MicroDec: Leveraging Graph-based Random Walks and Large Language Models for Microservice Decomposition

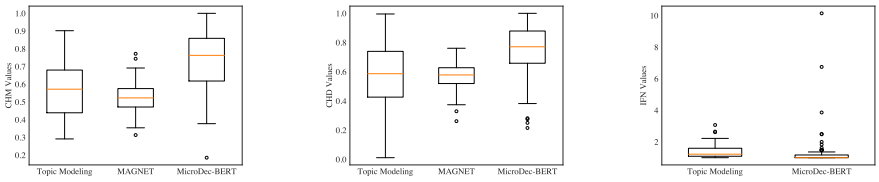
March 11, 2025

1 Q1- How does our approach enhance functional independence and modularity compared to baselines?

We further investigate the results to confirm ***MicroDec***-BERT's performance. Figure 1 shows the range of *CHM, CHD,* and *IFN* metrics for service candidates identified by ***MicroDec***-BERT, Topic modeling, and MAGNET across all applications.

In Figures 1a and 1b, ***MicroDec***-BERT achieves a higher median close to 0.8 for both *CHM* and *CHD* compared to other methods with median close to 0.6. For *IFN* as shown in Figure 1c, ***MicroDec***-BERT has outliers that extend up to a value of 10 for an application with 886 classes and 71 interfaces, while Topic Modeling shows fewer and less extreme outliers, with maximum value near 4. For the same application, Topic Modeling records 2.6 for *IFN*. However, this improvement comes at the cost of *CHM* and *CHD*, where ***MicroDec***-BERT shows improvements over Topic Modeling by 12.5% and 44.8%, respectively for this application. While we select the best results, our method still does not preform as well as expected in terms of *IFN* for this specific application. This observation supports the slight improvement of ***MicroDec***-BERT gain compared to Topic Modeling method. Despite these outliers, the overall distribution and the statistical result suggests that ***MicroDec***-BERT achieves better *IFN* with a lower median than Topic Modeling method.

Figure 2 conforms the performance of ***MicroDec***-BERT in achieving higher modularity. We can see in Figure 2a that ***MicroDec***-BERT achieves a median of about 0.2, which is notably higher than MAGNET and very close to Topic Modeling just below 0.2. This observation aligns with the results of our statistical test, which shows a significant difference $P < 0.05$ when compared to Topic modeling. In Figure 2b, ***MicroDec***-BERT shows three lower outliers around 0.2, with median around 0.8, which is higher than both Topic Modeling and MAGNET with median of 0.4 and close to 0.0, respectively.



(a) *CHM* (b) *CHD* (c) *IFN*

Figure 1: The range of *CHM, CHD,* and *IFN* metrics for service candidates across all 91 applications.

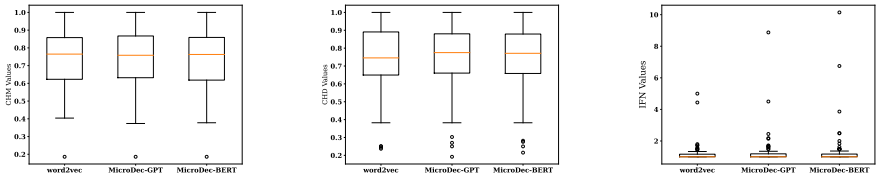
(a) *SMQ* (b) *CMQ*

Figure 2: The range of *SMQ* and *CMQ* metrics for service candidates across all 91 applications.

2 Q2- How does utilizing different LLMs affect the performance of identified service candidates?

The above results are further supported in Figure 3. In Figure 3a and 3b, the similar median values of *CHM* and *CHD*, at 0.75 and nearly 0.80 respectively for both models, show their equal performance at the message and domain levels of functional dependence. For *IFN*, Figure 3c shows ***MicroDec***-GPT has fewer extreme outliers, with maximum value near to 9 while the ***MicroDec***-BERT extend up to a value of 10 with more outliers than ***MicroDec***-GPT.

For *SMQ* and *CMQ* as shown in Figures 4a and 4b, the differences between two models become clearer. ***MicroDec***-BERT has a slight higher median than ***MicroDec***-GPT, while ***MicroDec***-GPT outperforms in terms of *CMQ*.



(a) *CHM* (b) *CHD* (c) *IFN*

Figure 3: The range of *CHM, CHD,* and *IFN* metrics for service candidates identified by word2vec, ***MicroDec***-GPT and ***MicroDec***-BERT across all 91 applications.

(a) *SMQ* (b) *CMQ*

Figure 4: The range of *SMQ* and *CMQ* metrics for service candidates identified by word2vec, ***MicroDec***-GPT and ***MicroDec***-BERT across all 91 applications.