Ensemble Bayesian Inference

Leveraging Small Language Models
to Achieve LLM-level Accuracy in Profile Matching Tasks

김찬우

August 6, 2025

Department of Statistics, Seoul National University

목차

Introduction

Introduction

- **Large Language Models (LLMs)** shows human-level accuracy in medical diagnosis
- Potential in cognitive tasks (e.g., diagnostic summarization)
- AI in psychology: Potential to replace human judgment?

Introduction

- Existing evaluation tasks focus mainly on surface-level accuracy.
- Such tasks fail to assess whether a model can make **human-like judgments**.
- We suggest:
  - structured matching task to evaluate human-like judgments
  - Ensemble model with SLM

Ensemble Bayesian Inference Framework

EBI Framework: Core Computation

For each small language model (SLM), we compute:

$$J_{ij}^{(1)} = s_{ij}^{(1)} \cdot P(a_j \mid b_i)^{(1)}$$

$$J_{ij}^{(2)} = s_{ij}^{(2)} \cdot P(a_j \mid b_i)^{(2)}$$

$$\vdots$$

$$J_{ij}^{(N)} = s_{ij}^{(N)} \cdot P(a_j \mid b_i)^{(N)}$$

- $s_{ij}^{(n)}$: confidence score output by SLM_n
- $P(a_j \mid b_i)^{(n)}$: estimated match likelihood from SLM_n
- $J_{ij}^{(n)}$: weighted judgment of candidate a_j for input b_i

EBI Framework

Type 1 Prompt (Answer Likelihood)

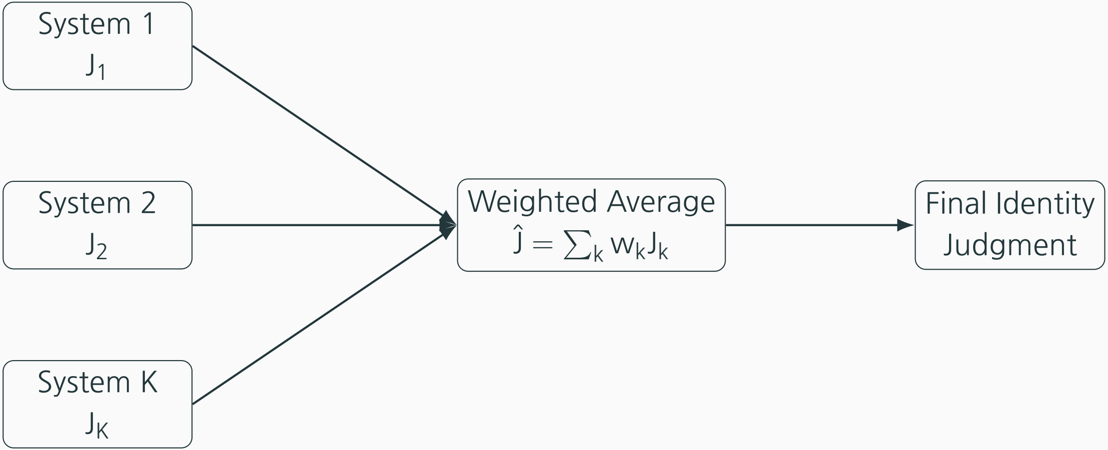
- Given input b_i and candidate a_j
- Ask the model to choose the best matching a_j
- Estimate $P(a_j \mid b_i)$ by frequency $\rightarrow c_{ij}/ni$

Type 2 Prompt (Confidence Estimation)

- Given b_i and candidates $\{a_j\}$, ask the model to rank them or assign confidence scores.
- Convert ranks or scores into a normalized confidence vector.
- s_{ij} reflects the model's subjective confidence: $\sum_j s_{ij} = 1$

EBI Framework

article tikz



Baseline Comparison Models

Baseline Model: Feedback-Reflect-Refine Mechanism

Key Characteristics:

- Step-by-step elimination based on pairwise judgment.
- Wrong early eliminations cannot be recovered later.
- Recursive review is used to correct inconsistencies.

Processing Structure
The model cycles through **Feedback** → **Reflect** → **Refine** until all conflicts in judgment are resolved.

Baseline Model: Sequential Processing Steps

Step-by-step Flow:

1. **S1 — Initialization (System Prompt)** Filter candidates A using demographic info (e.g., age). Update Aset; reset session if changed.
2. **S2 — Tournament-style Comparison** If multiple candidates remain ($n \geq 2$), compare and eliminate them sequentially.
3. **S3 — Recursive Review** When few candidates remain, re-check prior decisions for contradictions or inconsistencies.
4. **S4 — Conflict Resolution** If inconsistency is detected, re-evaluation is triggered. Loop continues until conflict is resolved.

Data Construction and Evaluation

Data construction

Source Data:

- 50 individuals' aptitude test results
- 14 psychological items: sociability, self-reflection, task persistence, risk avoidance, and others

Profile Generation via GPT-4o:

- **Profile A:** Generated with prompts encouraging self-improvement and personal growth → Perspective: ”Enhancing future performance”
- **Profile B:** Generated with prompts focused on work execution and professional evaluation → Perspective: ”Task behavior and observable weaknesses”

Experimental Results

Category	Description
Strength(A)	Highly responsible, perseverant
Weakness(A)	Prone to stress, averse to change
Assessment(A)	The person has a strong sense of responsibility and perseverance to complete assigned roles. They may feel uneasy about adapting to changes, but by clearly defining goals, schedules, and expectations, they can effectively lead team ...
Personnel(B)	As a manager, the employee leads the team and delivers the expected results. To further enhance the overall output of the team, please focus on strategic goal setting, progress management, and member development. It is also ...

Data construction

Design Principle:

- A B has 1 to 1 matching profile
- Different prompts ensure preventing simple word-matching
- Enables evaluation of deeper inference, not surface similarity

Evaluation Metrics

- **Accuracy (Acc)**

$$\mathrm{Acc} = \frac{n_c}{N} \quad (n_c\text{: number of correct matches, N: total samples})$$

- **Lift (Improvement over Human)**

$$\mathrm{Lift} = 100\left(\frac{n_c}{H} - 1\right) \quad (\text{H: number of human correct matches})$$

- **Reach (Relative to Reference)**

$$\mathrm{Reach} = 100 \cdot \frac{n_c}{\mathrm{Base}} \quad (\text{Base = H or G : G is number of LLM correct matches})$$

Experimental Results and Conclusion

Experimental Results

Table 3: Results of single BI systems for prof1_j (Japanese Aptitude Assessment).

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
37	gemma2-9b-it	t1∗-100	t2′-10	23	21.1%	104.5%
42	llama3-8b-8192	t1∗-100	t1∗-100	23	21.1%	104.5%
76	llama3.1-70b-versatile	t1∗-100	t2-10	23	21.1%	104.5%
64	gpt-4o-mini-2024-07-18	t2-10	t2-10	21	10.5%	95.5%
43	llama3-8b-8192	t1∗-100	t2′-10	21	10.5%	95.5%
25	gemma2-9b-it	t1∗-100	t2′-10	20	5.3%	90.9%
28	llama3-8b-8192	t1∗-100	t1∗-100	20	5.3%	90.9%
46	llama3-70b-8192	t1∗-100	t2-10	20	5.3%	90.9%
66	gpt-4o-mini-2024-07-18	t1∗-100	t2-10	18	-5.3%	81.8%
40	mixtral-8x7b-32768	t1∗-100	t2′-10	18	-5.3%	81.8%
12	mixtral-8x7b-32768	t1-500	t1-500	18	-5.3%	81.8%
13	llama3-70b-8192	t1∗-500	t1∗-500	17	-10.5%	77.3%

Experimental Results

Table 4: Results of EBI (ensemble systems) for prof1_j* (Japanese Aptitude Assessment), limited to top-performing systems (Lift$\geq$ 0).

system	components	weights	n_c	Lift	Reach
83,81	{37,40,43,46}	[1,1,1,1],[1,1,2,3]	26	36.8%	118.2%
50	{12,13,25,28,37,40,43,46}	[3,2,1,1,1,1,2,3]	25	31.6%	113.6%
55	{12,13,25,28,37,40,43,46}	[3,2,1,1,5,1,2,3]	23	21.1%	104.5%
78	{37,43,45,66,76}	[30,3,1,1,10]	23	21.1%	104.5%
71	{37,43,66}	[1,1,1]	22	15.8%	100.0%
82,84	{37,40,43,46,66,76}	[1,1,2,3,1,1],[1,1,1,1,1,1]	20	5.3%	90.9%
85	{37,40,42,46,64,59}	[1,1,1,1,1,1]	19	0.0%	86.4%

Conclusion

1. **Utilization of Weak Learners**
   Even SLMs with negative Lift contributed positively when appropriately included in ensembles.
2. **Effectiveness of the EBI Method**
   Weighting based on subjective scores (e.g., s_{ij}, c_{ji}) improved ensemble performance over simple averaging.
3. **Versatility Across Tasks and Languages**
   EBI-based SLM ensembles achieved consistent improvements across tasks (e.g., aptitude, purchase) and languages (Japanese, English).

Discussion

Discussion

- The model utilizes a Bayesian-like formulation ($s_{ij} \cdot P(a_j \mid b_i)$), but no actual Bayesian posterior estimation is performed.
- Unclear whether the task actually tests reasoning
- **Need for a metric to assess dataset matching difficulty and reasoning demand**

Appendix

Appendix B.1 - Prompt Type 1

##Analysis Approach
*Emulate human thinking processes and conduct qualitative analysis to draw conclusions.
*Directly interpret the data and make intuitive inferences from the context and expressions.
*Analyze the individual's behavioral traits, professional abilities, and personal characteristics in detail based on the comment from ## Personnel Evaluation Findings of id_B, and estimate the profile.
*Compare the inferred profile with the comment from ##Aptitude Assessment Findings of id_A and select the candidate id_A that most closely matches.
##Execution Method
*Describe the process of selecting the candidate id_A that most closely matches the inferred profile.
*Once the matching candidate id is found, output that id.
*Output the matching candidate id according to the specified ## Output Format.
##Output Format
Describe the process of selecting the candidate id_A that most closely matches the inferred profile.
id_B:{id_B number}, id_A:{matching candidate id_A number}
##Aptitude Assessment Findings
Comments from the assessment test for id_A. The data is as follows:
{id_A, Assessment(A) [repeat 7 sample data]}
##Personnel Evaluation Findings
Comments from the personnel evaluation for id_B. The data is as follows:
{id_B, Personnel evaluation(B) [repeat target id data]}
Based on the above requirements, please output the matching id according to the output format.

Appendix B.2 - Type 2 Prompt

##Guidelines
*Mimic human thought processes and derive results through qualitative analysis.
*Read the content of the data directly and intuitively infer from its context and expressions.
*Based on the Personnel evaluation of id_B, analyze the person's behavioral characteristics, professional abilities, and personal traits in detail to infer the persona.
*Compare the inferred persona with the Assessment test of id_A to determine the certainty level of a match.
##Detailed Requirements
*Describe the inferred persona. Compare the inferred persona with the Assessment test of id_A to find matching candidates. Calculate the certainty level (in percentage).
*List the matching candidate id_As in order of highest certainty level.
*Output up to the 7 matching candidate id_As.
*Display the certainty level next to each matching id_A.
*Output the results for all id_B (7 in total) in the specified ##Output Format without omitting any steps.
##Evaluation Method for Certainty Level
High certainty (e.g., 0.9 - 1.0): A very clear match between both texts.
Medium certainty (e.g., 0.5 - 0.8): Some commonalities exist, but it is not a perfect match.
Low certainty (e.g., 0.1 - 0.4): Not very confident, but it is a possible match.
Very low certainty (e.g., 0.0): Little to no matching points between the texts.
##Output Format
id_B:{id_B number} {Description of the inferred persona.}
1. id_B:{id_B number}, id_A:{matching candidate id_A number} {certainty level}
2. id_B:{id_B number}, id_A:{matching candidate id_A number} {certainty level}
(Omitted)