Comment on Guidelines 02/2026 on Anonymisation: text embeddings and vector representations

Yoshiaki SHIRAISHI, Kobe University

Affiliation listed for identification purposes only.

Summary. Text embeddings — the vector representations stored in retrieval-augmented generation (RAG) systems and vector databases — are record-level, high-dimensional data. The draft guidelines do not mention them. We submit empirical evidence that (i) personal names can be reconstructed from text embeddings with publicly available tools on commodity hardware, (ii) common transformations such as noise addition and 8-bit quantisation do not reliably prevent this, and (iii) the effectiveness of such transformations varies by an order of magnitude across embedding models and data distributions, so that it cannot be assessed once and generalised. We propose an explicit example, a sentence on attacker assumptions, and a minimal assessment record.

1. Why embeddings need explicit treatment. Converting documents into embeddings can easily be mistaken for anonymising them. Under the draft's framework, an embedding of a single sentence is record-level data of high dimensionality and resolution (Section 3.5.1), and published embedding-inversion attacks aim precisely to recover the underlying record, including its identifiers. Without an explicit reference, controllers may not recognise that Section 3 applies to these representations. We offer the evidence below in the spirit of para 84, which encourages scientifically grounded expertise on the evolving state of the art of re-identification techniques.

2. Reconstruction is feasible with means reasonably likely to be used (para 92). Using a publicly available inversion method (vec2text) and publicly released model checkpoints, we reconstructed personal names from sentence embeddings. On biography-style sentences carrying fictitious names, exact name recovery without protection exceeded 97% for both embedding models we tested (GTR-base and multilingual-e5-base). All experiments ran on a standard CPU workstation or a single desktop GPU system, without paid cloud computing; reconstructing one sentence took about three minutes on a CPU in our measurement. This illustrates para 92: time and cost are not prohibitive.

3. Noise and quantisation are not reliable safeguards (para 64). With Gaussian noise at a level that reduced retrieval quality on a standard benchmark (BEIR NFCorpus) by only 1.5%, 38% of names were still recovered exactly; an attacker who retrained the inversion model on noised embeddings recovered 56%. In a smaller evaluation on 20 synthetic sentences, 8-bit quantisation did not reduce reconstruction for either embedding model. Although embedding inversion is a different attack from the record linkage discussed in para 64, these results illustrate the same observation that noisy data may leak information in unexpected ways.

4. Results do not transfer across models or data (paras 29, 85–86). At the same noise level, the second embedding model (multilingual-e5-base) allowed only 3% name recovery (7% for the adaptive attacker) but lost 21.7% of retrieval quality. Name recovery depended strongly on how frequent the name components were in the model's training data. Numeric identifiers that appeared unrecoverable on synthetic templates were recovered 47–64% of the time on realistic sentence structures (counting only identifiers present in the input). An assessment performed for one model, one data distribution, or on synthetic data therefore does not establish the result for another.

5. Attacker assumptions (paras 30–31). On the full 120-item biography-style set, adapting the inversion model to the noise increased name recovery by about 4–18 percentage points at the two

noise levels tested, so evaluations against non-adapted attackers underestimated the risk. Paras 30–31 already ask that the range of relevant entities be considered without relying on a lack of motivation; building on this, we suggest that assessments assume an attacker who knows the transformation and can adapt to it. This is our proposal, not a requirement stated in the draft.

6. Proposals.

a. Add an example in Section 3.3 or 3.5.1 addressing text embeddings and other learned vector representations of personal data, stating that they are to be assessed under the three criteria like other record-level data.

b. Add a sentence stating that where a transformation (e.g. noise addition, quantisation, dimensionality reduction) is relied upon, its effectiveness should be evaluated against attackers who know and can adapt to that transformation, and for the specific model and data distribution concerned.

c. Illustrate a minimal assessment record (model and version, transformation and parameters, entity perspective, attacker assumptions including adaptation, data distribution, recovery rates distinguishing correct reconstruction from format-only matches, utility impact, cost, and date of assessment).

Limitations. These are results from a single study, not yet peer reviewed, on English-language text with two embedding models. A post-hoc screening found that name substitution had left identifying phrases of the original subjects in 10 of the 120 biography-style sentences; after a manual review, 47 sentences were set aside. Restricting the analysis to the remaining 73 changed name-recovery rates by at most about five percentage points (injected identifiers: about eight) and did not change any conclusion. The results are offered as an illustration of the risks the guidelines already describe, not as a general statement that all embeddings are personal data.

Data and ethics. No human participants were involved. Evaluation sentences contain either invented identifiers or the sentence structure of public Wikipedia biographies (the wiki_bio dataset), with the leading personal name replaced by a fictitious name and a fictitious date of birth or telephone number inserted. Because such sentences keep the wording of a public corpus, they could be matched back to the source biographies, and we treat them as potentially linkable rather than anonymous. A small set of unmodified public biographies served only as an upper-bound control for training-data memorisation and is reported in aggregate only. No individual evaluation records or reconstruction outputs are reproduced in this comment.

Data and code. Code, synthetic data and aggregate results will be released in a public repository. Until then they are available from the author on request.

Disclosure. This comment is submitted in a personal capacity. The views expressed are solely those of the author and do not represent the position of Kobe University or of any other organisation with which the author is affiliated, including a company that may develop tools related to the assessment described here.