Don’t Deploy Genomic Foundation Models Without Privacy Evaluation

Reconstruction Vulnerability Is Unpredictable Without Empirical Measurement

Reem Al-Saidi[1*], Erman Ayday[2], Ziad Kobti[1]

[1]School of Computer Science, University of Windsor
[2]Dept. of Computer and Data Sciences, Case Western Reserve University
*alsaidir@uwindsor.ca

NEURAL INFORMATION PROCESSING SYSTEMS

No genomic LLM should be deployed without first measuring whether its embeddings reconstruct the underlying sequences.

The threat

Input reconstruction from released inference-time embeddings

A data holder computes **z = f(x)** for a patient sequence x and releases z to an EaaS provider or collaborator. The adversary recovers x nucleotide by nucleotide. **Released embeddings cannot be recalled.**

Honest-but-curious adversary with black-box query access and auxiliary sequences from the same distribution. The target was never in training, so this sits **outside the DP-SGD threat model**.

The evidence base

Three studies, one chronology

Pan et al. 2020 [1] · pretrained baseline
Eight general-purpose transformers with an HS3D genome case study; up to 62.4% per-position accuracy on GPT-2.

Al-Saidi et al. 2025 [2] · fine-tuning effect
Six architectures, the same HS3D 20-nt sequences and MLP attacker; pretrained vs fine-tuned CLS / last-token embeddings.

Ouaari et al. 2026 [3] · Embeddings-as-a-Service
DNABERT-2, Evo 2 and NTv2 without fine-tuning; per-token and mean-pooled embeddings. Attacks generalise to 1000 Genomes individuals.

Attacks are lower bounds
An evaluation can establish that a model **is vulnerable, never that it is safe**. Attacker success certifies leakage (Fano’s inequality); attacker failure certifies nothing.

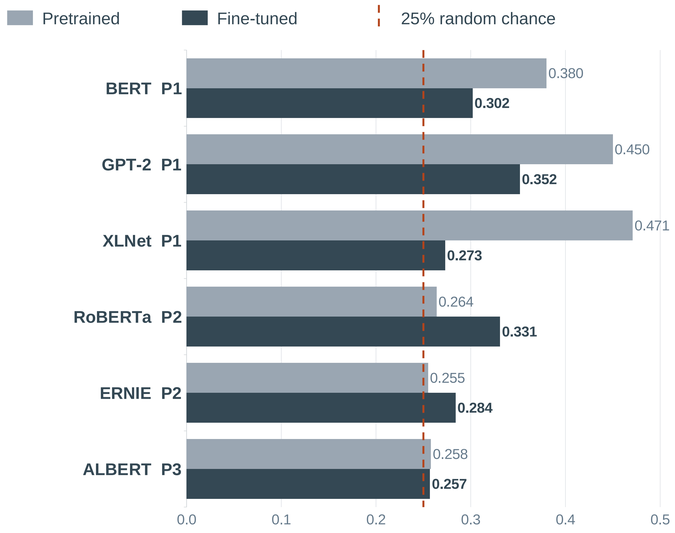
Reconstruction is the precondition; identification is the harm
30–80 independent SNPs suffice to identify a person [4]. One 20-nt window holds few, but an EaaS transfer holds thousands of embeddings: **the unit of risk is the release, not the window.**

Call to action

- Researchers: report reconstruction alongside task accuracy
- Venues: require its disclosure
- Developers: document privacy profiles in model cards
- GDPR Art. 9 and HIPAA: cover reconstructible embeddings
- Funders and clinical approval: require architecture-aware audits

Three fine-tuning profiles

Mean reconstruction accuracy (point estimates) · CLS / last-token · HS3D 20-nt · position-specific MLP [2]



P1 fine-tuning reduces vulnerability · **P2** increases it · **P3** redistributes it; the aggregate is unchanged

The profile cannot be derived from architecture descriptions or benchmark results; it must be measured.

Four interacting factors

Their joint effect cannot be predicted, only measured

Embedding type
Per-token: near-perfect reconstruction regardless of foundation model used [3]. Aggregated: reduced, not eliminated.

Fine-tuning effect
Reduces, increases or redistributes vulnerability, depending on how pretraining distributed information.

Pretraining & tokenization
Fixed before any practitioner acts. BPE makes DNABERT-2 more resilient to mean-pooled inversion than Evo 2 or NTv2.

Sequence length
Non-monotonic: Evo 2 scores 0.58 at 10 nt vs 0.98–0.99 at 15–25 nt; NTv2 falls from 90% to 57% between 10 and 100 nt. No single safe length.

A minimum evaluation standard

Three sequential decisions, two attacker tiers

Decision 1 · Embedding type

per-token
Never share

aggregated
Proceed to Decision 2

Decision 2 · Measure at the deployment embedding type

Fine-tuning context CLS / last-token

EaaS context mean-pooled

Profile 1 · BERT, GPT-2, XLNet
FT reduces ⇒ fine-tune before sharing

Pretrained measurement
DNABERT-2, Evo 2, NTv2: varies by model and length

Profile 2 · RoBERTa, ERNIE
FT increases ⇒ linear probe; do not fine-tune

If FT applied: re-evaluate
FT effect on genomic FMs unmeasured

Profile 3 · ALBERT
FT redistributes ⇒ position-specific audit; aggregate misleads

No safe default
Profile assignment requires measurement

Decision 3 · Match attacker strength to stakes

Tier 1 · research
Position-conditioned MLP (3 hidden layers, 200–400 units, ≥10^4 iterations) + nearest-neighbour baseline. Under 1 h per configuration on one GPU.

Tier 2 · clinical
GDPR Art. 9 / HIPAA. Vec2Text-style transformer inversion [5]; within a day on multi-GPU. Any release triggers Tier 2.

Reporting rules

- **Every claim names its attacker:** *“x% against an MLP attacker at length ℓ”*
- **A Tier 1 pass does not certify Tier 2 deployment**

DISCLOSE, AT MINIMUM

(i) embedding type
(ii) sequence length(s)
(iii) pretrained accuracy, avg + per-position
(iv) post-FT accuracy, per-position
(v) accuracy at polymorphic positions
(vi) reference-matching baseline
(vii) variant content of the window
(viii) embeddings released per individual

(i)–(iv) characterize the model; (v)–(viii) convert reconstruction into identification risk.

References

[1] X. Pan, M. Zhang, S. Ji, M. Yang. Privacy risks of general-purpose language models. IEEE S&P, 2020.
[2] R. Al-Saidi, E. Ayday, Z. Kobti. Comparing reconstruction attacks on pretrained versus full fine-tuned LLM embeddings on Homo sapiens splice sites genomic data. IEEE TrustCom, 2025.
[3] S. Ouaari, J. Kreuer, N. Pfeifer. How private are DNA embeddings? Inverting foundation model representations of genomic sequences. arXiv:2603.06950, 2026.
[4] Z. Lin, A. B. Owen, R. B. Altman. Genomic research and human subject privacy, 2004.
[5] J. Morris, V. Kuleshov, V. Shmatikov, A. M. Rush. Text embeddings reveal (almost) as much as text. EMNLP, 2023.