Evaluating Deletions with a DNA Masked Language Model

Project 6

December 30, 2025

Goal of the Evaluation

Goal: Quantify how **deletions** affect a DNA masked language model (GLM).

We compare two sequences:

- **Reference (ref):** undeleted sequence
- **Perturbed (alt):** same sequence with deletions (encoded as '-')

Both sequences are processed identically by the model.

Key idea: Any difference in model behavior can be attributed directly to the deletion.

Method 1: Fast Δ Log-Likelihood Score

Question: Does a deletion make the sequence globally less plausible for the model?

We compute a fast, global score by comparing the summed log-probabilities of the reference and perturbed sequences.

Definition:

$$\Delta = \sum_{i=1}^{L} \log p_{\text{model}}(x_i^{\text{alt}}) - \sum_{i=1}^{L} \log p_{\text{model}}(x_i^{\text{ref}})$$

Interpretation:

- $\Delta \ll 0$: deletion strongly disrupts the sequence
- $\Delta \approx 0$: deletion has little global effect

Why the Δ Log-Likelihood Score is Useful

- Requires only one forward pass per sequence
- Very fast and scalable
- Suitable for screening and ranking deletions

Note: Because the model is a masked language model, this is a *pseudo-likelihood*, not a true generative likelihood.

Nevertheless, it is effective as a **comparative score** between ref and alt.

Method 2: Influence / Probability-Shift Score

Question: Where and how does a deletion change the model's predictions?

Instead of scoring the whole sequence at once, we analyze **position-specific changes in predicted nucleotide distributions**.

For each target position j:

1. Mask position j in the reference sequence
2. Mask position j in the perturbed sequence
3. Compare the predicted distributions

This captures both **local and non-local effects** of deletions.

Influence Score: Mathematical Form

Let $p_{\text{ref}}(v)$ and $p_{\text{alt}}(v)$ be the predicted probabilities for nucleotide $v \in \{A, C, G, T, -\}$ at position j.

Default shift metric:

$$\text{shift}(j) = \max_{v} \left| \log p_{\text{alt}}(v) - \log p_{\text{ref}}(v) \right|$$

Final influence score:

$$\text{Influence} = \frac{1}{N} \sum_{j \in \text{targets}} \text{shift}(j)$$

Higher scores indicate stronger or more widespread effects of the deletion.

Why We Use Both Scores

The two scores answer complementary questions:

Score	Question	Strength
Δ log-likelihood	Global disruption?	Fast, robust
Influence score	Where/how predictions change?	Local, mechanistic

Together, they provide both a global and a local view of deletion effects.

Model Quality Score (Step 2)

Before interpreting deletion effects, we evaluate the model itself.

Using held-out sequences with random masking, we compute:

- Masked language modeling (MLM) loss
- Perplexity
- Masked-token accuracy

Purpose:

- Sanity check that the model learned meaningful sequence structure
- Enable comparison between different models or training setups

Next Steps

- Scale evaluation to larger datasets and more deletions
- Compare scores across different models or architectures
- Analyze positional patterns of high influence scores
- Relate model-based scores to biological annotations