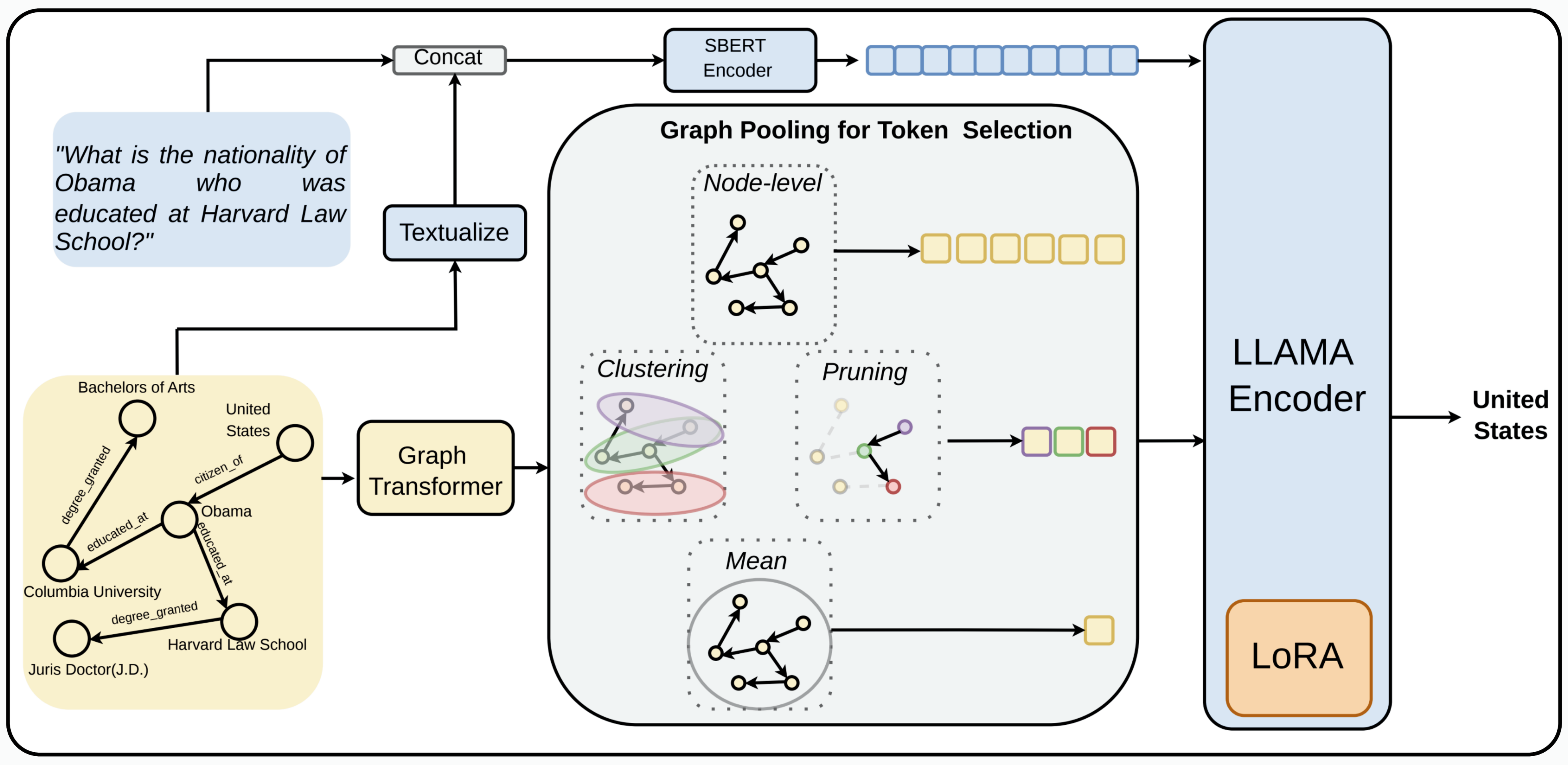
Is One Token All It Takes?

Graph Pooling Tokens for LLM-based Graph Question Answering

Ankit Grover[1], Lodovico Giaretta[2], Rémi Bourgerie[1], Sarunas Girdzijauskas[1,2]
[1]KTH Royal Institute of Technology, [2]RISE Research Institutes of Sweden AB

MOTIVATION

Graph-to-LLM methods pick an extreme. Feed **every node** (doesn't scale), or mean-pool into **one token** (loses topology, as in *G-Retriever*). Can **smarter compression recover** what mean-pooling loses? *Is one token all it takes?*



GRAPH POOLING TOKENISATION METHODS

- **Pruning:** TopK, Self-Attention Graph Pooling
- **Clustering:** DiffPool, MinCutPool
- **Virtual Nodes for Pooling** (VNPool)

Minimise both *Cross-Entropy Loss* for the LLM and *Pooling Loss* for dense graph pooling methods.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{pool}}$$

Pooling Loss is added only for DiffPool, MinCutPool.

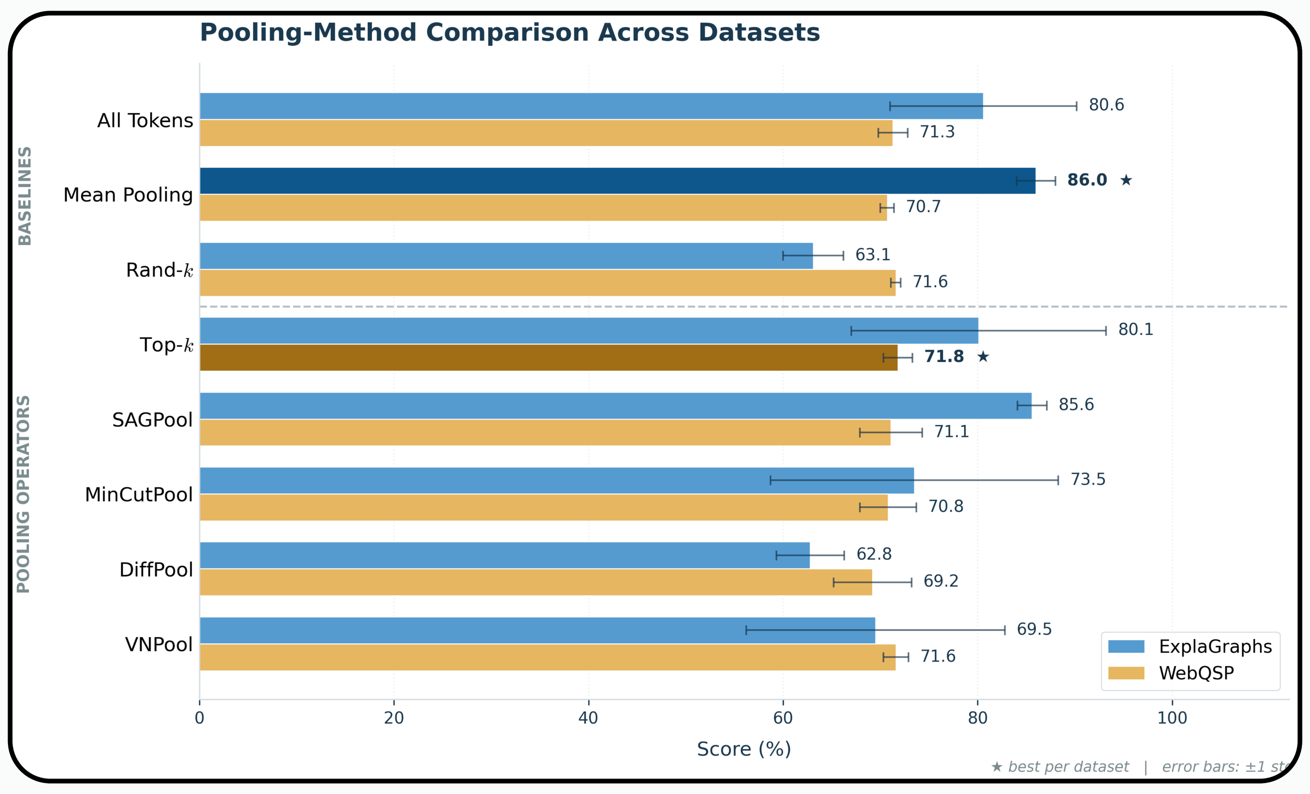
RQ1: Can we **replace textual graph tokens using graph pooling tokens** alone when training via Soft Prompt Tuning?

Method	ExplaGraphs	WebQSP
All Tokens	**88.3%**	56.7%
All Tokens + Textualization	87.3%	**71.4%**

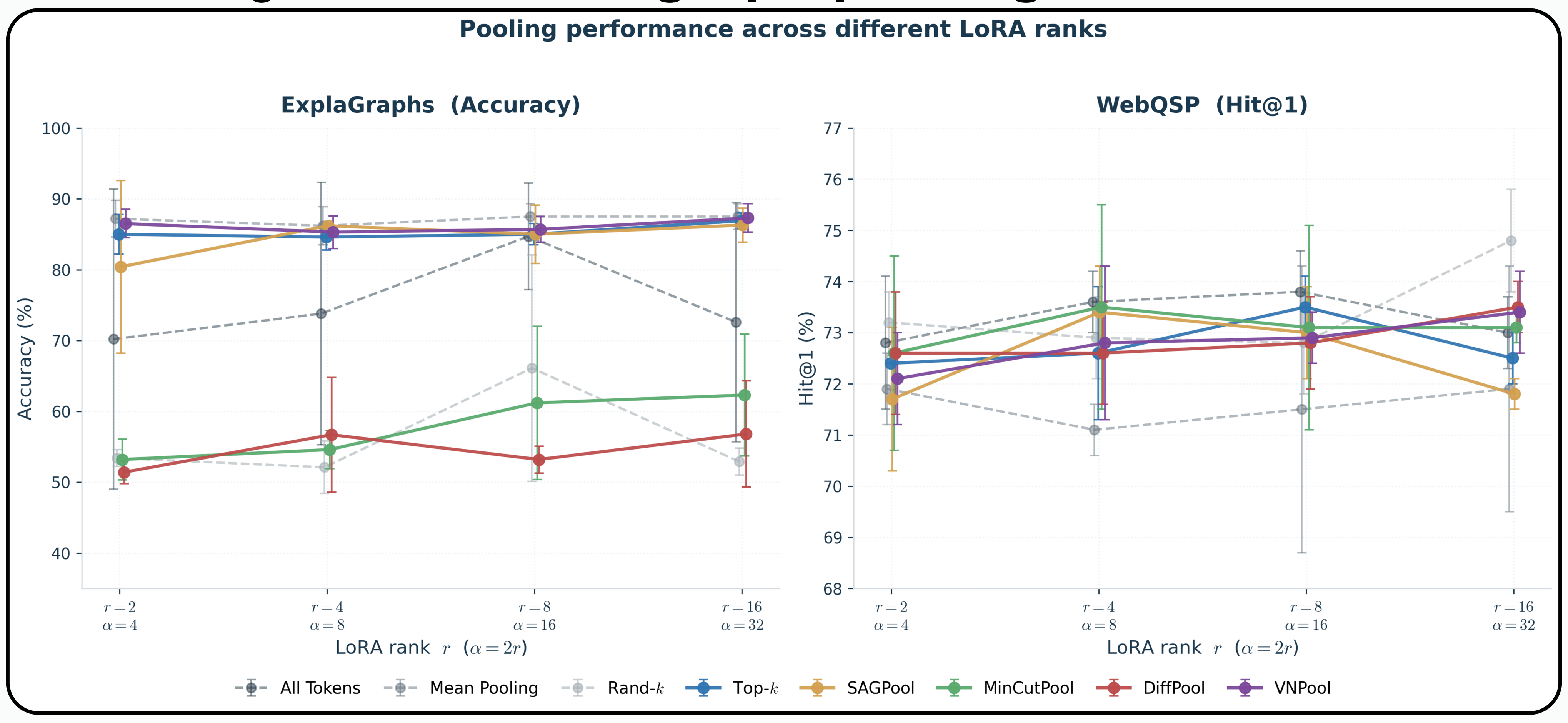
We use **k=8** graph tokens, **TransformerConv** as Graph Transformer and concatenating with **textualised graph** tokens via **PCST retriever** input to **LLAMA 2-7B.**

No, we require *both textualised subgraph tokens and graph tokens* for more stable Soft Prompt Tuning.

RQ2: Can **Low Rank Adapters***(***LoRA***)* fine-tuning **stabilise the graph pooling tokens**?

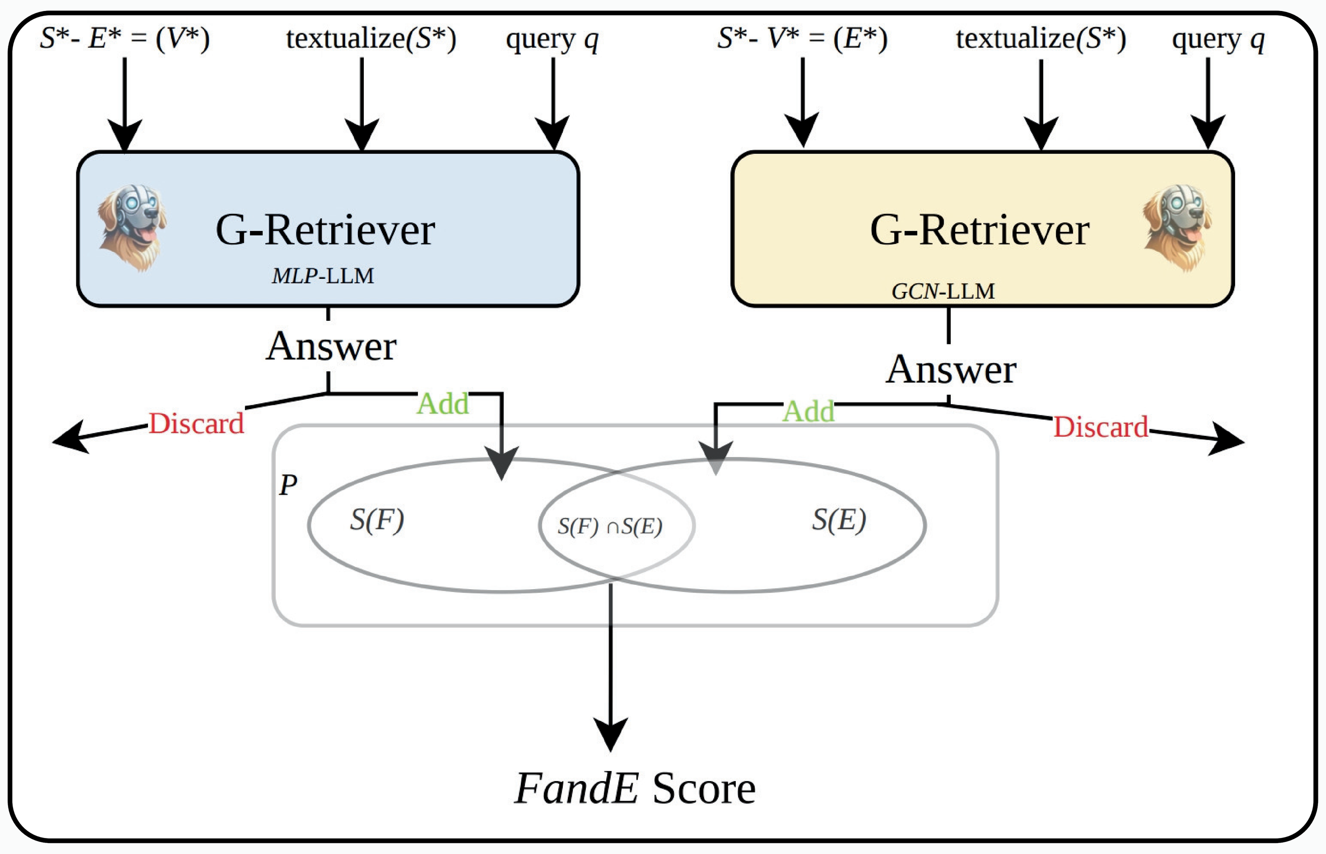


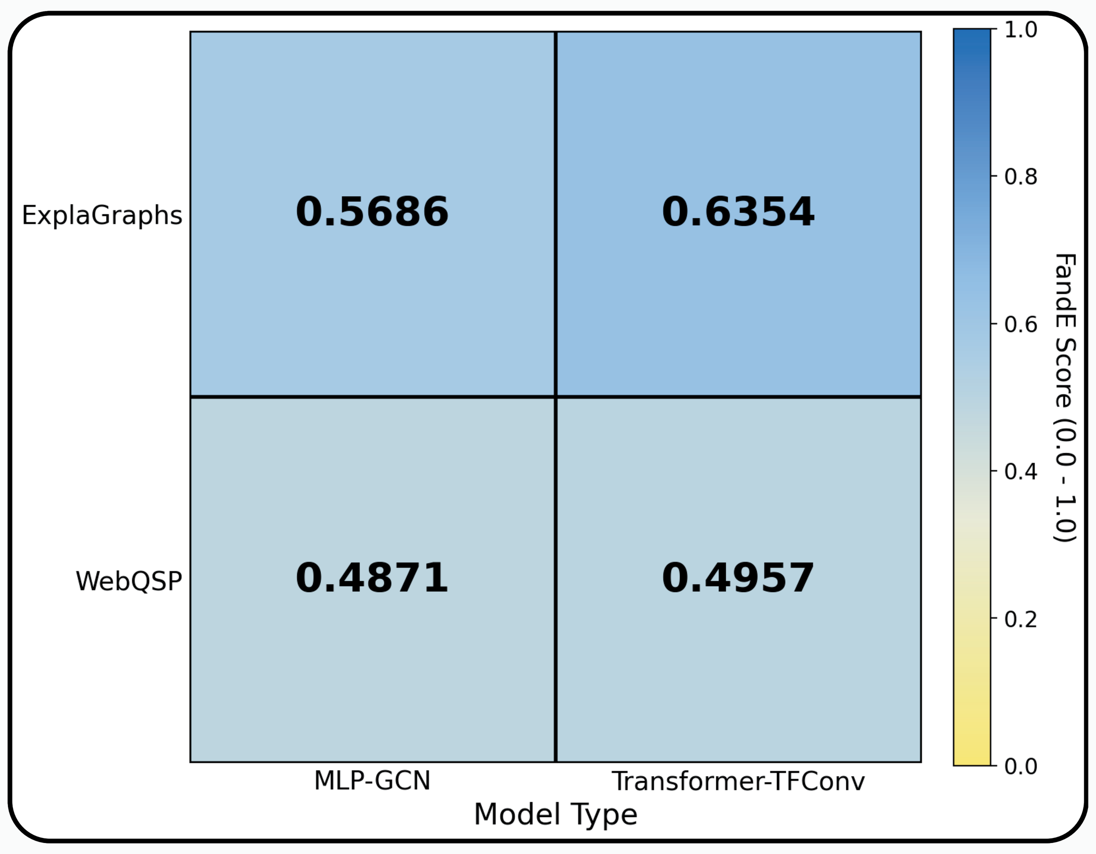
LLMs with graph pooling tokens trained via **Soft Prompt Tuning(PT)** remain **unstable**.



Inserting LoRA adapters into the **q_proj** and **v_proj** layers reveals that graph pooling tokens stabilise pruning-based methods such as TopK pooling and SAGPooling. Compared to soft PT it **improves VNPool**, but clustering-based approaches, such as **DiffPool** and **MinCutPool**, **remain unstable.**

RQ3: Can we **quantify redundant signals** from **node features or edges** alone in our datasets?





Calculate **FandE Score** as the normalised intersection of *solvable sets of predictions* of *Feature-only model* **S(F)** and *Edge-only model* **S(E),** keeping LLM frozen.

On ExplaGraphs, **315** examples are solvable by both MLP and GCN, showing **higher FandE Scores** *across different encoders.*

READ FULL PAPER
:)

CODE
;)

REFERENCES

1. He, X. et al. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and QA. NeurIPS 2024.
2. Kim, W., Park, B., Kim, W. Query-aware Learnable Graph Pooling Tokens as Prompt for LLMs. 2025. arXiv:2501.17549.
3. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.
5. Faber, L., Lu, Y., Wattenhofer, R. Should GNNs Use Features, Edges, or Both? 2021.
6. Ying, Z. et al. Hierarchical Graph Representation Learning with Differentiable Pooling. NeurIPS 2018.
7. Lee, J., Lee, I., Kang, J. Self-Attention Graph Pooling. ICML 2019.
8. Gao, H., Ji, S. Graph U-Nets. ICML 2019.
9. Bianchi, F. M., Grattarola, D., Alippi, C. Spectral Clustering with GNNs for Graph Pooling. ICML 2020.
10. Chen, R. et al. LLaGA: Large Language and Graph Assistant. ICML 2024.
11. Fatemi, B., Halcrow, J., Perozzi, B. Talk Like a Graph: Encoding Graphs for Large Language Models. ICLR 2024.
13. Southern, J., Di Giovanni, F., Bronstein, M., Lutzeyer, J. F. Understanding Virtual Nodes: Oversquashing & Node
Heterogeneity. 2024. arXiv:2405.13526.
14. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023.