Embedding-to-Text Decoding with Prefix Adapters

n Sean Gallagher

Abstract — We study whether a frozen sentence embedding can be decoded back into text using a lightweight prefix adapter in front of a frozen decoder LLM. Using E5-base-v2 embeddings and either a two-layer MLP or a linear projection that maps each embedding into K synthetic prefix tokens, we condition a Llama-3.2-1B-Instruct decoder with a fixed prompt and train only the adapter. On MS-MARCO v2.1 queries, the MLP adapter achieves best results at $K = 1$ (BLEU 47.99, F1 0.775), while the linear adapter peaks at $K = 32$ (BLEU 38.74, F1 0.722). These results suggest that nonlinear adapters excel under maximal compression, whereas linear projections benefit from additional prefix capacity.

1. Introduction

Sentence encoders produce compact representations that are routinely used as lossy summaries for retrieval. Embedding-to-text decoding provides a direct probe of how much surface form and content survive in those vectors. This paper summarizes our current pipeline, training objective, and early measurements, with an emphasis on reproducibility and clear separation between completed results and in-progress experiments.

We target three questions:

1. How much text can be reconstructed from a single embedding using a small prefix adapter?
2. How does reconstruction quality vary with the synthetic token budget K and input length?
3. Does partial decoder tuning (e.g., LoRA) improve reconstruction beyond adapter-only training?

2. Method

2.1. Data

We use MS-MARCO v2.1 queries with the standard train, validation, and test splits. The text field is `query`. Inputs are capped at 512 tokens. For datasets that ship with a single split (e.g., OpenWebText), we create deterministic train/validation/test splits by shuffling with a fixed seed and carving out fixed ratios, ensuring evaluations remain reproducible and non-overlapping.

2.2. Model

Our architecture consists of three components:

- **Encoder:** E5-base-v2 sentence embeddings (frozen, max-pool to a single vector).
- **Adapter:** Either a 2-layer MLP (hidden dim 2048, dropout 0.1) or a single linear projection, mapping each embedding into K prefix tokens.
- **Decoder:** Llama-3.2-1B-Instruct (frozen).

Conditioning is achieved by concatenating prefix tokens with prompt embeddings before decoding. The prompt consists of system message "You are a helpful assistant." and user prefix "Reconstruct the original text exactly."

2.3. Training Objective

We minimize masked cross-entropy on the target text tokens. Prompt tokens and padding are masked so the loss reflects reconstruction quality conditioned on the embedding.

2.4. Evaluation

We compute BLEU and token-level F1 on the validation split using greedy decoding. For auditability, we save per-example prompts, predictions, and references to Parquet. Additional diagnostics track average reference length, a token compression ratio (reference tokens divided by K), and length-bucketed performance.

3. Experimental Setup

Dataset: MS-MARCO v2.1 (queries). **Embeddings:** intfloat/e5-base-v2. **Decoder:** meta-llama/Llama-3.2-1B-Instruct. **Prompt:** "Reconstruct the original text exactly." $\boldsymbol{K}$ **sweep:** {1, 2, 4, 8, 16, 32} for both adapter types.

Adapter-only training uses 8 epochs for MS-MARCO and 3 epochs for OpenWebText. Stage B (decoder LoRA tuning) uses the best adapter checkpoint (MLP $K = 1$) with two configurations: adapter frozen vs unfrozen during LoRA training.

4. Results

MLP Adapter				**Linear Adapter**			
K	BLEU	F1	Comp.	K	BLEU	F1	Comp.
1	47.99	.775	7.81	1	27.69	.640	7.83
2	41.65	.734	3.91	2	30.85	.646	3.92
4	31.13	.639	1.95	4	31.52	.684	1.96
8	32.30	.662	0.98	8	38.13	.688	0.98
16	30.04	.628	0.49	16	37.12	.713	0.49
32	30.39	.629	0.24	32	38.74	.722	0.24

Table 1: K sweep results. MLP adapter (left) peaks at $K = 1$; Linear adapter (right) peaks at $K = 32$. Comp. = compression ratio (ref tokens / K).

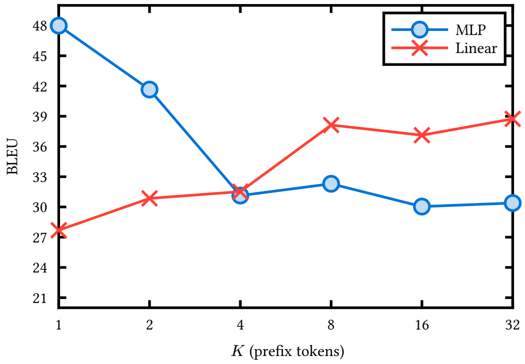

Figure 1: BLEU vs K for MLP and Linear adapters. The MLP adapter performs best at $K = 1$, while the Linear adapter improves with additional prefix capacity.

Table 1 shows the full K sweep results, visualized in Figure 1. Key observations:

- **MLP adapter:** Best performance at $K = 1$ (BLEU 47.99, F1 0.775), with quality degrading as K increases. Performance plateaus around $K = 4$–32 (BLEU 30–32), suggesting the nonlinear projection works best when forced to compress maximally.
- **Linear adapter:** Opposite trend—performance improves with larger K, peaking at $K = 32$ (BLEU 38.74, F1 0.722). The simpler projection benefits from additional capacity.
- **Compression ratio trade-off:** At $K = 1$, each prefix token must encode 7.8 reference tokens on average. The MLP adapter handles this compression better than the linear baseline.

4.1. Qwen3-4B Decoder Variants

Adapter	BLEU	Token F1
MLP (2-layer)	18.54	0.653
Linear	14.92	0.604

Table 2: Qwen3-4B decoder with Qwen3-Embedding-4B inputs (K=1, 500 samples).

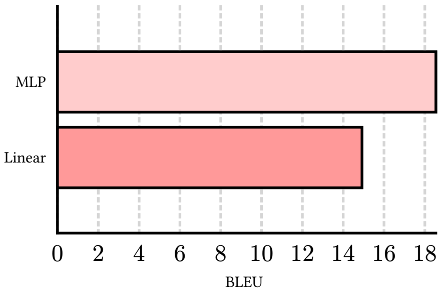

Figure 2: Qwen3-4B BLEU comparison.

4.2. Length-Bucketed Analysis

Reconstruction quality degrades with reference length across all configurations. We analyze this relationship at two scales: MS-MARCO queries (2–20 tokens) and OpenWebText passages (100–17K tokens).

4.2.1. Short Text (MS-MARCO)

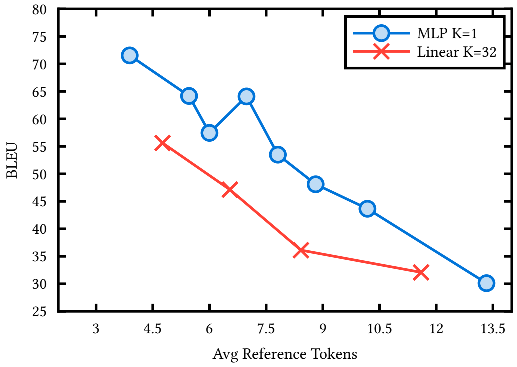

Figure 3: BLEU vs reference length on MS-MARCO. Both models decline with length; MLP K=1 starts higher but drops steeply.

Figure 3 shows a near-monotonic drop in BLEU with length for both best configurations. Pearson correlation between length and BLEU is strongly negative for MLP K=1 ($r = -0.97$) and Linear K=32 ($r = -0.95$), indicating length is the dominant predictor of reconstruction quality even within short queries.

4.2.2. Long Text (OpenWebText)

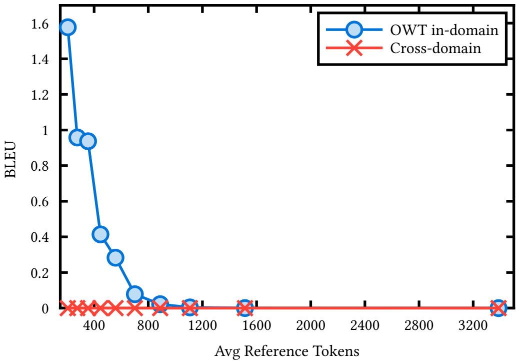

Figure 4: BLEU vs reference length on OpenWebText. In-domain training still collapses as length grows; cross-domain BLEU is effectively zero throughout.

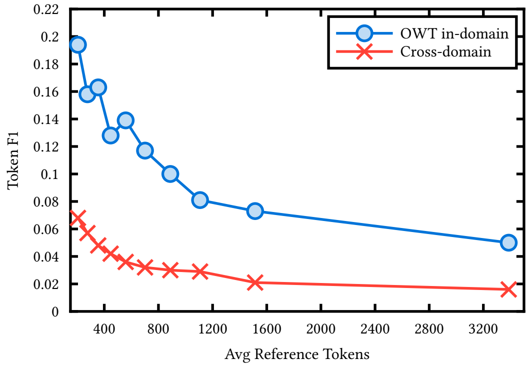

Figure 5: Token F1 vs reference length on OpenWebText. In-domain retains partial lexical overlap longer than cross-domain.

Ref Tokens	In-Domain BLEU	Cross-Domain BLEU	In-Domain F1	Cross-Domain F1
159–240	1.58	6.2e-7	0.194	0.068
240–315	0.96	5.2e-10	0.158	0.057
315–399	0.94	4.1e-13	0.163	0.048
400–503	0.41	3.1e-16	0.128	0.042
505–622	0.28	2.8e-18	0.139	0.036
627–800	0.08	8.0e-23	0.117	0.032
801–976	0.02	1.0e-29	0.100	0.030
979–1282	0.004	1.5e-30	0.081	0.029
1283–1807	0.00025	4.8e-35	0.073	0.021
1815–17245	0	0	0.050	0.016

Table 3: OpenWebText length-bucketed analysis. In-domain: model trained on OWT. Cross-domain: MS-MARCO adapter evaluated on OWT.

Figure 4, Figure 5, and Table 3 show a steep length dependence on OpenWebText. Even in-domain BLEU drops by three orders of magnitude as length grows, with length vs BLEU correlation $r = -0.57$. Token F1 remains non-zero but steadily declines, indicating partial lexical overlap without sequence-level fidelity. Cross-domain BLEU is effectively zero for all buckets, and cross-domain F1 decays toward zero with length.

4.3. LoRA Experiments

Early LoRA runs that trained both adapter and LoRA jointly increased loss relative to adapter-only training. We hypothesized that jointly training the adapter corrupts the learned prefix representations. To test this, we ran two configurations starting from the best adapter checkpoint (MLP $K = 1$).

Configuration	BLEU	Token F1
Adapter-only (baseline)	47.99	0.775
LoRA + unfrozen adapter	5.39	0.401
LoRA + frozen adapter	47.99	0.775

Table 4: LoRA configuration comparison. Results pending re-run.

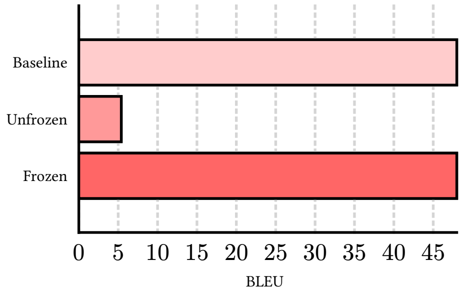

Figure 6: LoRA configuration BLEU scores. Frozen adapter preserves baseline quality; unfrozen collapses.

The results confirm the hypothesis: updating the adapter and LoRA simultaneously destroys the learned prefix mapping, while freezing the adapter preserves baseline quality. However, the frozen-adapter LoRA does not improve over baseline, suggesting the decoder is not the bottleneck under this objective.

4.4. Cross-Domain Evaluation

To test generalization, we evaluate adapters trained on one dataset against a different dataset.

Train Data	Eval Data	BLEU	Token F1
MS-MARCO	MS-MARCO	47.99	0.775
MS-MARCO	OpenWebText	0	0.038
OpenWebText	OpenWebText	0.012	0.120
OpenWebText	MS-MARCO	0.141	0.088

Table 5: Cross-domain evaluation. Transfer fails completely in both directions.

Cross-domain transfer largely fails (Table 5). The MS-MARCO adapter produces near-zero BLEU on OpenWebText, while the OpenWebText adapter achieves only marginal BLEU on MS-MARCO. This suggests the adapter learns dataset-specific decoding patterns rather than general embedding-to-text mappings. The fundamental limitation is the compression ratio: short queries (8 tokens) can be reconstructed, but multi-hundred-token passages cannot.

4.5. Embedding Model Comparison

Embedding Model	BLEU	Token F1
E5-base-v2	47.99	0.775
MiniLM-L6-v2	33.60	0.668
Qwen3-Embedding-4B	47.06	0.791

Table 6: Embedding model comparison (MLP $K = 1$) on MS-MARCO.

The smaller MiniLM model underperforms E5-base-v2 by 14 BLEU points (Table 6), suggesting that embedding capacity and/or pretraining quality directly impacts reconstruction quality. Qwen3 embeddings match E5 BLEU while slightly improving token F1.

5. Discussion

The K sweep reveals an unexpected divergence between adapter architectures. The MLP adapter performs best at $K = 1$ (BLEU 47.99), with quality degrading monotonically as K increases. In contrast, the linear adapter improves with larger K, peaking at $K = 32$ (BLEU 38.74). This suggests the nonlinear MLP can pack more information into a single prefix token, while the linear projection needs additional capacity to achieve comparable results.

The MLP adapter at $K = 1$ achieves a compression ratio of 7.83 (reference tokens per prefix token), substantially higher than the linear adapter at the same K. This compression advantage likely stems from the MLP's ability to learn nonlinear feature combinations that better exploit the decoder's prefix attention mechanism.

6. Limitations

- Results are limited to two datasets (MS-MARCO queries, OpenWebText passages) and a single decoder model (Llama-3.2-1B-Instruct).
- OpenWebText references are far longer than the generation cap (max_new_tokens=128), which depresses BLEU independently of domain shift.
- BLEU and token F1 penalize topical reconstructions that capture semantic content without exact wording.
- Evaluation uses 500 samples; a larger test set may reveal additional variance.

7. Conclusion

We demonstrate that frozen sentence embeddings can be partially decoded back into text using lightweight prefix adapters. The MLP adapter achieves BLEU 47.99 at $K = 1$ on MS-MARCO queries, while requiring only adapter parameters to be trained. However, the approach fails on longer texts and does not transfer across domains, suggesting fundamental limits to embedding-to-text reconstruction.