Supplementary Material

A. Supplementary Material for RQ1: LLM's performance in extracting architecture modules

Tables I–V present the performance metrics of GPT-3.5-Turbo, GPT-4-Turbo, and GPT-4o, Llama3.1:70b, Llama3.2:3b, DeepSeek-V3, respectively, in extracting architecture modules. These tables provide a detailed evaluation, including statistics for $Precision$, $Recall$, and F_1 score—covering their mean, and standard deviation—across 20 runs per project.

TABLE I: Precision, Recall, and F1 score Statistics for Modules Extracted by GPT-3.5-Turbo (PM: Precision Mean, PSD: Precision Standard Deviation, RM: Recall Mean, RSD: Recall Standard Deviation, FM: F1 Score Mean, FSD: F1 Score Standard Deviation)

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	76.17	12.80	83.86	9.76	78.96	8.25
JR	64.79	11.84	86.43	18.90	73.70	13.91
MS	74.18	4.59	89.47	2.29	81.03	3.00
TM	73.37	17.11	77.82	11.51	74.42	12.32
TS	96.25	7.37	99.58	2.64	97.71	4.22
Mean	76.95	10.74	87.43	9.02	81.17	8.34

* All values are in percentages.

TABLE II: Precision, Recall, and F1 score Statistics for Modules Extracted by GPT-4-Turbo

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	84.72	5.51	84.72	5.51	84.72	5.51
JR	62.31	11.64	75.45	14.25	67.58	10.98
MS	66.12	9.51	90.00	17.00	75.93	11.75
TM	74.42	1.82	90.00	0.00	81.46	1.12
TS	98.33	5.27	98.33	5.27	98.33	5.27
Mean	77.18	6.75	87.70	8.41	81.60	6.93

* All values are in percentages.

TABLE III: Precision, Recall, and F1 score Statistics for Modules Extracted by GPT-4o

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	72.31	12.70	81.82	12.16	76.00	9.64
JR	76.49	10.30	95.00	11.00	84.59	10.16
MS	76.39	3.88	90.00	3.24	82.59	3.13
TM	84.34	7.88	87.50	4.06	85.60	4.56
TS	98.57	4.40	100.00	0.00	99.23	2.37
Mean	81.62	7.83	90.86	6.09	85.60	5.97

* All values are in percentages.

We designated GPT-3.5-Turbo as the baseline model; it achieved average scores of 76.95% $Precision$, 87.43% $Recall$, and 81.17% F_1 score. GPT-4-Turbo demonstrated modest improvements, with averages of 77.18% $Precision$ (+0.23%), 87.70% $Recall$ (+0.27%), and 81.60% F_1 score (+0.43%) over the baseline. GPT-4o exhibited more substantial enhancements, recording 81.62% $Precision$ (+4.67%), 90.86% $Recall$ (+3.43%), and 85.60% F_1 score (+4.43%). Among the Llama variants, Llama3.1:70b outperformed the baseline GPT-3.5-Turbo, achieving 81.63% $Precision$ (+4.68%), 89.88% $Recall$ (+2.45%), and 84.75% F_1 score (+3.58%). Conversely, Llama3.2:3b underperformed relative to all other models, attaining only 66.32% $Precision$ (-10.63%), 83.80% $Recall$ (-3.63%), and 72.00% F_1 score (-9.17%).

TABLE IV: Precision, Recall, and F1 score Statistics for Modules Extracted by Llama3.1:70b

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	64.66	9.53	91.61	4.40	75.42	6.61
JR	85.37	9.44	99.46	3.29	91.60	5.75
MS	78.75	3.68	87.50	4.42	82.84	3.38
TM	83.54	17.25	75.00	0.00	78.06	9.52
TS	95.83	8.47	95.83	8.47	95.83	8.47
Mean	81.63	9.67	89.88	4.11	84.75	6.74

* All values are in percentages.

TABLE V: Precision, Recall, and F1 score Statistics for Modules Extracted by Llama3.2:3b

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	56.06	16.90	88.77	14.16	66.70	10.33
JR	77.81	10.82	98.00	8.83	86.43	8.97
MS	72.56	6.68	82.50	5.88	77.06	5.11
TM	52.62	29.11	57.14	12.47	49.56	13.49
TS	72.57	17.97	92.59	9.30	80.24	13.30
Mean	66.32	16.30	83.80	10.13	72.00	10.24

* All values are in percentages.

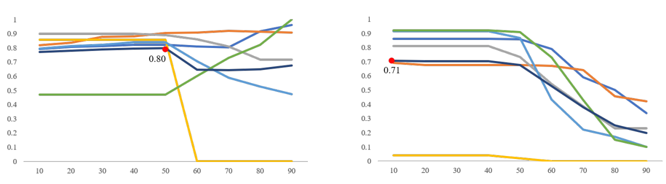
Most models exhibited consistent performance across the projects. Specifically, the baseline GPT-3.5-Turbo had standard deviations of ±10.74% $Precision$, ±9.02% $Recall$, and ±8.34% F_1 score. GPT-4-Turbo showed reduced variability with standard deviations of ±6.75% $Precision$, ±8.41% $Recall$, and ±6.93% F_1 score. GPT-4o maintained low variability, with standard deviations of ±7.83% $Precision$, ±6.09% $Recall$, and ±5.97% F_1 score. Llama3.1:70b also demonstrated relatively low variability, exhibiting standard deviations of ±9.67% $Precision$, ±4.11% $Recall$, and ±6.74% F_1 score. However, Llama3.2:3b demonstrated notably higher variability in its results, with standard deviations of ±16.30% $Precision$, ±10.13% $Recall$, and ±10.24% F_1 score, likely due to its smaller model size. This suggests less stable performance compared to other LLMs.

The results indicate that GPT-4o significantly outperforms the baseline GPT-3.5-Turbo in extracting architecture modules. GPT-4o achieves higher $Precision$, $Recall$ and F_1 score, highlighting its enhanced ability to accurately identify documented modules essential for comprehensive architecture analysis. GPT-4-Turbo, while showing only slight improvements over the baseline, still performs better than GPT-3.5-Turbo in all metrics. In terms of the Llama models, Llama3.1:70b performs better than the baseline GPT-3.5-Turbo, which has also been demonstrated the better performance on reasoning, tool use, math, and coding tasks [1]. In contrast, Llama3.2:3b shows the lowest performance metrics and greater variability, which may be attributed to its smaller model size and limited capacity, leading to less effective results compared to other LLMs.

Evaluating LLMs for extracting documented modules shows that GPT-4o outperforms all tested models, with an average of 81.6% in $Precision$, 90.9% in $Recall$, and 85.60% in F_1 score, along with low variability. Among the Llama models, Llama3.1:70b ranks highest, closely matching GPT-4o in effectiveness. These results highlight GPT-4o and Llama3.1:70b as the most reliable and effective choices for module extraction.

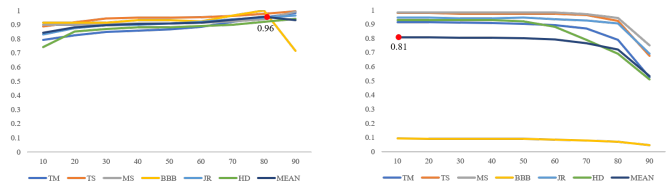
B. Supplementary Material for RQ2: Precison and Recall of BM25 and LLM-S on DM mapping

Figure 1 and Figure 2 respectively present the changes in $Precision$ and $Recall$ for these two methods under a range of similarity thresholds (10 to 90 for BM25-based retrieval and 10% to 90% for LLM-based scoring). As the threshold increases, $Precision$ tends to decrease while $Recall$ increases, with the F_1 score reaching its optimal value at a threshold of 50. This is because increasing the threshold causes more mappings with lower similarity scores to be filtered out and thus treated as UM mappings. As a result, some files that should be mapped to documented modules are mistakenly excluded, leading to a decrease in $Precision$. Meanwhile, since more mappings are classified as UM due to the higher threshold, the likelihood of retrieving true UM mappings increases, which improves $Recall$.



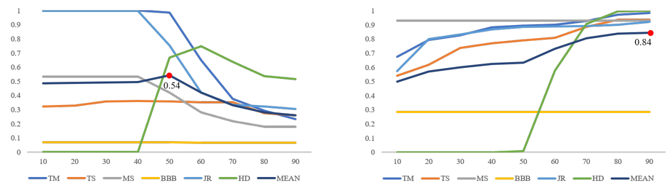
(a) Precision (b) Recall

Fig. 1: Performance of BM25 in generating DM mappings



(a) Precision (b) Recall

Fig. 2: Performance of LLM-S in generating DM mappings



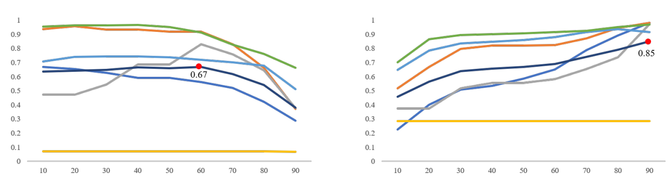
(a) Precision (b) Recall

Fig. 3: Performance of BM25 in generating UM mappings

C. Supplementary Material for RQ3: Precison and Recall of BM25 and LLM-S on UM mapping

REFERENCES

[1] (2024) Gpt-3.5-turbo-vs-llama-3.1:70b. [Online]. Available: https://www.vellum.ai/comparison/gpt-3-5-vs-llama-3-1-70b



(a) Precision (b) Recall

Fig. 4: Performance of LLM-S in generating UM mappings