Named Entity Recognition (NER) – Model Evaluation & Analysis

1. Gold Labeling Approach

The gold dataset was manually curated based on human interpretation of each conversation. Each chat was independently reviewed and labeled to reflect the most explicit and unambiguous user intent.

Assumptions Made
• Only explicitly stated information was labeled.
• Budgets were normalized to absolute INR values.
• Buying timelines expressed in months were converted to weeks (1 month = 4 weeks).
• Professions were labeled conservatively (e.g., “business owner” vs inferred roles).
• Relative dates (e.g., “next Saturday”) were made null.

Handling Ambiguity
• Vague timelines (“soon”, “in future”) → null
• Implicit professions (“I run a shop”) → null unless explicit
• Location expansions (“Kandarpada, Mumbai”) were normalized only if city was explicitly mentioned

Final Gold Dataset (20 Chats – Hard Labeled)

Chat	First, Last Names	Budget	Current Location	Preferred Location	Timeline (Weeks)	Profession	Visit Date
01	rahul mehta	25000000	goregaon west	andheri west	12		
02	anjali	20000000	thane	powai	24		
03	amit	5500000	Kandarpada	mira road			
04	suresh kumar	11000000	panvel	Vashi			
05	arjun malhotra	30000000	delhi	lower parel	10		
06	rakesh	16000000	kandivali east	malad west	10	service sector	2026-03-12
07	sunil	18000000	ghatkopar	chembur east		business	2026-03-15
08		14000000	borivali east	borivali west	48	retired	2026-04-05
09	kunal	30000000	Worli	lower parel	1	finance	2026-03-20
10		10000000	mulund	thane west			2026-02-28
11	rahul sharma	30000000	goregaon west	andheri west	12	it services	2026-03-18
12	anjali mehta	22000000	thane west	powai		private bank	2026-04-02
13	sandeep kulkarni	27000000	baner, pune	lonavala	24	manufacturing business owner	2026-03-25
14	kamla deshpande	12000000	nerul	vashi	16	retired	2026-02-20
15	arjun verma	35000000	worli	lower parel	24	consulting firm	2026-03-10
16	rohit malhotra	30000000	goregaon east	andheri west	10	private company	2026-03-08
17	ajay verma	21000000	kanjurmarg	powai	8	business owner	2026-03-05
18	rahul mehra	30000000	goregaon East	andheri west	16	it services	2026-03-22
19	ananya iyer	22000000	mulund West	powai	20	private bank	2026-04-06
20	suresh patil	27000000	wakad, pune	lonavala	24	logistics business owner	2026-03-30

2. Model Comparison Results

Entity	gpt-4.1-nano	gpt-4.1-mini	gpt-5-mini
Email / Phone	100%	100%	100%
Budget	83.3%	96.7%	90.0%
Timeline	66.7%	56.7%	70.0%
Profession	45.0%	50.0%	58.3%
Visit Date	88.3%	90.0%	95.0%

Latency & Token Usage
Exact token and latency values were not logged. Relative behavior observed:
• gpt-4.1-nano → Lowest latency & cost
• gpt-4.1-mini → Balanced
• gpt-5-mini → Highest latency but best semantic accuracy

3. Error Analysis

Common Failure Patterns
• Timeline hallucination from vague phrases
• Profession paraphrasing (“business” → “business owner”)
• Location over-expansion (adding city names)
• Date normalization errors for relative dates

4. Final Recommendation

Best Model + Prompt
gpt-5-mini with Prompt 3

Why it Worked Best
• Better semantic grounding
• Reduced over-paraphrasing
• Strong handling of temporal expressions

Trade-offs
Accuracy ↑ vs Cost ↑ vs Latency ↑

Future Improvements
• Entity normalization layer post-extraction
• Ontology-based profession mapping

- Explicit timeline resolution module