Supplementary Material

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

Tables I–VI 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 architectural 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 F_1 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 F_1 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 F_1 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%). DeepSeek-V3 achieved 74.47% $Precision$ (–2.48%), 93.14% $Recall$ (+5.71%), and 82.26% F_1 score (+1.09%) relative to

TABLE IV: Precision, Recall, and F_1 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 F_1 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.

TABLE VI: Precision, Recall, and F_1 score Statistics for Modules Extracted by Deepseek-V3

Project	PM*	PSD*	RM*	RSD*	FM*	FSD*
BBB	74.41	9.18	88.18	8.62	80.61	8.52
JR	68.75	4.31	100.00	0.00	81.41	3.10
MS	75.00	0.00	90.00	0.00	81.82	0.00
TM	62.10	10.56	87.50	0.00	72.17	7.47
TS	92.07	14.14	100.00	0.00	95.30	8.61
Mean	74.47	7.64	93.14	0.00	82.26	5.54

* All values are in percentages.

the baseline GPT-3.5-Turbo.

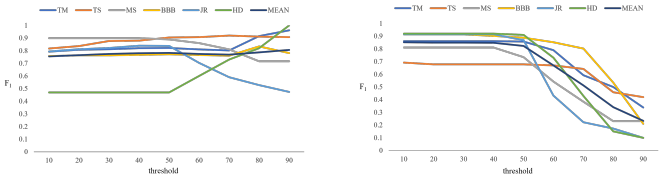
Most models exhibited consistent performance across the projects. Specifically, the baseline GPT-3.5-Turbo had standard deviations of ±10.74% *Precision*, ±9.02% *Recall*, and ±8.34% F_1 score. GPT-4-Turbo showed reduced variability with standard deviations of ±6.75% *Precision*, ±8.41% *Recall*, and ±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. DeepSeek-V3 records moderate *Precision* variability (±7.64%) and an anomalously zero *Recall* standard deviation, and exhibits the smallest F_1 spread (±5.54%), indicating reasonably consistent effectiveness.

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. 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. Finally, DeepSeek-V3 attains the highest *Recall* among all evaluated models and registers an F_1 score slightly above the baseline, accompanied by a modest reduction in *Precision*.

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

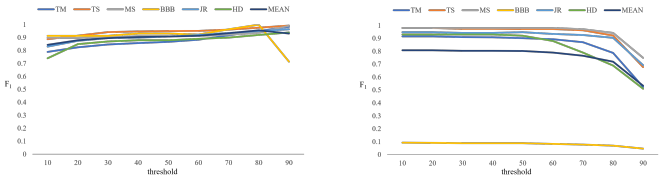
Figure 1 and Figure 2 plot the variation of Precision and Recall in DM mapping for BM25 and LLM-S as the similarity threshold increases (10–90 for BM25-based retrieval, 10%–90% for LLM-based scoring).

As the threshold increases, *Precision* increases whereas *Recall* decreases. This trend occurs because a higher threshold accepts only those mappings that show stronger semantic similarity to the candidate documented module (DM), thereby improving classification reliability and raising *Precision*, while the stricter criterion simultaneously discards numerous true DMs and thus lowers *Recall*. Across the entire threshold range, LLM-S exhibits lower variance than its BM25 counterpart and consistently achieves higher *Precision* and *Recall*, indicating that LLM-based semantic scoring provides a more robust foundation for DM mapping than traditional lexical similarity.



(a) Precision (b) Recall

Fig. 1: Performance of BM25 in generating DM mappings

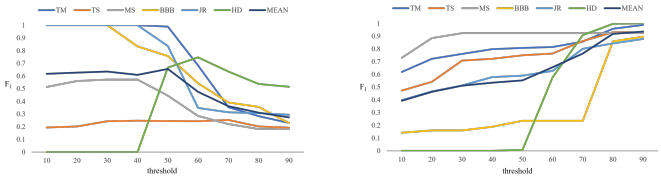


(a) Precision (b) Recall

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

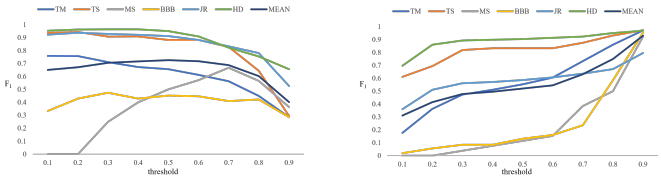
C. Supplementary Material for RQ3: Precison and Recall of BM25 and LLM-S on UM mapping

Figure 3 and Figure 4 show the *Precision* and *Recall* of BM25 and LLM-S in generating UM mappings as the similarity threshold increases. As the threshold increases, *Precision* falls and *Recall* rises. This is because increasing the threshold causes more mappings with high similarity to be classified 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. 3: Performance of BM25 in generating UM mappings



(a) Precision (b) Recall

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