Comment on Guidelines 02/2026 on Anonymisation

1. Standpoint

I write as a researcher working on confidential, meaning-preserving de-identification of conversational text: counselling and healthcare conversations, and synthetic conversation outputs derived from them. My interest is practical. When a conversation is de-identified, what remains is not a table of fields. It is a sequence of turns in which a person describes events, relationships, routines, and ways of speaking. The question the Guidelines must answer for this kind of data is whether that sequence can still single someone out, link to other information about them, or support an inference about them, after names and direct identifiers are gone.

I welcome the draft's framework, and in particular the contextual approach. Anonymity judged from each relevant entity's perspective, rather than in the abstract, is the only version of the test that fits conversational data, because the same output is far more identifying in the hands of someone who already knows the speaker than in the hands of a stranger. I also welcome the three criteria, No Record Isolation, No Linkage, and No Inference, as the right structure. My comments below are endorsements with reasons, and four specific requests. They are about the framework, not a claim that any current method achieves anonymity for conversational text.

2. What the draft gets right, and why it matters for conversational data

Prompting and querying as inference (paragraph 82). The draft states that inferences can be drawn from data representing correlations rather than the original records,

in particular for AI models or synthetic data, by querying or, for AI models, prompting the given data with additional information. This is the correct test for conversational outputs. A de-identified transcript, or a synthetic conversation generated from one, does not need to contain a record of the person to leak information about them. A prompt that supplies a known fact about the person can elicit further facts that are correlated with it in the output. The identifying signal need not be a field in the output at all. I endorse this paragraph, and I ask that it be kept in the final version without narrowing it to record-level data.

AI agents as a re-identification technique (paragraph 83). The draft's range of techniques, from a basic search to special-purpose AI agents deploying a sophisticated combination of probabilistic techniques, correctly describes the attacker class that matters most for conversational data. An agent can gather additional information from multiple public sources, combine it, and test candidate identities against a conversation's episodic facts, relational structure, and discourse patterns. That is linkage across semantic layers, not matching of tokens. Paragraphs 92 and 93 make the same point from the cost side: techniques are increasingly accessible on commodity hardware, and agentic AI will likely further reduce the time and cost of gathering and combining information from multiple sources. I endorse both, and I endorse the consequence drawn in paragraph 88, that contextual assessments should incorporate adequate safety margins so the assessment remains robust as entities' capabilities evolve. A test passed against today's attacker is not a test passed.

The leave-one-out test (paragraph 71(b)). The draft's test for whether an inference relies on the given data, comparing inference from data produced with a person's records against inference from data produced without them, is the most falsifiable standard in the document. For model and synthetic outputs, where tracing how an

inference was drawn is often impossible because of opacity, this comparison is the practical alternative to assertion. I endorse it, and I ask that the final Guidelines recommend it explicitly as the evaluation method for AI-model and synthetic outputs, rather than leaving it as one option among three.

Sufficiency rather than perfection (paragraph 87). The draft states that a re-identification technique need not be perfect or certain, only sufficiently close and confident enough to distinguish and treat a person differently, and that it can succeed against a single individual in a larger set. This matters because it blocks a common misreading: that an attacker method which fails most of the time is therefore safe. For conversational data, a method that identifies one person out of many, with confidence, is a failure of anonymity for that person, not a partial success for the method.

The caution on motivation (paragraph 31). The draft cautions against using lack of motivation as a factor in the analysis. I endorse this caution. For conversational data, the relevant entities include people who know the speaker, former colleagues, family members, and journalists, whose motivation cannot be assumed away. A contextual assessment that discounts motivated entities will understate linkage risk for exactly the outputs most likely to be shared, such as de-identified transcripts used for research or training.

Linkage as inference, including for aggregate data (paragraphs 80 and 81). The draft notes that linkage can be seen as inference, that information need not match on a one-to-one basis to support relevant inferences, and that this applies to aggregate data as well as record-level data. This is directly relevant to surrogate replacement and noise-based methods for conversations: changing surface forms, names, dates, or wording does not by itself defeat

linkage, because the remaining structure, the sequence of events, the relationships described, and the pattern of speech, can still be matched to additional information. A method should be tested for linkage, not assumed anonymous because its surface changed.

3. Four requests

First, add a conversational worked example. The draft's examples are drawn from tables, histograms, and location tracks. None shows a de-identified conversation transcript or a synthetic conversation output. I ask the EDPB to add one: a short transcript from which direct identifiers have been removed, followed by an analysis of how episodic facts (a distinctive event, a workplace, a timeline), relational structure (who is described in relation to whom), and discourse patterns (recurring phrases, ways of describing a problem) could still support linkage to additional information, and how paragraph 71(b) would be applied to test whether an inference drawn from it relies on the person's records. Without such an example, practitioners working with text will map the criteria onto fields, and will miss the threat class that matters most for their data.

Second, recommend an evaluation protocol for model and synthetic outputs. For any claim that an AI model or synthetic dataset derived from personal data is anonymous, the final Guidelines should recommend reporting three things: the candidate-set size against which identification was tested, the confidence threshold applied, and the paragraph 71(b) comparison between outputs produced with and without the person's records. A claim reported without these three items cannot be checked, and an uncheckable claim should not support a finding of anonymity.

Third, state expressly that surrogate replacement is a

method to be tested, not a result. The draft already implies this through paragraphs 80 to 82, but practitioners need it stated. Replacing names, places, and dates with surrogates, or perturbing wording, changes surface forms. It does not by itself satisfy No Linkage or No Inference, because linkage can proceed through structure and content that survives replacement. The final Guidelines should say so in terms, so that a surrogate pipeline is not treated as anonymisation by construction.

Fourth, retain the safety-margin and motivation provisions in the final version. Paragraph 88 on safety margins and paragraph 31 on motivation are the provisions that keep a contextual assessment honest over time, as capabilities improve and as outputs leak or are shared beyond their intended recipients. Weakening either would make the framework easier to pass and less protective in exactly the cases where conversational outputs circulate most widely.

4. Limits of this comment, stated plainly

I offer no claim that any method currently available achieves anonymity for conversational text, and no result from my own program should be read as such a claim. In two pre-registered evaluations, a local-model attacker track built to test de-identified conversation outputs failed to meet its own pre-registered ceiling, and that track is parked pending a stronger local model. I cite that failure because it is the relevant evidence: the threat class described in paragraphs 82 and 83 is real, and the instruments for testing it are not yet good enough to certify safety. A comment that cited only successes would misrepresent the state of the work, and the Guidelines should be read with that limitation in mind. The contribution I claim for this comment is narrower: that the framework's tests, applied honestly and with the evaluation protocol requested above, are the right tests for

conversational data, and that the worked example requested above would make them usable for it.

5. Closing

I thank the Board for the consultation and for a draft that already contains, in regulator form, the core distinctions this field needs: anonymity relative to entity, linkage across datasets, and inference by prompting. My requests are intended to make those distinctions operational for conversational and synthetic outputs, where the identifying signal lives in meaning rather than in fields. I would welcome a worked example and an evaluation protocol that let practitioners test their outputs against the draft's own standards, and report the results in a form others can check.

Luis Vasquez, M.S., Optimust Intelligence, LLC