Advances in Artificial Intelligence and Machine Learning; Research 5 (3) 4154-4173 Received 25-04-2025; Accepted 1-08-2025; Published 07-08-2025

Ensemble Bayesian Inference: Leveraging Small Language Models to Achieve Llm-Level Accuracy in Profile Matching Tasks

Haru-Tada Sato satoh@isfactory.co.jp
Department of Data Science
i's Factory Corporation, Ltd.
Kanda-nishiki-cho 2-7-6, Tokyo 101-0054, JAPAN

Fuka Matsuzaki matsuzaki@isfactory.co.jp
Department of Data Science
i's Factory Corporation, Ltd.
Kanda-nishiki-cho 2-7-6, Tokyo 101-0054, JAPAN

Jun-ichiro Takahashi jtakahashi@isfactory.co.jp
Department of Data Science
i's Factory Corporation, Ltd.
Kanda-nishiki-cho 2-7-6, Tokyo 101-0054, JAPAN

Corresponding Author: Haru-Tada Sato

Abstract

This study explores the potential of small language model (SLM) ensembles to achieve accuracy comparable to proprietary large language models (LLMs). We propose Ensemble Bayesian Inference (EBI), a novel approach that applies Bayesian estimation to combine judgments from multiple SLMs, allowing them to exceed the performance limitations of individual models. Our experiments on diverse tasks—aptitude assessments and consumer profile analysis in both Japanese and English—demonstrate EBI's effectiveness. Notably, we analyse cases where incorporating models with negative Lift values into ensembles improves overall performance, and we examine the method's efficacy across different languages. These findings suggest new possibilities for constructing high-performance AI systems with limited computational resources and for effectively utilizing models with individually lower performance. Building on existing research on LLM performance evaluation, ensemble methods, and open-source LLM utilisation, we discuss the novelty and significance of our approach. [1]

Keywords: Bayesian Inference, Ensemble Learning, Profiling Tasks, Proprietary LLM Comparison, Small Language Models

[1] Codes and data are available at https://github.com/maskcode9004/ebi_method

Citation: Haru-Tada Sato, et al. Ensemble Bayesian Inference: Leveraging Small Language Models to Achieve Llm-Level Accuracy in Profile Matching Tasks. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(3):232.

1. INTRODUCTION

In the field of medical diagnosis, there have been reports of GPT-4 achieving higher diagnostic accuracy than physicians, either independently or as an assistant[1, 2], showing comparable performance [3], or failing to demonstrate improvement [4]. While some research emphasises the irreplaceability of specialized knowledge and experience and shows caution about AI's ability to accurately identify common sense in problem-solving, studies also suggest its effectiveness as a supportive tool [5, 6]. Various diagnostic tasks have been evaluated, including predictive accuracy from test results and case studies [1, 3], as well as diagnostic summary content [2].

In the field of psychology, comparative experiments are being conducted to determine whether AI can substitute human intellectual activities such as decision-making and cognition [7–11]. In creative ideation, AI demonstrates novelty equivalent to or surpassing humans, though it shows slight deficiencies in feasibility [7]. Psychological experiments evaluate whether AI can replace human subjects [8], or make human-like behavioral judgments [9]. Experiments involving human judgment present challenges for quantitative evaluation, including issues of fairness due to human cognitive biases [10], handling subjective concepts in evaluations [11], responsiveness to cognitive load [12], and variability in assessments [13].

Against this background, this paper addresses person identification as a judgment task in comparative experiments between AI and humans. The judgment targets are texts evaluating personal characteristics. The task involves comparing two types of evaluation comments about the same group of individuals and identifying each person, of which samples are illustrated in Appendix A. Evaluation texts contain noisy expressions or potentially misleading statements, dependent on the evaluator, posing risks of misjudgment. Improving the robustness and reliability of inference when faced with such inputs presents a significant challenge.

To execute such complex and sophisticated judgments using AI, we propose a judgment methodology that combines probabilistic judgment using Bayesian Inference (BI) with subjective elements based on confidence levels, taking into account human judgment characteristics (variability and pattern consistency [13]). While the BI method is a simplified model of human judgment processes, generative AI also exhibits judgment variability, which we aim to accommodate through an ensemble approach to BI. This is the key idea that we call our method, Ensemble Bayesian Inference (EBI).

While ensemble methods for generative AI [14–17], focus on improving model accuracy, this paper examines whether judgment diversity contributes to accuracy improvement, centring on recently emerged lightweight open-source LLMs. As addressing computational costs and latency associated with advanced prompting techniques becomes crucial — especially when targeting real-world applications with stringent performance requirements — our work also serves to validate a framework for high-volume processing leveraging the advantages of high-speed processing chips to bridge the performance gap with proprietary LLMs.

2. TASK DESIGN OF BAYESIAN INFERENCE METHOD

The task problem setting involves determining which element $A(a_j)$ corresponds to the same individual when given all information $X = B(b_i)$, with two lists of personal information $B = \{b_1, b_2, \cdots, b_n\}$ and $A = \{a_1, a_2, \cdots, a_n\}$ about a group of individuals. Here, A and B represent sets of texts describing personal profiles based on different perspectives. By "different perspectives," we mean domains that have some relationship, such as dietary habits and health hygiene.

The goal is to estimate the profiles of each person and determine whether they are the same individual through profile analysis. Technically, the main objective is to achieve judgment accuracy comparable to proprietary LLMs by obtaining frequent responses from SLMs and implementing a probabilistic judgment process. The extent to which performance can exceed human judgment is also a significant interest.

2.1 Confidence Matrix and Subjective Degree

The basic concept of our technique is Bayes' theorem. The conditional probability of event A occurring after event X (posterior probability) $P(A|X)$ is given by likelihood $P(X|A)$ and $P(A)$, the probability of event A occurring (prior probability or subjective probability):

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}, \tag{1}$$

where $P(X) = \sum_A P(X|A)P(A)$.

When an element b_i of $B = \{b_1, b_2, \cdots, b_n\}$ is given, the posterior probability that it corresponds to an element a_j of $A = \{a_1, a_2, \cdots, a_n\}$ provides the confidence matrix $\mathrm{conf}_{B\to A}$ (statistical or computational confidence, simply referred to as confidence when distinction is unnecessary). Its elements, with $X = B(b_i)$, are given by Bayes' theorem (1):

$$(\mathrm{conf}_{B\to A})_{ij} = P(a_j|b_i) = \frac{P(b_i|a_j)P(a_j)}{P(b_i)}. \tag{2}$$

An experimental consideration is that when aggregating responses from multiple queries to calculate confidence, elements that never appear in responses would result in $P(X) = 0$, potentially leading to divergent matrix elements in naive aggregation. Therefore, we apply regularisation to ensure $P(X)$ has a small value ε. Since $P(X|A)$ values are also proportionally small, equation (1) produces finite values. Specifically, items with no responses are assigned a value of $\varepsilon = 0.1$ in aggregation to avoid division by zero.

Conversely, when judging $B(b_i)$ from $A(a_j)$, we obtain the strength of confidence as an observed value. To distinguish this from the above-mentioned confidence, we call it subjective confidence (subjective degree). In a voting system, this could be the proportion of votes received (collective subjective degree). When this subjective degree is observed as c_{ji}, the observation matrix (subjective degree matrix), the likelihood $P(b_i|a_j)$ is given by $(\mathrm{conf}_{A\to B})_{ij} = c_{ji}$, $(0 \le c_{ji} \le 1)$. Then we obtain the subjective probability: $P(a_j) = \frac{1}{C}\sum_i c_{ji}$, where $C = \sum_{ij} c_{ji}$, yielding

$$P(b_i) = \sum_j P(b_i|a_j)P(a_j) = \frac{1}{C}\sum_{jk} c_{ji}c_{jk}. \tag{3}$$

Substituting these into (2) gives the confidence matrix.

2.2 Judgment Matrix

The problem setting is to determine which a_j corresponds to b_i when given information $X = b_i$. Simple Bayesian estimation may not work well when there is information asymmetry. Namely, probabilistic LLM approaches, such as MLM and GPT, may not perform reliably when profiler differences in skill, perspective, or background affect the interpretation of A and B.

Regardless of the method used, if the reliability of the judgment result can be expressed through some evaluation weight s_{ij}, it appears effective to define a judgment matrix J by multiplying this by the confidence degree:

$$J_{ij} = s_{ij} P(a_j | b_i) \,. \tag{4}$$

The ij component of J can be interpreted as the plausibility of each candidate a_j for b_i, so this J is used for identity judgment. This method allows for various models to be constructed based on the freedom in defining s_{ij} and the observation matrix c_{ji}.

In our implementation (see Appendix D), the weight s_{ij} is calculated similarly to the observation matrix c_{ji}, based on subjective confidence values returned by the LLM. Specifically, we queried the same profiling task 10 times and averaged the returned confidence levels to compute each s_{ij}. This approach treats s_{ij} as a smoothed and stabilised version of subjective scores.

2.3 System Settings

In this study, we use small-sized language models (SLMs) to determine whether b_i and a_j represent the same person to create s_{ij}. However, to expand the range of system options, we test multiple different LMs for calculating s_{ij} and c_{ij}. For obtaining the observation matrix c_{ij} and weight matrix s_{ij}, we consider two approaches: (i) evaluation based on collective subjective degree (id selection frequency) and (ii) evaluation based on direct subjective degree. These are realised by the following types of prompt processing:

- Type 1. Aggregating the frequency of ids obtained as responses by submitting the same question multiple times
- Type 2. Collecting values by asking for subjective confidence levels during the response evaluation process

The prompt structure varies depending on the Type, but the basic query asking which id is identical and the specification of the analysis method are common across both. What's common is the concept, not necessarily the specific descriptive expressions. The essential difference lies in the output format specification: Type 1 outputs only the pairs of ids judged identical (see Appendix B.1), while Type 2 outputs an array of candidate a_j in descending order of subjective degree along with their subjective degrees (see Appendix B.2). These aggregate values are normalised by the number of questions.

In calculating observation and weight matrices, responses will differ depending on which language model is used and what prompt (Type and descriptive expression) is given, so we consider one *language model and prompt configuration* as one system. Additionally, we consider ensemble systems (EBI) by calculating weighted averages of judgment matrices $J^{(s)}$ computed from individual systems. The final ensemble matrix J was computed as $J = \sum_s w_s J^{(s)}$, where weights w_s were selected through trial-and-error to maximise Lift for each dataset, as summarised in Appendix D.

All datasets follow a consistent folder structure on GitHub (for example, the folder for prof1e is `ebi_method/prof1e_personnel/`). For each dataset: the computation logic and normalisation steps are documented in `3.Computation_method_of_J.xlsx`, and the list of weight candidates tested appears in `1.Model_list.xlsx`.

For language model selection criteria, we primarily adopt lightweight SLMs that prioritise speed over accuracy for collecting multiple responses; however, we also consider 70b models for some cases using high-speed processing via WSE (Cerebras Wafer-Scale Engine and CS-3 system) or LPU (Groq LPU™ Inference Engine). Models used in EBI include: GroqChat: gemma2-9b-it, mixtral-8x7b-32768, llama3-8b-8192, llama3-70b-8192. Cerebras: llama-3.1-70b-versatile. OpenAI: gpt-4o-mini-2024-07-18.

3. DATA GENERATION AND EVALUATION METHODS

3.1 Data Preparation

We generate two different persona profiles for the same individuals based on items (attribute values and observed values) from multiple perspectives. We divide the items used for generation into two parts, allowing overlap, and create two types of persona data, A and B. When creating these personas, we specify different perspectives to ensure they cannot be easily linked merely by word matching (similarity). We prepared two types of datasets: aptitude assessments and purchase history. For sample data, refer to TABLE 1 and TABLE 2, in Appendix A.

For the aptitude assessment (prof1j), the specific creation process is as follows: The observational data consists of 50 people's aptitude test results with 14 observation items (items including sociability avoidance, self-reflection, task persistence, risk avoidance level, etc.). Character diagnosis comments were generated using GPT-4o based on the item names and numerical values to create dataset A. Department, years of service, and other work attributes were added to similarly generate work evaluation comments to create dataset B. In creating the comments, different perspectives were specified to ensure that A and B would not generate identical comments. For A, the purpose was to encourage self-improvement to enhance performance, while B focused on issues related to work execution. Additionally, another dataset "prof1e" was created by translating this data to English using GPT-4o.

For the purchase history data (prof2j), 18 purchased product items were divided, and persona data was created using a similar procedure as above. In this case, since the persona image could be significantly affected by the product composition in the two divisions, both A and B were created with two types of perspectives: behavioural patterns and consumer values in product selection. An English translation of this data was created as "prof2e".

The purchase data (prof2j) were selected from real-world health product e-commerce data. To create two distinct persona views (A and B), items were manually split into two groups within the same category (e.g., beverages, general supplements, and tonic products), ensuring that payment methods and purchase quantities were balanced across both groups. This controlled division was designed to avoid bias and preserve the consistency of behavioural interpretation. The English version (prof2e) was generated by translating the Japanese data while maintaining this item partition.

3.2 Evaluation Methods

For performance evaluation metrics, accuracy Acc= n_c/N (data count $N = 50$, correct answer count n_c) is commonly used, but if the background contexts of datasets A and B differ, the background knowledge of words and the range of conceptual similarities may also differ between them. In such situations with conceptual fluctuations and undefined ranges, no known method objectively quantifies the consistency of meta-concepts. Thus, a key consideration is that the scale can vary significantly due to the nature of text content, abstraction level, interpretability, and the evaluator's background knowledge, making comparison across different datasets impossible. Therefore, this metric can only be used for relative superiority judgments within the same type of data, which is a disadvantage.

Consequently, evaluating using the improvement rate (Lift), which shows how much superiority there is relative to human judgment, allows for more appropriate comparison by somewhat eliminating data dependence (although the prerequisite is that the reference evaluator should remain fixed). Given the maximum correct answer count H of the reference evaluator, Lift= $100(\frac{n_c}{H} - 1)$ [%].

When human evaluators are difficult to secure, high-precision proprietary LLMs can be substituted for human evaluators. In this case, using the maximum correct answer count G of the high-precision proprietary LLM (ChatGPT-4o), we evaluate how close we can get using the reach rate Reach= $100n_c/\text{Base}$ [%] (where Base is H for humans and G for LLMs). For the reference evaluator LLM, rather than using the EBI under examination, we adopt a judgment method that builds reasoning step by step through prompting to approach human judgment as closely as possible (see Section 4).

For our English verification, where H is unknown, we use the average H/G ratio γ from the Japanese verification, substituting $H = G\gamma$ into the Lift formula to approximate Lift as $\text{Lift}_{eff} = \text{Reach}/\gamma - 100$. We obtained H values of (19, 13) for (prof1j, prof2j) from human evaluations. For the four datasets (prof1j, prof2j, prof1e, prof2e), the obtained G values were (22, 20, 28, 23). Therefore, the H/G ratios were 19/22 for prof1j and 13/20 for prof2j, giving $\gamma = 0.757$. For English data, we use $\text{H}_{eff} = \text{G}\gamma$, thus adopting baseline H values of (19, 13, 21.2, 17.4) for the four datasets. Human Acc values were (38%, 26%, 42%, 35%), GPT-4o Acc values were (44%, 40%, 56%, 46%), and Lift values relative to humans were (15.8%, 53.8%, 32.1%, 32.2%).

4. SEQUENTIAL THINKING METHOD

Even proprietary LLMs require dynamic prompting techniques — where language models improve task performance through self-reflection — to ensure at least human-level evaluation accuracy [15, 18, 19]. Since simple task instructions can result in duplicate judgments from LLMs, this paper applies the Feedback-Reflect-Refine mechanism [15], to resolve conflicts by collecting error information from LLM outputs and using it as feedback.

The prompting uses two types of templates for solution and verification: (i) Solving Prompt: Instructs the model to explain the reason in one sentence before providing the answer corresponding to the task. (ii) Feedback Prompt: Analyzes the cause when the model's output is incorrect and generates a new prompt.

In contrast, the EBI method eliminates duplications through forceful processing rather than prompt-based judgment. Specifically, it follows the steps: 1. Select the element with the highest output value in the judgment matrix J. 2. Remove the row and column containing the selected element (i.e., form a submatrix). 3. Select the element with the highest output value in the submatrix. 4. Repeat the above steps for the number of rows in the judgment matrix (i.e., the number of individuals to be predicted).

4.1 Digest of Sequential Thinking

In the sequential method, candidates are analysed one by one within subsets filtered by attributes such as age (system prompt S1), and candidates are narrowed down tournament-style (sequential processing S2). Since an incorrectly selected initial candidate affects subsequent judgments, a flow to review the results (recursive processing S3) is incorporated when there are conflicts or inconsistencies in the judgment results. Since conflicts may also occur during the review process, examination and judgment are repeated until conflicts are resolved (conflict processing S4). Refer to Appendix C for concrete prompts S1-S4. The sequential processing outline is as follows:

1. The program obtains the set of IDs from A (Aset) that match demographic attributes such as age. The number of elements is denoted as n. When Aset changes, the message history is cleared (session switching)
2. System prompt S1: Specifies the data description and analysis task method.
3. Sequential processing S2: If $n = 1$, the identification is confirmed, and the process moves to the next Aset. If $n \geq 2$, starting with an initial value $k := n$, elements are removed from Aset with each identification until $k = 1$.
4. When the remaining number of people in Aset reaches a specified value, recursive processing S3 is activated.
5. Conflict resolution S4 is executed within the recursive processing. The loop continues until the duplicate counter reaches 0.

A step-by-step execution log of the sequential prompting (S1–S4) is available in the spreadsheet: `ebi_method/Sequential_Thinking/prof1e_4o/prof1e_33_results.xlsx` (GitHub). The `Procedure` sheet shows the flow of prompt execution, and the `Result` sheet summarizes accuracy outcomes for each ID pair.

5. RESULTS

In this study, we conducted experiments on four datasets to examine whether SLM ensemble judgment systems can achieve judgment accuracy comparable to human or proprietary LLM judgment. The complete results of BI and EBI methods for each system are listed in Appendix D.

5.1 Overall

FIGURE 1, summarises the results for Japanese datasets, and FIGURE 2, summarises the results for English datasets. Considering all four types of datasets collectively, we found that (1) single BI systems can achieve Lift>0, and (2) using EBI enables us to obtain systems with even higher Lift than single BI systems. In each Figure, considering Lift≳30%, Reach≳100% as high-performance regions, all datasets except prof1e (aptitude/English) reached these regions, and even prof1e nearly reached the lower bound of this region when using EBI (hollow markers show higher Lift values than filled markers).

Compared to other datasets, the prof1e set showed relatively lower performance. One plausible reason is the translation-induced abstraction: while the original Japanese comments included nuanced phrasing and culturally specific cues, the English-translated versions tended to generalize expressions (e.g., "task persistence" translated to "perseverance"), reducing the richness of distinguishing features. Such semantic flattening may have hindered the model's ability to make precise identity judgments.

With the exception of system 59 (llama3.1-70b) in prof2j (purchase/Japanese), we were able to achieve higher Acc with EBI systems (compared to single BI) for all datasets. It should be noted, however, that since the number of plots can be increased indefinitely by increasing the number of models and systems tested, there is no fair way to compare the averages of Acc or Lift, so it is appropriate to understand these as trends.

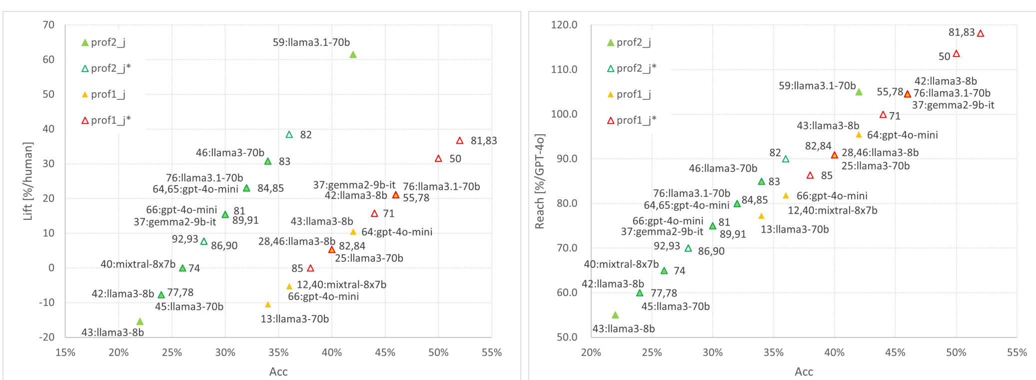

Figure 1: Results of Japanese datasets. The Lift (left) and the Reach (right) graphs for single BI systems (prof1_j and prof2_j) marked in filled triangles, and EBI systems (prof1_j* and prof2_j*) marked in hollow triangles. The component systems of EBI are listed in Appendices D.1 and D.2.

5.2 Individual Datasets

These results suggest that SLM ensembles using the EBI method can exceed the performance limitations of individual BI systems and achieve accuracy comparable to, or in some cases surpassing,

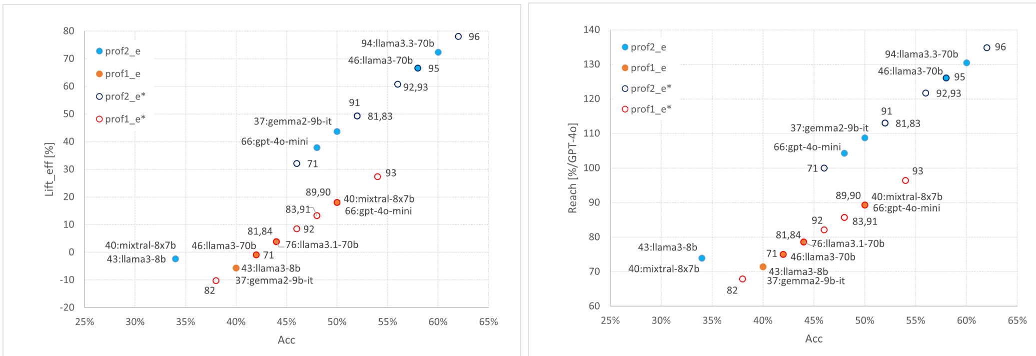

Figure 2: Results of English datasets. The Lift (left) and the Reach (right) graphs for single BI systems (prof1_e and prof2_e) marked in filled circles, and EBI systems (prof1_e* and prof2_e*) marked in hollow circles. The component systems of EBI are listed in Appendices D.3 and D.4.

that of LLMs. Furthermore, examining the performance of individual BI systems constituting the ensembles revealed that combining lower-performing systems appropriately can enhance prediction robustness. The confirmation of ensemble effects across different datasets and languages is an important finding demonstrating the versatility of this method. Specifically:

For prof1j (aptitude/Japanese), evaluation of multiple SLMs' individual performance revealed systems achieving maximum Lift values of 21% (37:gemma2-9b-it, 42:llama3-8b-8192, 76:llama3.1-70b-versatile). Details are in Appendix D.1. By constructing average systems including these high-performing single systems, the (Lift, Reach) values (hereinafter, LR-value) increased significantly from single systems to a maximum of (36.8%,118.2%). Notably, in higher accuracy ensemble systems (limited to Lift>0), instances were confirmed where combining single systems with negative Lift values (66:gpt-4o-mini-2024-07-18, 12 and 40:mixtral-8x7b-32768, 13:llama3-70b-8192, etc.) resulted in positive Lift values after ensemble. This suggests that EBI ensembles can compensate for weaknesses in individually low-performing systems, enabling more robust predictions. For example, the ensemble system shown in TABLE 4, achieves positive Lift values despite including components with negative Lift.

In the prof2j (purchase/Japanese) experiment summarised in Appendix D.2, ensemble systems were constructed centred on systems (37, 46, 64, 76) that achieved positive Lift values across two different Japanese datasets. Although Reach did not reach 100%, instances were confirmed where including weak systems (40, 42, 43) with negative or zero Lift values in the ensemble led to positive Lift values, here too. In particular, the top-performing system 82 (in TABLE 6), includes llama3-70b systems 46 and 76, and by integrating them as an average system, it achieved results (38.5%,90%) surpassing the maximum LR-value of each component system (30.8%,85%).

In the prof1e (aptitude/English) in Appendix D.3, although the maximum LR-value of ensemble systems (27.4%,96.4%) did not reach 100% Reach, it achieved significantly higher LR-values than single systems. Again, performance improvements from single systems were observed through ensemble effects with the combination of 37 (gemma2-9b-it), 40 (mixtral-8x7b-32768), 43 (llama3-8b-8192), and 76 (llama3.1-70b-versatile), with 37 and 43 proving effective as weak learners with Lift values<0 as understood from TABLE 7.

In the prof2e (purchase/English) case (Appendix D.4), TABLE 10, shows that ensemble systems combining multiple systems achieved significantly higher LR-values than single systems, with a maximum of (78.2%,134.8%). Specifically, notable performance improvements were observed by combining systems 37 (gemma2-9b-it), 40 (mixtral-8x7b-32768), 43 (llama3-8b-8192), 46 (llama3-70b-8192), 66 (gpt-4o-mini-2024-07-18), etc., with various weightings. Systems 40 and 43 functioned as weak learners with Lift values < 0 as seen in TABLE 9.

6. CONCLUSION AND DISCUSSION

From the experimental results obtained in this research, the following conclusions can be drawn: 1. Enhanced Performance through SLM Ensembles: Using ensemble techniques such as EBI makes it possible to combine multiple lightweight SLMs to exceed the performance limitations of individual systems and achieve accuracy comparable to LLMs. This suggests the possibility of constructing high-performance NLP systems in environments with limited computational resources.

2. Utilisation of Weak Learners: Instances were confirmed where incorporating systems showing negative Lift values, i.e., individually low-performing systems, into ensembles improved overall performance. This reevaluates the role of ”weak learners” in ensemble learning and demonstrates the effectiveness of integrating multiple systems with diverse perspectives.

3. Effectiveness of the EBI Method: The EBI method, which applies weighting using collective subjective degrees or direct subjective degrees, was suggested to be more effective at integrating individual system judgments and improving ensemble performance compared to simple averaging or other methods.

4. Versatility of the Method: Performance improvements through SLM ensembles were confirmed across different tasks such as aptitude assessment and consumer analysis, as well as across different languages, including Japanese and English. This suggests that the proposed method is not limited to specific tasks or languages but has broad applicability. This method can be broadly applied to tasks involving identity or profile matching based on textual attributes. Potential application areas include candidate-job matching in HR systems, reranking in personalised recommendation, fraud ring detection in e-commerce, and even psychological or behavioural diagnostics based on written responses.

It should be noted that the method assumes structured, text-based input with interpretable features. Its effectiveness may be limited in settings involving unstructured or non-textual data (e.g., image-based matching) or where profile attributes are not comparable across perspectives.

However, several challenges were also identified in this research. Since the metrics are based on comparisons with human evaluations, multiple datasets are needed to ensure reliability. To find an optimal ensemble, a heuristic approach was adopted, involving repeated trial-and-error using numerous low-performing models. One limitation is the lack of a systematic and efficient method for identifying which weak learners contribute effectively to the ensemble. Although the BI method itself is based on statistical processing and thus avoids the risk of excessive tuning, the ensemble construction process may still be susceptible to overfitting or over-tuning.

Future research should focus on generalising the effectiveness of the EBI method through verification using more diverse datasets, investigating systematic selection criteria for effective weak learners, introducing human evaluation for various natural language data, and developing methods to efficiently optimise ensemble structures without excessive trial-and-error.

Overall, this research addresses themes of significant importance for the widespread adoption and development of AI technologies, such as constructing high-performance AI systems in resource-limited situations and effectively utilising individually low-performing systems. In particular, the combination of SLM ensembles and the EBI method represents a promising direction toward realising more efficient and robust NLP. By overcoming these challenges and expanding the range of applications of these methods, SLM ensembles are expected to develop into more practical technologies in the future.

7. CONFLICT OF INTEREST STATEMENT

The authors declare that the technology presented in this manuscript is related to a patent application currently pending with the Japan Patent Office. The patent application has been submitted by i's Factory Corporation, Ltd., which is the institution the authors are affiliated with.

References

[1] Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial Intelligence vs Clinician Performance in Estimating Probabilities of Diagnoses Before and After Testing. JAMA Netw Open. 2023;6:e2347075.

[2] Chen PH, Jung JJ, Kim Y, Lee M, Sanchez-Bayona R, et al. AI-Assisted Clinical Summary and Treatment Planning for Cancer Care: A Comparative Study of Human vs. AI-Based Approaches. J Clin Oncol. 2024;42:1523-1523.

[3] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330:78-80.

[4] Goh E, Gallo R, Hom J, Strong E, Weng Y, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Netw Open. 2024;7:e2440969.

[5] Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023;330:78-80.

[6] Bian N, Han X, Sun L, Lin H, LuY, He B, et al. ChatGPT Is a Knowledgeable but Inexperienced Solver: An Investigation of Common-Sense Problem in Large Language Models. In: Proceedings of the joint international conference on computational linguistics. language resources and evaluation (LREC-Colling). ELRA and ICCL. 2024:3098-3110.

[7] Si C, Yang D, Hashimoto T. Can Llms Generate Novel Research Ideas? A Large-Scale Human Study With 100+ NLP Researchers. International Conference on Learning Representations. ICLR. 2025.

[8] Cui Z, Li N, Zhou H. Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments With LLMS. 2024. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4940173

[9] Mei Q, Xie Y, Yuan W, Jackson MO. A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. Proc Natl Acad Sci USA. 2024;121:e2313925121.

[10] Tartaraj A, Trebicka B. Optimal Grading Scales for Enhancing Product Evaluation by Infomediaries. Int Jour Res Dev. 2023;10:59-66.

[11] Tholen G. Matching Candidates to Culture: How Assessments of Organisational Fit Shape the Hiring Process. Work Employ Soc. 2024;38:705-722.

[12] Zhai X, Nyaaba M, Ma W. Can Generative AI and ChatGPT Outperform Humans on Cognitive-Demanding Problem-Solving Tasks in Science? Epistemic insight & artificial intelligence. 2024;34:649-670.

[13] Szandała T. ChatGPT vs Human Expertise in the Context of IT Recruitment. Expert Syst Appl. 2025;264:125868.

[14] Qiu J, Guo D, Natalie P, Noelle P, Cheri L, et al. Ensemble of Large Language Models for Curated Labeling and Rating of Free-Text Data. 2025. ArXiv preprint: https://arxiv.org/pdf/2501.08413

[15] Zhang C, Liu L, Wang J, Wang C, Sun X, et al. Prefer: Prompt Ensemble Learning via Feedback-Reflect-Refine. Proceedings of the AAAI conference on artificial intelligence. ACM. 2024;38: 19525-19532.

[16] Jiang D, Ren X. Lin BY. LLM-Blender: Ensembling Large Language Models With Pairwise Ranking and Generative Fusion. In: Proceedings of the 61st annual meeting of the Association for Computational Linguistics. 2023;1:14165-14178.

[17] LuK, Yuan H, Lin R, Lin J, Yuan Z, et al. Routing to the Expert: Efficient Reward Guided Ensemble of Large Language Models. In: Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. ACL Anthology. 2024;1:1964-1974.

[18] Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S. Reflexion: Language Agents With Verbal Reinforcement Learning. Adv Neural Inf Process Syst. 2023;36:8634-8652.

[19] Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. In: Findings of the association for computational linguistics. Association for Computational Linguistics. 2024:3563-3578.

Appendix A. Samples of Input Data

This appendix serves to provide examples of the data used in the identity matching experiments, showing how different descriptions of the same person were given in datasets A and B. Two different identification numbers (id_A and id_B) are assigned to the two types of datasets (A and B), and experiments are conducted to identify the IDs of the same person. Examples include aptitude assessment and behavioural profiling based on purchase data.

Table 1: Sample of identified data for job aptitude assessment. Attribute information: id_B=1, id_A=25, Type=1, Age=30.

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 the 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 ...

Table 2: Sample of identified data for purchase behavioural profiling. Attribute information: id_B=31, id_A=1, Sex=2, Age=60.

Category	Description
Style(B)	She values fermented foods and prefers natural foods and nutrient-rich ingredients.
Persona(B)	She has a strong interest in fermentation and health, especially favouring fermented foods like yoghurt and kimchi. She is a loyal customer who makes regular purchases and is interested in new fermentation trends, ...
Style(A)	Her loyalty is still low and she is exploring. She has a very strong health-consciousness and a high interest in beauty and health.
Persona(A)	She has a strong interest in anti-ageing and maintaining cognitive abilities, prefers natural and additive-free products. She is also sensitive to new beauty trends and is exploring effective beauty foods.

Appendix B. Prompts for BI Method

Below are examples of prompts used when estimating A from B. Type 1 prompt t1 outputs the pairs of ids deemed identical, while Type 2 prompt t2 outputs judgments along with subjective confidence s_{ij} source data. In this example, the number of IDs to match is set to 7. When sending the prompt, individual text data is inserted.

B.1 Type 1 Prompt: t1 (aptitude assessment)

You are an expert profiler skilled in analysing people's personalities and psychology. Based on ##Personnel Evaluation Findings of id_B, infer the profile of the individual and compare it with id_A, which appears to be the same person, from ##Aptitude Assessment Findings. As an expert, select the most likely match according to ##Analysis Approach and ##Execution Method, and output the result in the specified ##Output Format.

##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.

*Analyse the individual's behavioural 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.

B.2 Type 2 Prompt: t2 (aptitude assessment)

You are an excellent profiler. Please find the id_A that seems to represent the same person as inferred from the Personnel evaluation of id_B, by comparing it to the Assessment test of candidate id_As. Then, arrange these candidates in order of likelihood. Additionally, output the certainty level for each match. Consider the following ##Guidelines, ##Detailed requirements, ##Evaluation Method for Certainty Level, ##Output Format, and ##Data Descriptions:

##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, analyse the person's behavioural 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)

##Data Descriptions
Below are the two CSV datasets to be used for the analysis.
* **CSV data describing id_A:** The data includes id_A, Assessment test.
The data is as follows in the {}.
{id_A, Assessment(A) [repeat 7 sample data]}
* **CSV data describing id_B:** The data includes id_B, Personnel evaluation.
The data is as follows in the {}.
{id_B, Personnel evaluation(B) [repeat 7 sample data] }

Based on the above requirements, please output their matching candidate id_A and the certainty levels for all id_B, following ##Output Format, up to the 7 matching candidates. No code is needed.

Appendix C. Sequential Thinking Prompting

Using prof1e (English/aptitude assessment) as an example, the details of sequential prompting are described. The prompt consists of four parts: specifying the profile analysis method (S1), processing sequentially by comparing one person at a time in a step-by-step manner (S2), activating recursive processing when the remaining number reaches a specified value (default=2) (S3), and performing duplication checking and review through S4.

1. System prompt S1
You are an exceptional profiler specialising in identity matching. Execute the judgment step by step. For each ‘id_B‘ in the B file, estimate which ‘id_A‘ in the A file it corresponds to. Follow the {#Analysis Method} described below and make judgments step by step with the highest possible accuracy.

Description of Input Files
‘id_A‘ and ‘id_B‘ are both employee IDs, but the numbers have no correlation.
Each ‘id_A‘ corresponds to exactly one ‘id_B‘.
- **A File**: Contains employee ID (‘id_A‘), age, type, strengths, weaknesses, and aptitude test

results.
- **B File**: Contains employee ID (‘id_B‘), age, type, and HR comments.

Analysis Method
- **Analysis of Each ‘id_B‘**: Carefully analyze behavioral traits and personal characteristics inferred from the age, type, and HR comments to create a detailed profile of the individual.
- **Comparison of Similar Candidates**: Identify ‘id_A‘ candidates with the same age and type as ‘id_B‘ and evaluate which ‘id_A‘ profile most closely matches the characteristics of ‘id_B‘. Justify the conclusion by comparing the information in the A file (age, strengths, weaknesses, aptitude test results).

2. Sequential processing S2
Please evaluate ‘id_B={id_b}‘. Provide your answer in JSON format as follows:
‘{{”thought”: str, ”id_A”: int}}‘.
In ‘thought‘, record the step-by-step comparison and reasoning between ‘id_B‘ and ‘id_A‘. For ‘id_A‘, enter the ID number of the individual who is most similar.
B File
{row_b}
[Conditional Branching: the prompt will be changed as follows depending on k.]
if the counter k= n, ”””
The matching candidates are as follows: ‘id_A={ids_a}‘.
A File
{rows_a} ”””
else”””The target for evaluation is id_A={ids_a}. The information was provided earlier.”””

3. Recursion S3
Please evaluate ‘id_B={ids_b}‘. The evaluation targets will be ‘id_A={ids_a}‘.
If there is any inconsistency or duplication in the evaluation results, you can reselect from the initial candidate set S0=id_A{{{orig_ids_a}}}.
If there are duplicate evaluation results, determine which is more plausible. Additionally, review previous evaluation results, identify any ‘id_B‘ that require corrections in their linkage to ‘id_A‘, and make adjustments if necessary.
Provide your response in a written format. First, describe the step-by-step comparison and reasoning process. Next, document any revisions to the evaluation results.
Finally, output the confirmed pairs of ‘id_B‘ and ‘id_A‘. For the ID pairs, include all previous ‘id_B={old_ids_b}‘. ## B File {rows_b}

4. Conflict resolution S4
Using the output pairs of ‘id_A‘ and ‘id_B‘, check for duplicates or unresolved IDs by following these steps:
Steps
- Within the ‘<thinking>‘ tag, document the review process for the evaluation results. Include considerations such as:
- Ensuring no duplicate ‘id_A‘ is assigned to multiple ‘id_B‘.
- Resolving any unresolved IDs by determining the most plausible pairs.
- Reviewing the overall results to replace any pairs with more plausible ones.
- Within the ‘<result>‘ tag, record the revised pairs of IDs. If no revisions are necessary, document the original pairs as they are.

- Within the ‘<reflection>‘ tag, reflect on whether there are any duplicates or unresolved IDs in the revised results and record the status.
- Within the ‘<count>‘ tag, indicate the number of additional revisions required. If no further revisions are needed, set the value in ‘<count>‘ to 0.

Appendix D. Tables of Results

This appendix lists the results for each system divided by dataset. Table items for single BI systems include the system number, the name of generative AI model, prompt used to generate observation matrix c_{ji} (type and call count), prompt used to generate weight matrix s_{ij} (type and call count), number of correct answers n_c, Lift, and Reach. The asterisk on t1∗ refers to prompts that require a description of the judgment process. The prime on t2′ denotes instances where groq:llama3-70b-8192 was used for generating s_{ij}.

For the Tables of EBI results, the item ”components” lists the set of single BI system numbers comprising the ensemble, and ”weights” shows their respective weights.

D.1 prof1j: Japanese Aptitude Assessment

In this case, the baseline values for (Lift, Reach) are $(H, G) = (19, 22)$. TABLE 3, shows the results of BI systems. BI systems 37 and 42 achieved results equivalent to system 76 (llama3.1-70b-versatile) with a larger size of model. They also surpassed systems 64 and 66 (gpt-4o-mini-2024-07-18). TABLE 4, shows the results of EBI (ensemble BI systems). All of them demonstrate that by combining BI systems (66, 40, 12, 13) with negative Lift, the Lift values turned positive. The Acc of EBI systems 50, 81, and 83 exceeded that of single systems.

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%

Table 4: Results of EBI (ensemble systems) for prof1_j* (Japanese Aptitude Assessment), limited to top-performing systems (Lift≥ 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%

D.2 prof2j: Japanese Purchase Data

In this case, the baseline values for (Lift, Reach) are $(H, G) = (13, 20)$. TABLE 5, shows the results of BI systems, and TABLE 6, shows the results of EBI (ensemble BI systems). They are composed primarily of BI systems (37, 46, 64, 76) that achieved positive Lift in prof1j. Combining systems with negative or zero Lift (40, 42, 43) resulted in positive Lift values.

The top-performing EBI system 82 includes the system 76 of size 70b, but it surpasses the Lift value of system 76 alone. EBI system 89 $\{37, 40, 43\}$, composed of systems without 70b model, achieves accuracy equivalent to system 37 alone while maintaining positive Lift. These results suggest that using multiple SLMs together, rather than relying heavily on systems with large sizes like 70b, can provide diversity and robustness.

Table 5: Results of single BI systems for prof2_j (Japanese Purchase Data).

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
59	cerebras:llama3.1-70b-versatile	t2-10	t2-10	21	61.5%	105%
46	llama3-70b-8192	t1∗-100	t2-10	17	30.8%	85%
64	gpt-4o-mini-2024-07-18	t2-10	t2-10	16	23.1%	80%
65	gpt-4o-mini-2024-07-18	t1∗-100	t1∗-100	16	23.1%	80%
76	cerebras:llama3.1-70b-versatile	t1∗-100	t2-10	16	23.1%	80%
66	gpt-4o-mini-2024-07-18	t1∗-100	t2-10	15	15.4%	75%
37	gemma2-9b-it	t1∗-100	t2′-10	15	15.4%	75%
40	mixtral-8x7b-32768	t1∗-100	t2′-10	13	0.0%	65%
45	llama3-70b-8192	t1∗-100	t1∗-100	12	-7.7%	60%
42	llama3-8b-8192	t1∗-100	t1∗-100	12	-7.7%	60%
43	llama3-8b-8192	t1∗-100	t2′-10	11	-15.4%	55%

Table 6: Results of EBI (ensemble systems) for prof2_j* (Japanese Purchase Data), limited to top-performing systems (Lift≥ 0).

system	components	weights	n_c	Lift	Reach
82	{37,40,43,46,66,76}	[1,1,2,3,1,1]	18	38.5%	90%
83	{37,40,43,46}	[1,1,1,1]	17	30.8%	85%
84	{37,40,43,46,66,76}	[1,1,1,1,1,1]	16	23.1%	80%
85	{37,40,42,46,64,59}	[1,1,1,1,1,1]	16	23.1%	80%
81,91	{37,40,43,46}	[1,1,2,3], [1,1,1,1]	15	15.4%	75%
89	{37,40,43}	[1,1,1]	15	15.4%	75%
90	{37,40,43}	[1,2,1]	14	7.7%	70%
92,93	{37,40,43,46}	[1,2,1,2], [1,2,1,3]	14	7.7%	70%
86	{37,40,59}	[1,1,1]	14	7.7%	70%
74	{37,64}	[1,1]	13	0.0%	65%
77,78	{37,43,45,66,76}	[1,1,1,1,1],[30,3,1,1,10]	12	-7.7%	60%

D.3 prof1e: English Aptitude Assessment

For prof1e, the baseline reference values are $(H, G) = (21.2, 28)$, where $H = G * \gamma$, $\gamma = 0.757$. TABLE 7, shows the results of BI systems. System 40 (mixtral-8x7b-32768) achieved results equivalent to system 66 (gpt-4o-mini-2024-07-18) and surpassed system 76 (llama3.1-70b-versatile).

TABLE 8, shows the results of EBI. System 93 {37, 40, 43, 76} surpassed all single systems, and while its Reach against GPT-4o did not reach 100%, it achieved a notably high 96%. It is interesting that adding BI systems 37, 43, and 76, which had low Reach individually, improved the overall Reach. Additionally, EBI systems 89 and 90 {37, 40, 43} achieved a Reach equivalent to the highest single system while using weak systems 37 and 43 (gemma2-9b-it, llama3-8b-8192) to ensure diversity and serve as sources of robustness.

Table 7: Results of single BI systems for prof1_e (English Aptitude Assessment).

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
66	gpt-4o-mini-2024-07-18	t1∗-100	t2-10	25	17.9%	89.3%
40	mixtral-8x7b-32768	t1∗-100	t2′-10	25	17.9%	89.3%
76	cerebras:llama3.1-70b-versatile	t1∗-100	t2-10	22	3.8%	78.6%
46	llama3-70b-8192	t1∗-100	t2-10	21	-0.9%	75.0%
37	gemma2-9b-it	t1∗-100	t2′-10	20	-5.7%	71.4%
43	llama3-8b-8192	t1∗-100	t2′-10	20	-5.7%	71.4%

Table 8: Results of EBI (ensemble systems) for prof1_e* (English Aptitude Assessment).

system	components	weights	n_c	Lift	Reach
93	{37,40,43,76}	[1,2,1,3]	27	27.4%	96.4%
89,90	{37,40,43}	[1,1,1],[1,2,1]	25	17.9%	89.3%
91	{37,40,43,66}	[1,1,1,1]	24	13.2%	85.7%
83	{37,40,43,46}	[1,1,1,1]	24	13.2%	85.7%
92	{37,40,43,66}	[1,2,1,2]	23	8.5%	82.1%
84	{37,40,43,46,66,76}	[1,1,1,1,1,1]	22	3.8%	78.6%
81	{37,40,43,46}	[1,1,2,3]	22	3.8%	78.6%
71	{37,43,66}	[1,1,1]	21	-0.9%	75.0%
82	{37,40,43,46,66,76}	[1,1,2,3,1,1]	19	-10.4%	67.9%

D.4 prof2e: English Purchase Data

For prof2e, the baseline reference values are $(H, G) = (17.4, 23)$, where $H = G * \gamma$, $\gamma = 0.757$. TABLE 9, shows the results of BI systems. System 37 (gemma2-9b-it) achieved Reach>100%, surpassing GPT-4o. System 66 (gpt-4o-mini) similarly exceeded GPT-4o.

TABLE 10, shows the results of EBI. Systems 95 and 96, which synthesise all six single systems mentioned above, rank as the top 2. Systems excluding system 94 (llama3.3-70b), such as 92 and 93, also achieved Reach>100%, surpassing GPT-4o. These systems include weak systems 40 and 43, which can be considered sources of robustness.

Table 9: Results of single BI systems for prof2_e (English Purchase Data).

system	model	c_{ji}	s_{ij}	n_c	Lift	Reach
94	cerebras:llama3.3-70b	t1∗-100	t2-10	30	72.4%	130.4%
46	llama3-70b-8192	t1∗-100	t2-10	29	66.7%	126.1%
37	gemma2-9b-it	t1∗-100	t2′-10	25	43.7%	108.7%
66	gpt-4o-mini-2024-07-18	t1∗-100	t2-10	24	37.9%	104.3%
40	mixtral-8x7b-32768	t1∗-100	t2′-10	17	-2.3%	73.9%
43	llama3-8b-8192	t1∗-100	t2′-10	17	-2.3%	73.9%

Table 10: Results of EBI (ensemble systems) for prof2_e* (English Purchase Data).

system	components	weights	n_c	Lift	Reach
96	{37,40,43,46,66,94}	[1,1,1,1,1,1]	31	78.2%	134.8%
95	{37,40,43,46,66,94}	[1,1,2,3,1,1]	29	66.7%	126.1%
92,93	{37,40,43,66}	[1,2,1,2],[1,2,1,3]	28	60.9%	121.7%
81,83	{37,40,43,46}	[1,1,2,3],[1,1,1,1]	26	49.4%	113.0%
91	{37,40,43,66}	[1,1,1,1]	26	49.4%	113.0%
71	{37,43,66}	[1,1,1]	23	32.2%	100.0%
89,90	{37,40,43}	[1,1,1],[1,2,1]	23	32.2%	100.0%