Fusing **Knowledge-Graphs** and **Large Language Models** using **Bi-directional Cross-Attention**

Supervisors:
Prof. Rex Ying, Prof. Dr. Roger Wattenhofer

York von Schlabrendorff – 15-944-333 – 02.02.2024

Motivation

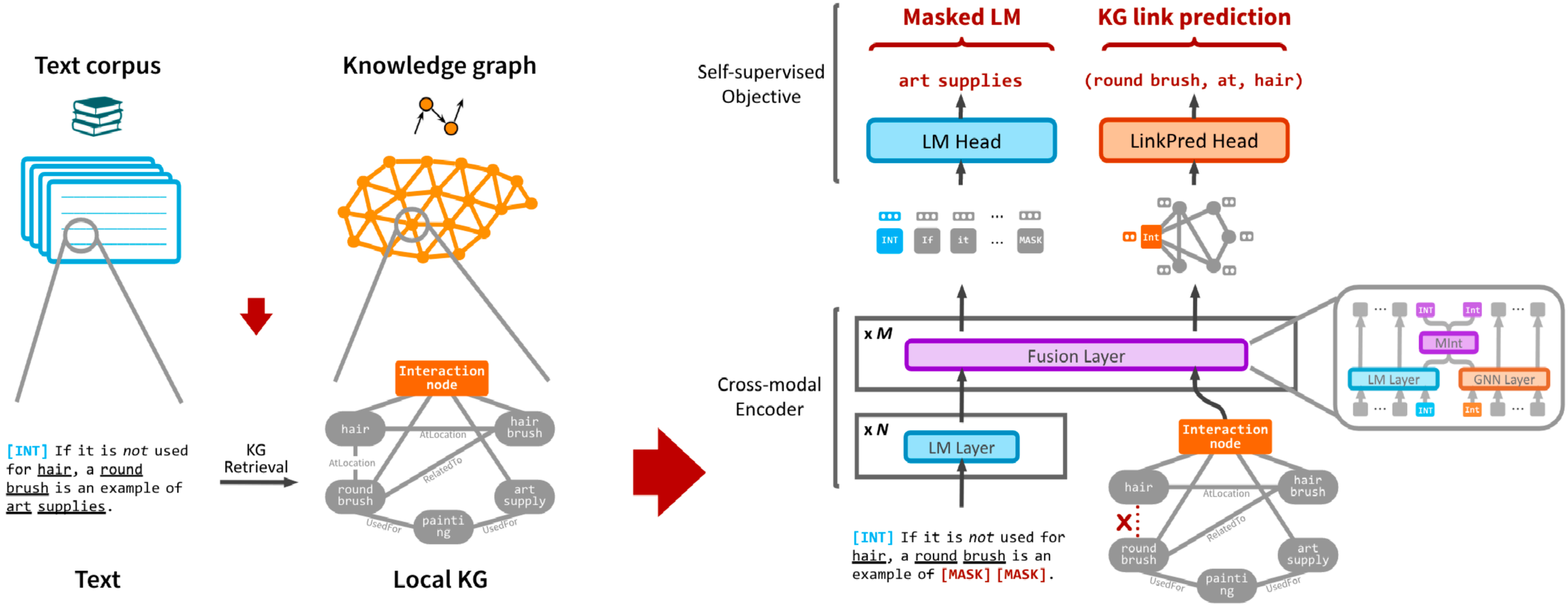
- LLMs are great at reproducing Patterns, but struggle with Facts
- Integrating **Knowledge Graphs** with **LLMs** is a promising approach to improve Fact Reproduction
- We identified potential “improvement possibilities” of previous **KG+LLM** architecture **DRAGON**
    - Dated components: RoBERTa 2019, Graph Attention Networks 2018
    - information exchange bottleneck between the modalities

➡ Goal: improve with **Bidirectional Cross-Attention**

- Multiple-Choice Question Answering as prediction task (MedQA)

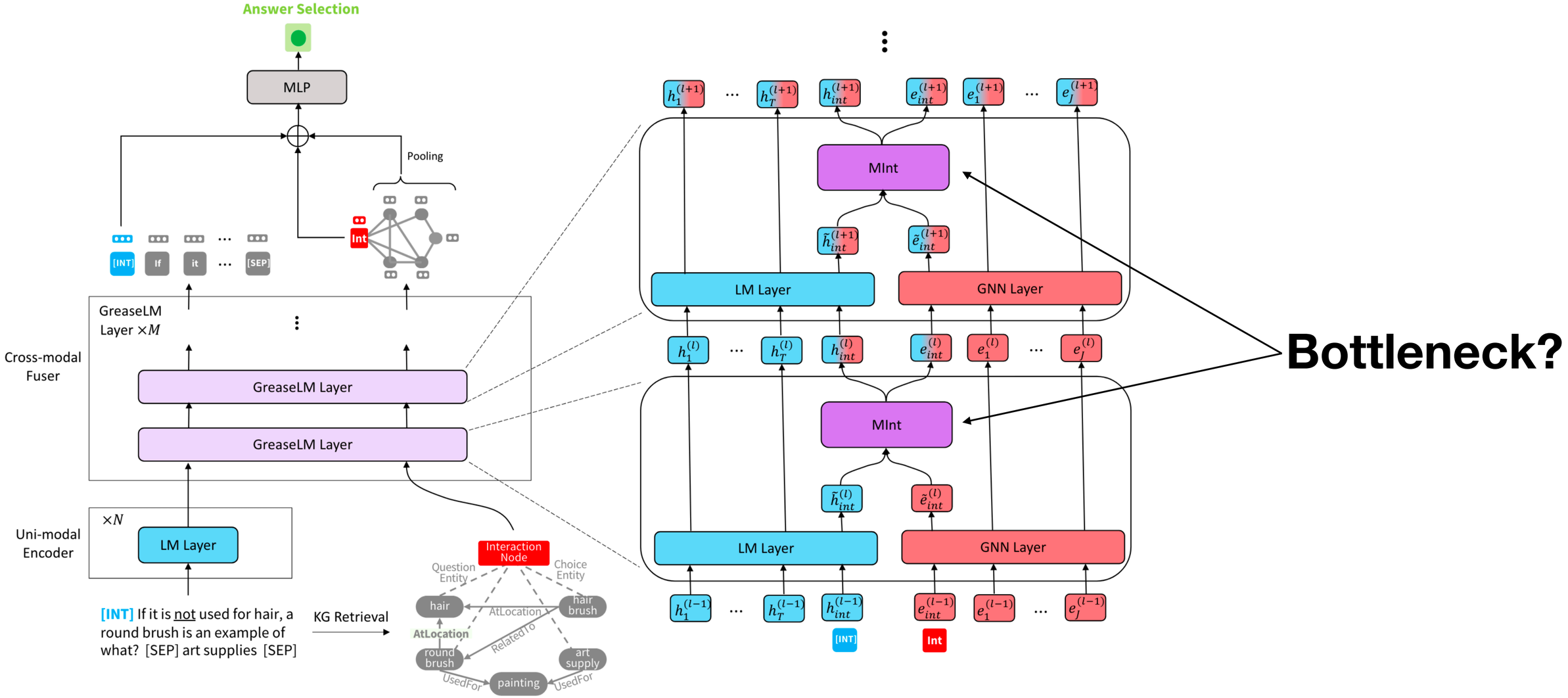
Background

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining



Background

DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining



Background

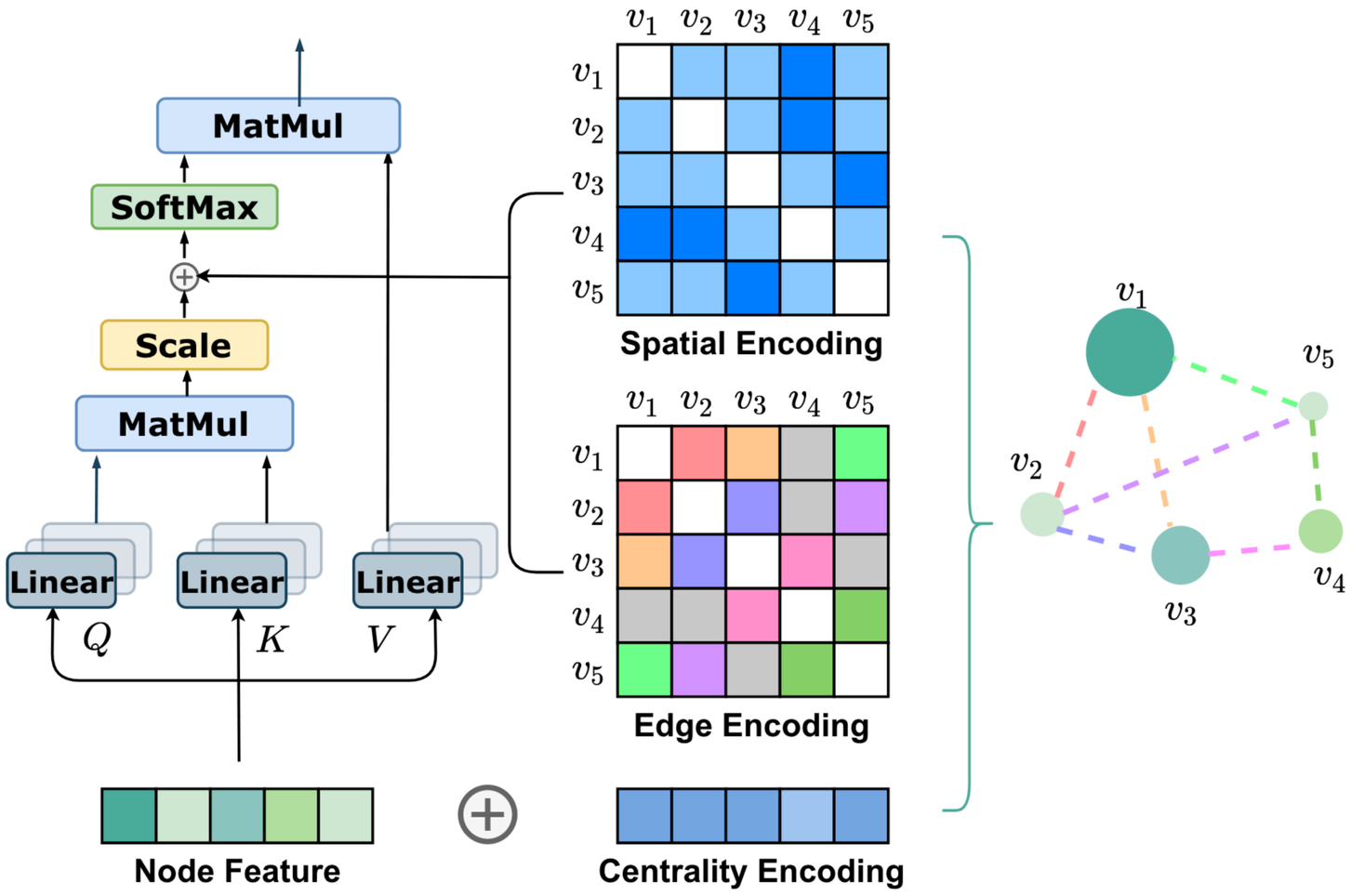
DRAGON: Deep Bidirectional Language-Knowledge Graph Pretraining

Method	MedQA	PubMedQA	BioASQ
BioBERT [74]	36.7	60.2	84.1
PubmedBERT [75]	38.1	55.8	87.5
BioLinkBERT [19]	44.6	72.2	94.8
+ QAGNN	45.0	72.1	95.0
+ GreaseLM	45.1	72.4	94.9
DRAGON (**Ours**)	**47.5**	**73.4**	**96.4**

+2.9pp

Graphormer

- unmodified Attention
- Encodes Nodes as Tokens
- Custom Positional Encoding (Centrality Encoding)
- Encodes Edges and Paths between Nodes as Attention Biases
- O(|V|)

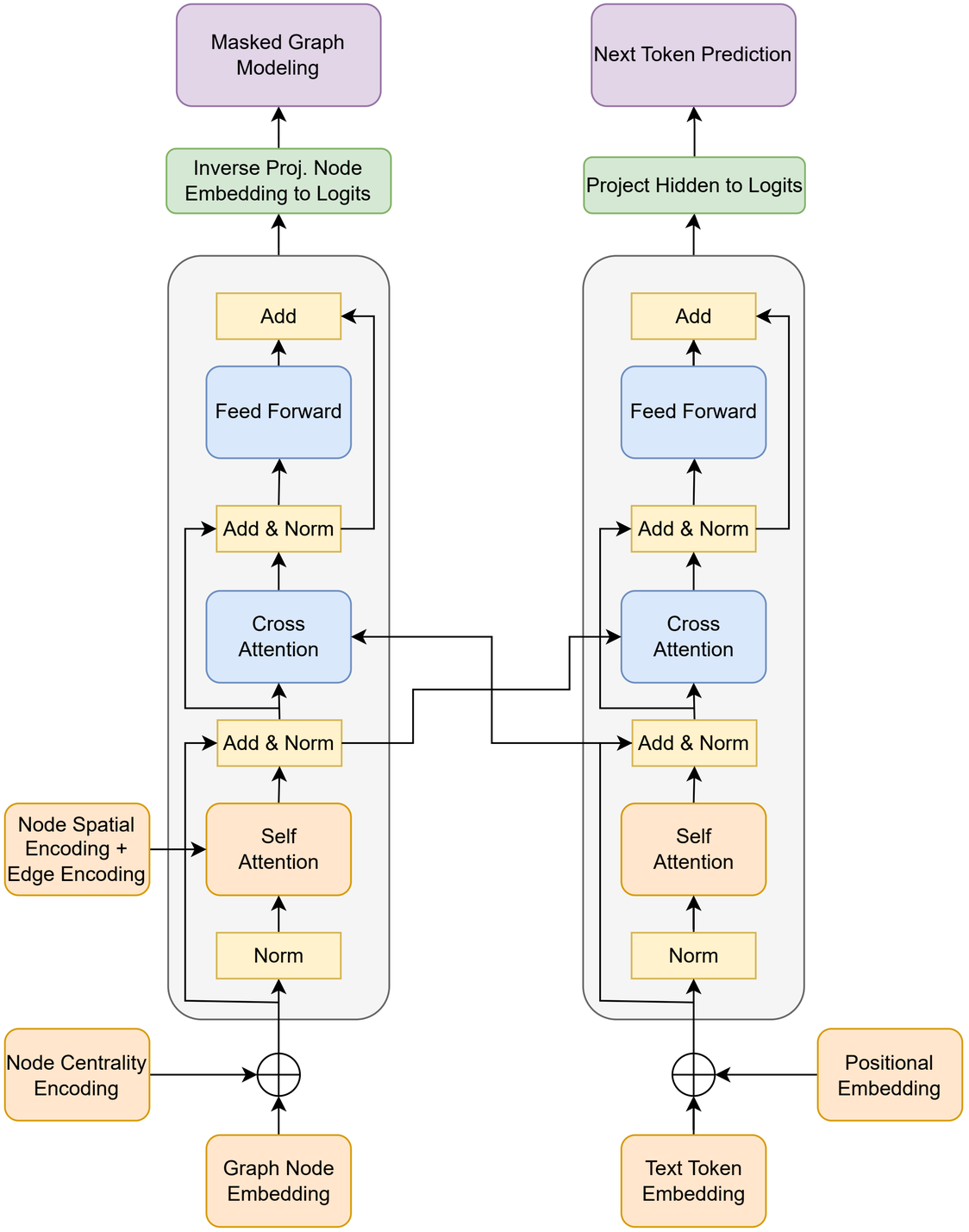


Method

Replace:

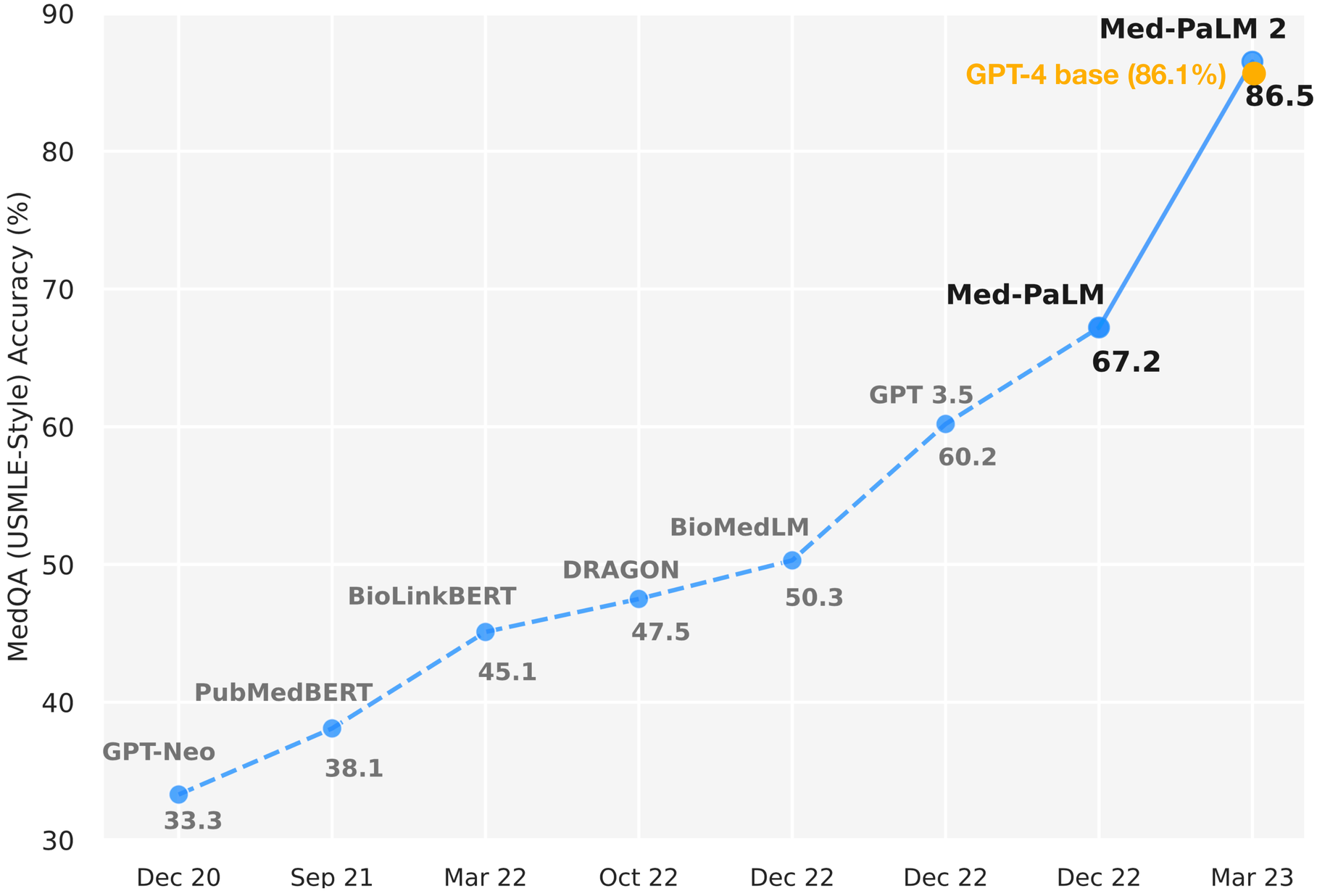
- Dated LM -> Mistral 7B
- Dated GNN -> Graphormer
  - DistMult Link-Prediction -> Masked Graph Modeling
- GreaseLM Modality Interaction -> Bidirectional Cross-Attention

Bidirectional Cross-Attention (pretraining)



MedQA

State of the Art



Baselines

Language Model only

		Accuracy
Frozen LM + Seq Class Head	Without prompt template	~25 %
LoRA-ft LM + Seq Class Head	Without prompt template	~25 %
Frozen LM + Log Prob Eval	With prompt template	41.5 %
Frozen LM + Seq Class Head	With prompt tempalte	41.1%
LoRa-ft LM + Seq Class Head	With prompt Template	**50.7 %**

Prompt Template

Question: *A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7 F (36.5C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?*

A. *Ampicillin*
B. *Ceftriaxone*
C. *Doxycycline*
D. *Nitrofurantoin*

Answer: *D. Nitrofurantoin*

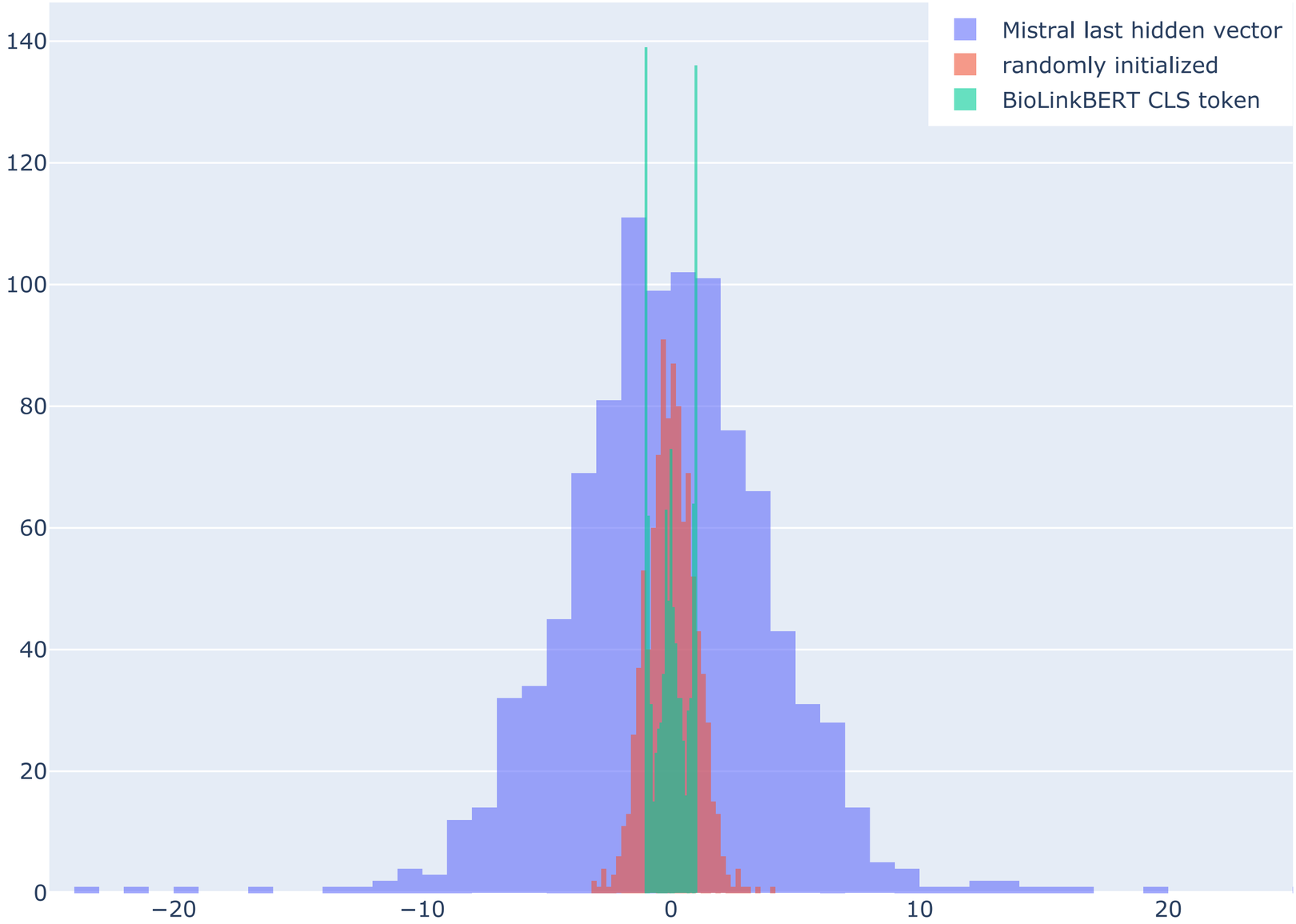
Datasets

- Pretraining
  - PubMed Abstracts (25% subset) + UMLS Metathesaurus
  - ~1T Text Tokens + 280M Nodes
- Finetuning & Evaluation
  - MedQA + UMLS Metathesaurus
  - MedQA: 40’000 examples

Embedding Parameters

- Mistral: 32,000 token vocabulary
  - Byte-Pair Encoding Tokenization enables representation of complex words with multiple tokens
  - ~130M embedding parameters (4096 dimension)
- UMLS: 300,000 concepts
  - No GNN architecture supports “sub*node* tokenization”
  - ~1.2B embedding parameters (4096 dimension)
  - ~300M embedding parameters (1024 dimension)
  - Infeasible to train => use pre-computed embeddings & freeze

Pre-Computed Embeddings



Last-Token from Mistral <30% acc

pre-computed from BioLinkBERT 49.2% acc

Too Many Parameters!

Total: 8.2B – 9.1B params

- 7B from Mistral (mostly frozen, ~2M trainable by LoRA)
- Graphormer:
  - Concept Embedding ~300M-1.2B
  - Node Degree Embedding ~400K
  - Edge Embedding ~30K
  - Dimension Adapter ~3M
- Bidirectional Cross-Attention Layer
  - LLM Self-Attention 25M
  - LLM Cross-Attention 42M
  - Graphormer Self-Attention 6M
  - Graphormer Cross-Attention 1.5M

Embedding Parameters

total number of concepts	297,927 (100%)
MedQA	19 %
PubMed	21 %
MedQA ∪ PubMed	95,093 (32 %)
MedQA ∩ PubMed	25,113 (8 %)

=> 48M parameters with dimension 512

Too Many Parameters!

Graphormer Mini:

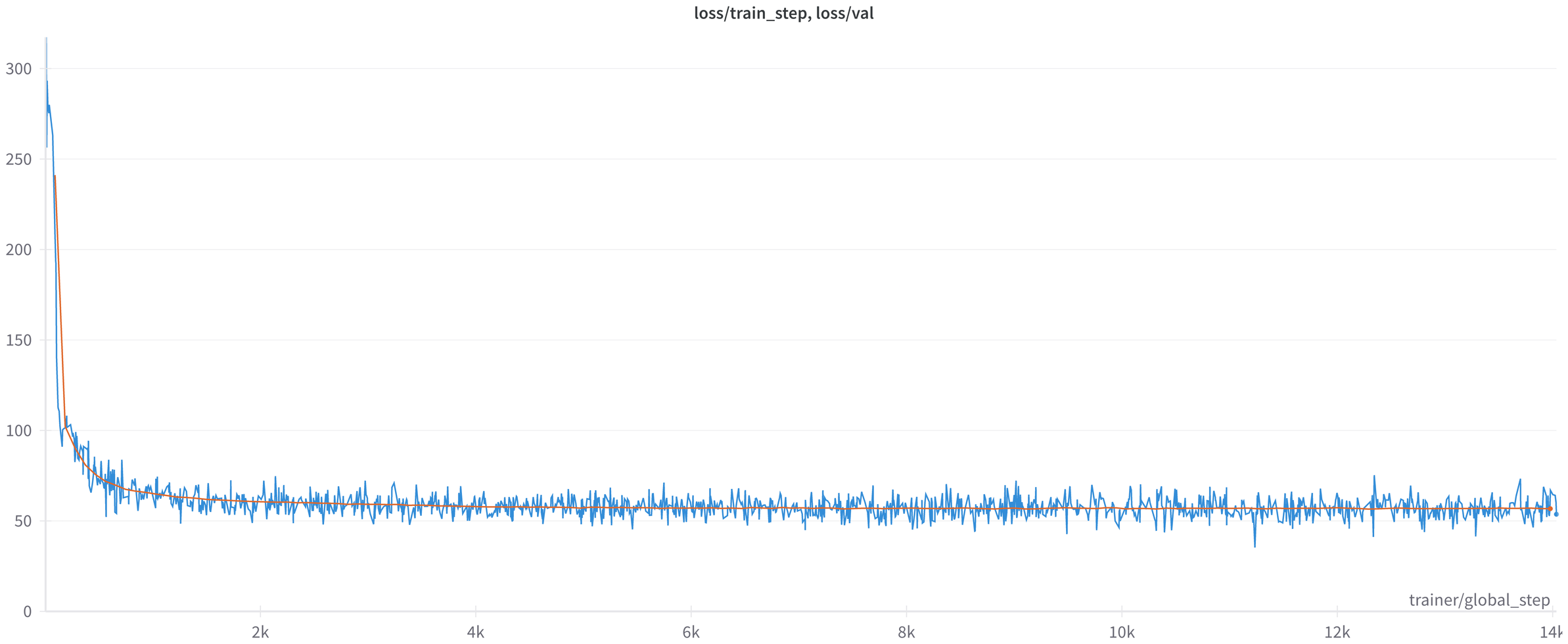
- Layers 12 -> 6
- Embedding Size 756 -> 512
- MLP intermediate size 3072 -> 2048

+ Training the embedding matrix

Total params: 7.5B (180M trainable)

Accuracy 50.9% (beating LM Baseline 50.7%, Dragon 47.5%)

Pretraining Loss



Outlook

- Longer Pre-Training
- Better Graph-Question alignment
  - Combining all possible choice entities in a single graph
  - Is UMLS the best knowledge graph for supplementation?
- Matching hidden Dimensions between Graph and Text modality
  - Closer coupling of Bidirectional Cross-Attention by weight sharing
- BPE for Graph Nodes?
- Alternatives to Graphormer: Ultra? TokenGT?