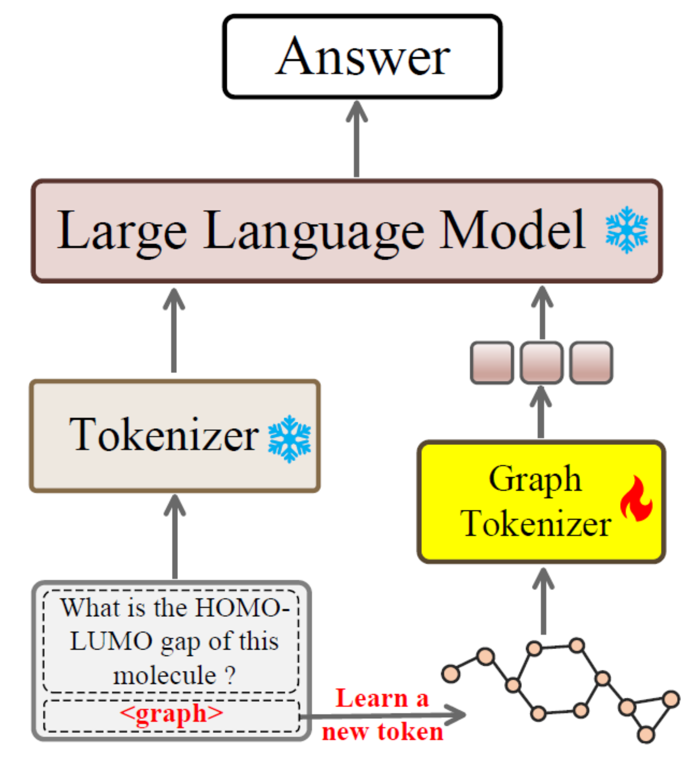
# Graph2Token: Make LLMs Understand Molecule Graphs

Runze Wang[1], Mingqi Yang[2], Yanming Shen[1]
[1]Dalian University of Technology [2]National University of Singapore

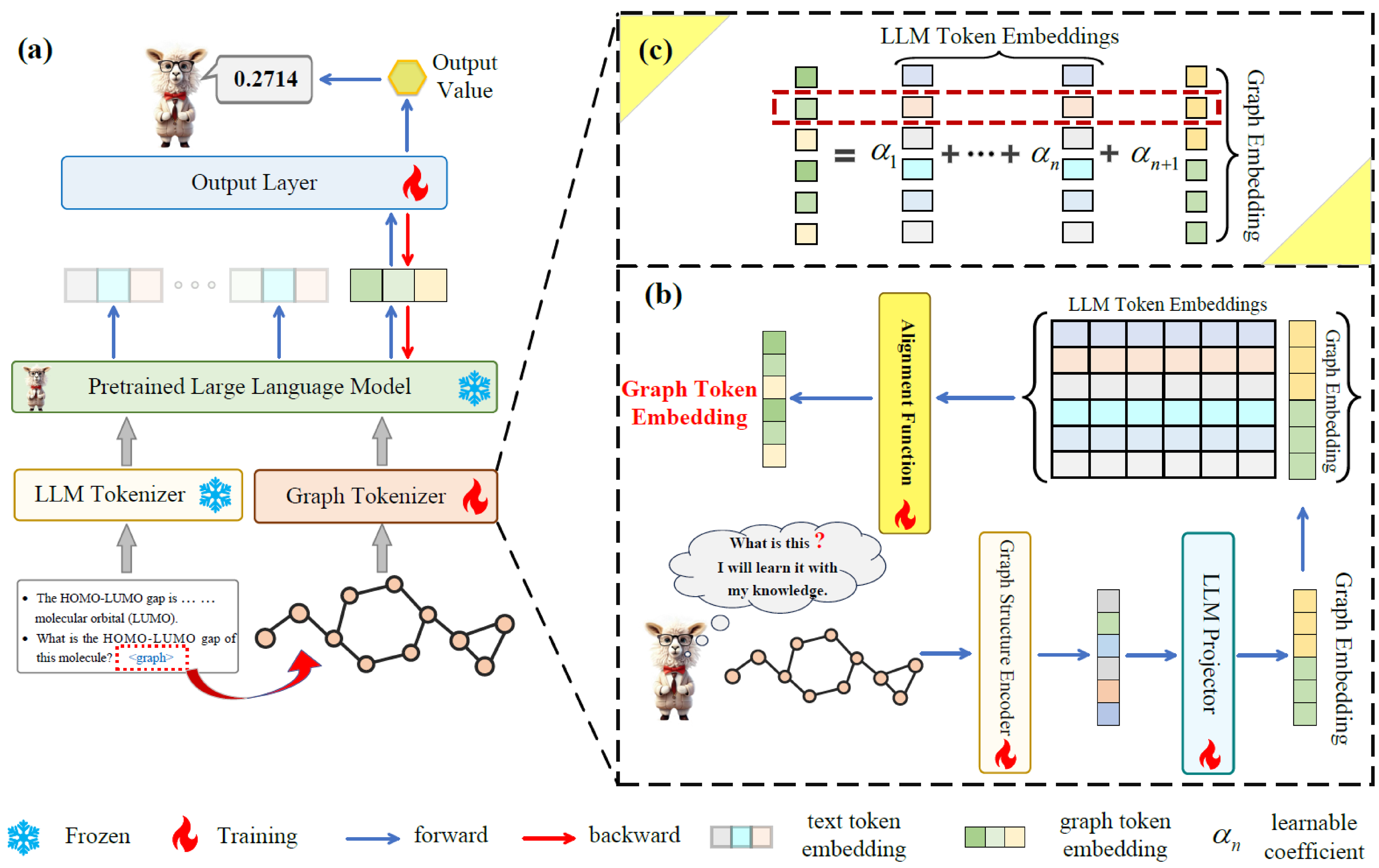
## Motivation

- Large language models (LLMs) excel at various text-related tasks.
- It is still challenging for LLMs to process molecular graph data.
- Textualized graph as input ⟶ insufficient graph reasoning
  Instruction fine-tuning ⟶ change the semantics of LLM backbone
- **Graph2Token**, an efficient solution that aligns a graph token to LLM tokens.




Graph2Token

## Aligning a graph token with the LLM token vocabulary




a) The architecture of Graph2Token with frozen LLM tokenizer and trainable graph tokenizer.

b) The trainable graph tokenizer learns the unknown graph token representation using LLM token vocabulary.

c) Alignment function utilizes a learnable combination of tokens pre-trained by LLM to represent the graph tokens.

## Finetune and few-shot learning performance on molecular datasets

| Method Type | Method | BBBP ↑ | BACE ↑ | HIV ↑ | TOX21 ↑ | Avg↑ |
|---|---|---|---|---|---|---|
| *Supervised Learning* | GIN | 67.8 | 76.8 | 76.5 | 73.9 | 73.8 |
| | GT | 68.7 | 77.2 | 74.2 | 75.5 | 73.9 |
| *Graph Pretrain Finetuning* | GraphMVP-C | 72.4 | 81.2 | 77.0 | 74.4 | 76.3 |
| | Mole-BERT | 70.8 | 79.3 | 76.0 | 75.9 | 75.5 |
| | MolFM | 72.9 | 83.9 | 78.8 | 77.2 | 78.2 |
| | SimSGT | 72.3 | 83.6 | 77.7 | 75.7 | 77.3 |
| *LLM-Based Tuning* | Llama-2-7B-chat | 65.6 | 74.8 | 62.3 | - | 67.6 |
| | Vicuna-v1.3-7B | 60.1 | 68.3 | 58.1 | - | 62.6 |
| | MolCA-S | 70.8 | 79.3 | - | 76.0 | 75.4 |
| | MolCA-GS | 70.0 | 79.8 | - | 77.2 | 75.7 |
| | InstructMol-G | 64.0 | **85.9** | 74.0 | - | 74.6 |
| | InstructMol-GS | 70.0 | 82.3 | 68.9 | - | 73.7 |
| | Graph2Token | **73.5** | 85.0 | **79.4** | **79.2** | **79.3** |

- Results (ROC-AUC) of finetune learning on molecular classification tasks on different datasets compared with LLM-based methods and graph learning methods.

| Ratio | Dataset | Graph2Token | GIN | GCN |
|---|---|---|---|---|
| 5% | BBBP | **64.7** | 61.8 | 64.4 |
| | BACE | **73.2** | 64.4 | 65.1 |
| | HIV | **68.5** | 66.2 | 62.7 |
| | TOX21 | **70.6** | 62.6 | 58.4 |
| 10% | BBBP | **69.5** | 66.9 | 67.0 |
| | BACE | **74.6** | 68.1 | 64.6 |
| | HIV | **69.7** | 66.9 | 60.0 |
| | TOX21 | **71.2** | 66.7 | 68.4 |

- Few-shot learning using 5% and 10% training data on different molecular datasets.

GET CODE

GET PAPER