Capability Floor 70B parameters is the minimum threshold. Smaller models exhibit "reasoning collapse" on complex clinical tables.	3B: 25% ✗ 70B: 75% ✓ Visit Accuracy
3	Human-in-the-Loop is a Feature, Not a Limitation 75% accuracy with 90% schema compliance = a pre-fill engine that maintains human oversight for regulatory compliance.	60–70% Manual Entry Reduced

Thank You!

To learn more, email us at
angu.krishnakumar@saama.com
ramanathan.srinivasan@saama.com

