

<https://doi.org/10.31891/csit-2026-3-2>

Iegor PAVLENKO

doctoral student of Department of
Information Technology and Computer
Engineering,
Dnipro University of Technology,
Dnipro, Ukraine.

<https://orcid.org/0009-0004-0600-3090>
pavlenko.yeh.v@nmu.one

Volodymyr HNATUSHENKO

Doctor of Technical Sciences (Dr.-Ing.
habil.), Professor, Head of Department of
Information Technology and Computer
Engineering,
Dnipro University of Technology,
Dnipro, Ukraine.

<https://orcid.org/0000-0003-3140-3788>
hnatushenko.v.v@nmu.one

Received: 24/08/2026

Accepted: 14/09/2026

Published: 30/09/2026

© Copyright
2026 by the author(s)



This is an Open Access article distributed
under the terms of the [Creative Commons
CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)

UDC 004.4:004.8

**SEMANTIC EMBEDDINGS AS A CANDIDATE-
GENERATION CHANNEL FOR SANCTIONS
NAME SCREENING**

Sanctions name screening is an open-set identification problem with asymmetric error costs: a missed designee is more consequential than an extra alert. We evaluate OpenAI text-embedding-3-large and Google gemini-embedding-001 as candidate-generation channels over a fixed OFAC SDN individuals snapshot (7,426 entries; primary firstName + lastName only), using a controlled stress test in which the same 150 designated individuals are rendered in four Slavic-name regimes — direct Latin spelling, ICAO-style romanisation, French-convention transliteration, and a Russian form converted to Ukrainian and officially romanised — at two specificity levels, against easy and deliberately hard look-alikes, at three Matryoshka dimensions (768, 1536, 3072). Gemini led in every non-trivial regime and its margin grew with transliteration distance, reaching 34.6 pp of Recall@1 on two-part Russian-to-Ukrainian forms (entity-cluster bootstrap 95% CI 27.1–42.1 pp; exact McNemar $p \approx 1 \times 10^{-16}$), with no systematic loss down to 768 dimensions, whereas OpenAI degraded with both transliteration distance and reduced dimension. Top-1 similarity did not separate transliterated positives from hard look-alikes: for OpenAI on the hardest regime the positives scored below the negatives (AUC 0.318, 95% CI 0.251–0.385). At an equal candidate budget, unioning both models' top-5 lists matched Gemini alone at $k = 10$ exactly (99.36%), and one designee sat at rank 60 despite a one-character surname difference. Embedding similarity is therefore a useful candidate-generation channel but not a reliable accept/reject rule; on a limited, partly synthetic corpus queried against mutable hosted APIs these are benchmark-specific findings, not a no-miss guarantee. Data, results and figures are released.

Keywords: sanctions screening; entity resolution; name matching; candidate generation; semantic search; vector search; dense retrieval; text embeddings; recall; AML/CFT compliance.

Introduction

Obligated entities must screen customers and counterparties against sanctions watchlists. For individuals the core task is name matching: deciding whether an incoming personal name refers to a listed person despite spelling variation, transliteration, missing or reordered name parts, and patronymics. Incoming names arrive in many orthographies — a Russian designee may appear under an ICAO transliteration, a French rendering or an official Ukrainian one.

Screening decomposes into blocking (candidate generation), pairwise verification, open-set identification, and the regulated disposition an analyst makes using identifiers beyond the name. This paper measures blocking only, and reports score separability solely to test whether the score could serve as a decision rule. We evaluate hosted, general-purpose services rather than reimplemented research systems, and use a synthetic corpus because it buys control over the one variable of interest — the same 150 individuals re-rendered under progressively more distant conventions — at the price of a query set that is small, partly artificial and internally dependent (see Dataset audit and remaining controls; Threats to Validity).

• *RQ1*. How does candidate recall change across the transliteration regimes and specificity conditions, and what failure modes appear in the rank tail?

• *RQ2*. What is the trade-off between embedding dimension, recall and observed latency?

• *RQ3*. Can the top-1 score, alone or ensembled, act as an accept/reject rule against hard look-alikes?

We frame the negative result narrowly: on this benchmark, embedding similarity is a strong candidate-generation channel for in-script variation but not a sufficient decision rule for cross-transliteration screening.

Background and Related Work

Watchlist screening is record linkage under an extreme cost asymmetry: a missed designee carries regulatory and legal liability, a false alert costs analyst time [1]. Review is human-in-the-loop by requirement [2, 3], and OFAC frames a "valid match" as a judgement over multiple identifiers rather than a similarity threshold [4, 5], so a score is a queue-ordering signal, not a disposition. The layer measured here is blocking: a recall-oriented filter that must place the correct record in a small candidate set [6], formalised as a recall/candidate-set-size trade-off [7]. TF/IDF blocking is a strong, cheap baseline [8]; pre-trained embeddings [9] and retrieval-style formulations [10] have been analysed for it; one-to-one graph matching is a later, distinct stage [11].

Sentence and text embeddings [12, 13] map text to vectors for nearest-neighbour search. MTEB [14] establishes that no single model dominates across tasks — hence a task-specific benchmark rather than a general ranking — while Kamaloo et al. [15] raise the domain-generalisation and reproducibility questions behind Dataset audit and remaining controls, BGE-M3 [16] is the most obvious open baseline to add next, and both evaluated services document Matryoshka representations [17, 18, 19] that trade output dimension against quality.

Transliteration is a modelling problem in its own right: cross-script person-name matching including Cyrillic [20], context-aware romanisation in another language family [21], and a cross-lingual named-entity corpus for six Slavic languages [22] are the nearest prior work, none of them a sanctions benchmark. Our regimes follow ICAO Doc 9303 [23], the official Ukrainian romanisation of Cabinet of Ministers Resolution No. 55 [24] and BGN/PCGN [25], the last only as a stress-test transformation.

Two gaps motivate the study: blocking work rarely reports candidate recall and score separability jointly against hard look-alikes under controlled transliteration variation, and hosted-service evaluations rarely characterise the rank tail, which decides whether a person can be missed. We compare two hosted services to each other, not to string, phonetic, sparse or PLM-based matchers; large-scale benchmarks such as OpenSanctions Pairs [26] (755,540 labelled pairs from 293 sources) are complementary rather than superseded.

Task Formalisation

Let L be the indexed list — the OFAC SDN individuals, one vector per entry — and q a personal-name query; the channel returns the k nearest entries $C_q(k)$ by cosine similarity, with scores $s(q, e)$, where $e \in L$ denotes one indexed list entry. For a positive query, G_q — the entries an analyst would accept as the same person — is here a singleton, the entity id the query was generated from, matched as $\text{str}(\text{ground_truth.entity_id}) == \text{str}(\text{result.entity_id})$. That fits an index built without aliases (see Data and Experimental Setup). We report candidate recall@k ($G_q \cap C_q(k) \neq \emptyset$), the true-entity rank, right-censored at the depth run, and the alert score $\max_e s(q, e)$, whose usability as a decision rule is RQ3 rather than an assumption. Negatives carry `is_sanctioned = false`, for which recall is undefined and the question is false-alert pressure.

Data and Experimental Setup

Snapshot, index and models. Table 1 states the configuration held fixed for every run reported below.

Table 1

Fixed experimental configuration: snapshot, index, models and query runs

Item	Value
Source	https://sanctionslistservice.ofac.treas.gov/api/PublicationPreview/exports/SDN_XML.ZIP
Snapshot taken	14 May 2025, 18:34:19 (SDN_XML.ZIP)
Individual entries indexed	7,426
Fields embedded	firstName + lastName of the list entry only; aliases/AKAs not indexed
Store	vector DB; one collection per model × dimension; filter entity_type = Individual; cosine similarity
Models	OpenAI text-embedding-3-large; Google gemini-embedding-001
Dimensions	3072, 1536, 768 (positives and negatives); additionally 512 and 128 on negatives only
Vectors per query	1 (the full name string, as issued)
Depth	k = 5 standard; full runs at k ∈ {10, 20} and single-query probes at k ∈ {40, 80} for rus_ukr two-part, Gemini, 3072
Query runs	30 May – 27 June 2026; 96 JSONL run files, 13,932 query records

The snapshot is the fixed experimental corpus, not the live SDN list [27]. Aliases are out of scope by design, which isolates the channel's behaviour on the primary name form but bounds the external validity of every number in Results (see Threats to Validity). The negative-only 512- and 128-dimensional runs are reported in metrics_recomputed.csv rather than dropped.

Query design: regimes and specificity. Positive queries were generated from a fixed set of 150 designated individuals in four regimes ordered by increasing orthographic distance from the OFAC Latin spelling (Table 2).

Table 2

Transliteration regimes (positives), ordered by distance from the OFAC Latin spelling		
Regime	Description	Example query → OFAC target
lat	Direct Latin/Slavic spelling (closest to OFAC)	SECHIN Igor → <i>Igor Ivanovich SECHIN</i>
icao	ICAO-style romanisation of the Russian form	Sechin Igor → <i>Igor Ivanovich SECHIN</i>
french	French-convention transliteration	Chevtchenko Iouri → <i>Yuriy Valeryevich SHEVCHENKO</i>
rus_ukr	Russian form converted to a Ukrainian form, then official Ukrainian romanisation	Shoihu Serhii → <i>Sergei Kuzhugetovich SHOIGU</i>

Each regime ran at two specificity conditions, labelled 3-2 and 3-3 in the artefacts: two-part (surname given-name, e.g. Aliksieiev Oleh) and three-part (adding the patronymic). They are paired, not independent samples — the same 150 entities in both conditions and across all four regimes — so they form a controlled name-completeness factor, analysed as a paired contrast in Query specificity, and their agreement (Figure 3) is stability rather than replication (see Conclusions).

Negatives were drawn at three levels of difficulty (Table 3).

Table 3

Negative sets		
Set	Description	n (2-part / 3-part)
ns	random non-sanctioned Slavic names	150 / 150
ns_similar_nb	look-alikes sharing a name component with a real SDN individual	148 / 150
ns_similar_part	hardest construction: several name parts collide with a real designee	146 / 150

The hard sets recombine components of real SDN individuals with non-matching ones, so a component-level match exists but the person does not — adversarial by construction. They anchor a stress condition rather than estimating operational traffic, so hard-negative false-positive rates are not forecasts or bounds on review load.

Dataset audit and remaining controls. All 96 raw run files were audited programmatically (dataset_audit.csv); irregularities follow from the design and are reported rather than repaired. The lat and three-part negative runs are clean at 150 rows; elsewhere a designee may have several defensible transliterations (Razorienov / Razoronov / Rozoronov for entity 42874), giving 151 rows for icao and french and 156 for rus_ukr over the same 150 entities, while dropping the patronymic collapses distinct negatives onto one string (146–148 remain distinct). Query-level counts therefore weight entities unequally, so we report both units of analysis; they agree to within 0.84 pp of Recall@1 across the sixteen 3072-dimension cells, and one row moves a rate by ≈ 0.64 pp. Hosted APIs bring further limits [15]:

Model mutability. Both services are versioned only by marketing name; runs carry timestamps and the returned model identifier, not immutable snapshots, so a rerun may differ (see Threats to Validity).

Dimension procedure. Each dimension is a separately built collection: OpenAI requests the length via the dimensions parameter and relies on the returned unit-normalised vectors [17, 18], while Gemini requests outputDimensionality and both paths L2-normalise every vector below 3072 dimensions, as Google requires [19] — provider-supported Matryoshka outputs, not separately trained models or a post-hoc projection.

ANN layer. In-memory HNSW, $m = 16$, $ef_construct = 100$, quantisation disabled; the search-time parameter was not recorded and ANN recall was never measured against exact search, so reported recall is that of *the configured system* (see Conclusions).

Latency environment. Client-observed end-to-end latencies from one network location, confounded with network path, region, throttling and load (see Output dimension and observed latency).

Metrics and Statistical Treatment. We report Recall@1/@5, MRR, the true-entity rank distribution from the deep- k probes (right-censored at the depth run), top-1 score distributions with the rank-1 versus rank-2 gap, the AUC of the top-1 score used as an alert score per positive regime $\times$ negative set, operating points in-sample and held out, and latency percentiles.

Query-level rates carry Wilson score intervals; entity-level recall, MRR and AUC use a percentile bootstrap over entity clusters (2,000 resamples; 1,000 for AUC), reflecting the number of genuinely independent units rather than rows. Comparisons are paired — both models saw the same queries against the same snapshot — so we report exact two-sided McNemar tests with the paired entity-level Recall@1 difference and its cluster interval [28]; because

McNemar is query-paired and does not remove the within-entity dependence intrinsic to this design, the cluster interval is the primary uncertainty statement. At $n \approx 150$ one entity is ≈ 0.67 pp, so differences such as 99.34% versus 98.68% are not a stable advantage. One caution governs every score-based result: thresholding the maximum similarity asks whether a query looks like *someone* on the list, not whether the returned record is the right person, so the AUC figures describe *alert-triggering* behaviour, never identification accuracy.

Results

Candidate recall by regime (RQ1). Gemini matched or exceeded OpenAI in every cell, and the margin grew across the prespecified ordering of regimes (Table 4): both sit at ceiling on direct Latin names, while on rus_ukr two-part forms Recall@1 is 59.62% (Wilson 95% CI 51.8–67.0) for OpenAI against 94.23% (89.4–96.9) for Gemini. At depth 5 OpenAI misses the true entity entirely for 21–27% of rus_ukr queries against Gemini's 0.6–1.3% (MRR 0.652 versus 0.959), so on this benchmark OpenAI is insufficient as a sole candidate-generation channel for the far transliterations.

Table 4

Recall@1 and Recall@5 (%) at 3072 dimensions, query level. Δ is the paired entity-level Recall@1 difference (Gemini – OpenAI) with an entity-cluster bootstrap interval; p is the exact McNemar test. Wilson intervals are in metrics_entity_ci.csv

Regime	Condition	n	OpenAI R@1	Gemini R@1	OpenAI R@5	Gemini R@5	Δ @1 pp [95% CI]	p
lat	two-part	150	99.33	99.33	100.00	100.00	0.0 [–2.0, 2.0]	1.00
lat	three-part	150	100.00	100.00	100.00	100.00	0.0 [0.0, 0.0]	1.00
icao	two-part	151	92.05	97.35	97.35	98.68	+5.3 [0.7, 10.0]	0.039
icao	three-part	151	97.35	99.34	98.01	99.34	+2.0 [0.0, 4.0]	0.25
french	two-part	151	84.11	94.70	92.72	98.68	+10.7 [4.7, 16.7]	8.6×10^{-4}
french	three-part	151	84.77	97.35	95.36	98.68	+12.3 [7.3, 17.7]	3.8×10^{-6}
rus_ukr	two-part	156	59.62	94.23	73.08	98.72	+34.6 [27.1, 42.1]	1.1×10^{-16}
rus_ukr	three-part	156	65.38	97.44	79.49	99.36	+31.1 [23.3, 38.9]	2.4×10^{-14}

The paired reading is narrower: on icao three-part (3 discordant entities of 150) and on lat (0–2 discordant entities) the data support no claim of difference. The services are indistinguishable near the indexed spelling and diverge sharply as the rendering moves away from it; 157 of the 164 discordant pairs favour Gemini, and entity- and query-level recall differ by at most 0.84 pp.

Query specificity: two-part versus three-part names. A patronymic adds a distinctive token and helps everywhere, rarely by enough to be individually significant (Table 5): all eight paired estimates are positive, four intervals exclude zero, and only one (french, OpenAI) reaches below it. The gain is largest where the model is otherwise weakest — OpenAI on icao and rus_ukr — as expected if the failure mode is too little distinctive signal in a short string (see Ensembling at an equal budget, and the rank tail).

Table 5

Effect of adding a patronymic: paired entity-level Recall@1 difference (three-part – two-part) at 3072 dimensions, with discordant entities and the exact McNemar p .

Regime	Model	Discordant entities	Δ pp [95% CI]	p
lat	Gemini	1	+0.7 [0.0, 2.0]	1.00
lat	OpenAI	1	+0.7 [0.0, 2.0]	1.00
icao	Gemini	3	+2.0 [0.0, 4.7]	0.25
icao	OpenAI	8	+5.3 [2.0, 9.3]	0.008
french	Gemini	4	+2.7 [0.7, 5.3]	0.125
french	OpenAI	7	+1.0 [–2.0, 4.3]	1.00
rus_ukr	Gemini	5	+3.3 [0.7, 6.7]	0.063
rus_ukr	OpenAI	25	+6.8 [0.4, 13.1]	0.108

Output dimension and observed latency (RQ2). Gemini shows no systematic sensitivity to dimensional reduction: Recall@1/@5 and MRR are identical at 3072, 1536 and 768 on icao and french. OpenAI loses recall as the vector shrinks, at two parts 92.05 $\rightarrow$ 91.39 $\rightarrow$ 90.07% on icao, 84.11 $\rightarrow$ 81.46 $\rightarrow$ 78.15% on french and 59.62 $\rightarrow$ 55.13 $\rightarrow$ 51.28% on rus_ukr for 3072 $\rightarrow$ 1536 $\rightarrow$ 768. Latency — client-observed in one environment and confounded with network path, region, throttling and load — favours OpenAI (median 171 versus 273 ms, maximum 2,602 versus

25,338 ms over 6,966 requests each, interquartile ranges within 1 ms), which under a recall-oriented objective is secondary to a 35 pp recall difference (see Implications for architecture and model choice).

Score distributions and separability (RQ3). The models occupy different score scales — mean top-1 on negatives at 3072 dimensions is 0.583–0.686 for OpenAI against 0.900–0.938 for Gemini, whose scores are high and compressed — so no threshold transfers between models, dimensions or negative populations. The failure driver is that transliterated positives score low while hard look-alikes score high (Table 6): for OpenAI on rus_ukr the mean positive score (0.651) falls *below* the mean hard-negative score (0.686). AUC (Table 7) is excellent against easy negatives and collapses against look-alikes as transliteration distance grows — for OpenAI on french the interval covers 0.5 and on rus_ukr it lies wholly below it, an inverted rather than merely uninformative ordering — and is flat across dimensions (Gemini rus_ukr versus ns_similar: 0.745, 0.744, 0.748), so the failure tracks the transliteration gap, not the vector size. Figures 1 and 2 give the ROC and DET curves, which describe alert-triggering behaviour only (see Metrics and Statistical Treatment).

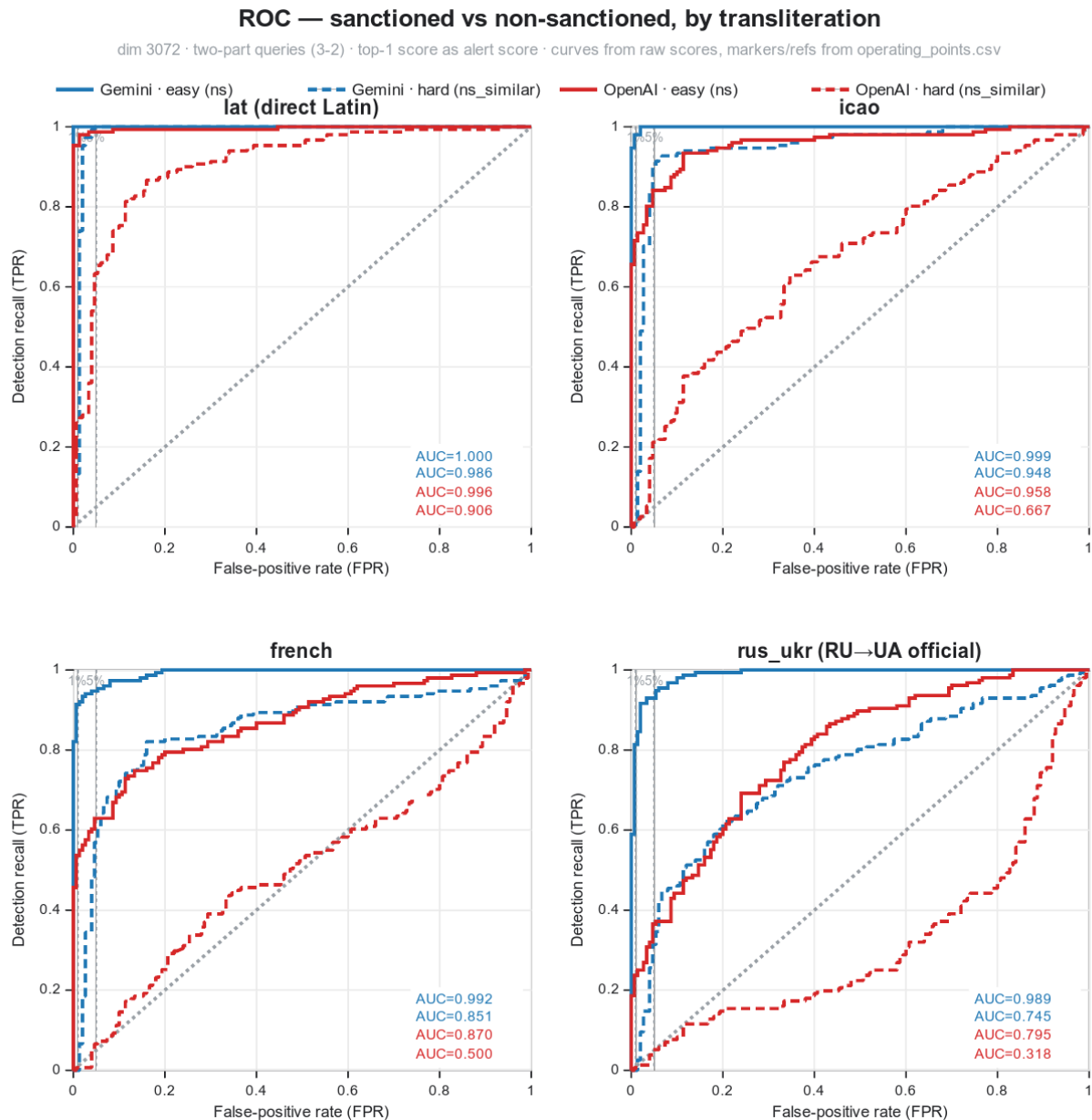


Fig. 1. ROC of the top-1 score used as an alert score, 3072 dimensions, two-part queries: {Gemini, OpenAI} × {easy ns, hard ns_similar}. Gemini's easy-negative curves hug the top-left corner; the OpenAI hard-negative curve on rus_ukr runs below the chance diagonal (AUC 0.318).

DET — sanctioned vs non-sanctioned, by transliteration

dim 3072 · two-part queries (3-2) · top-1 score as alert score · curves from raw scores, markers/refs from operating_points.csv

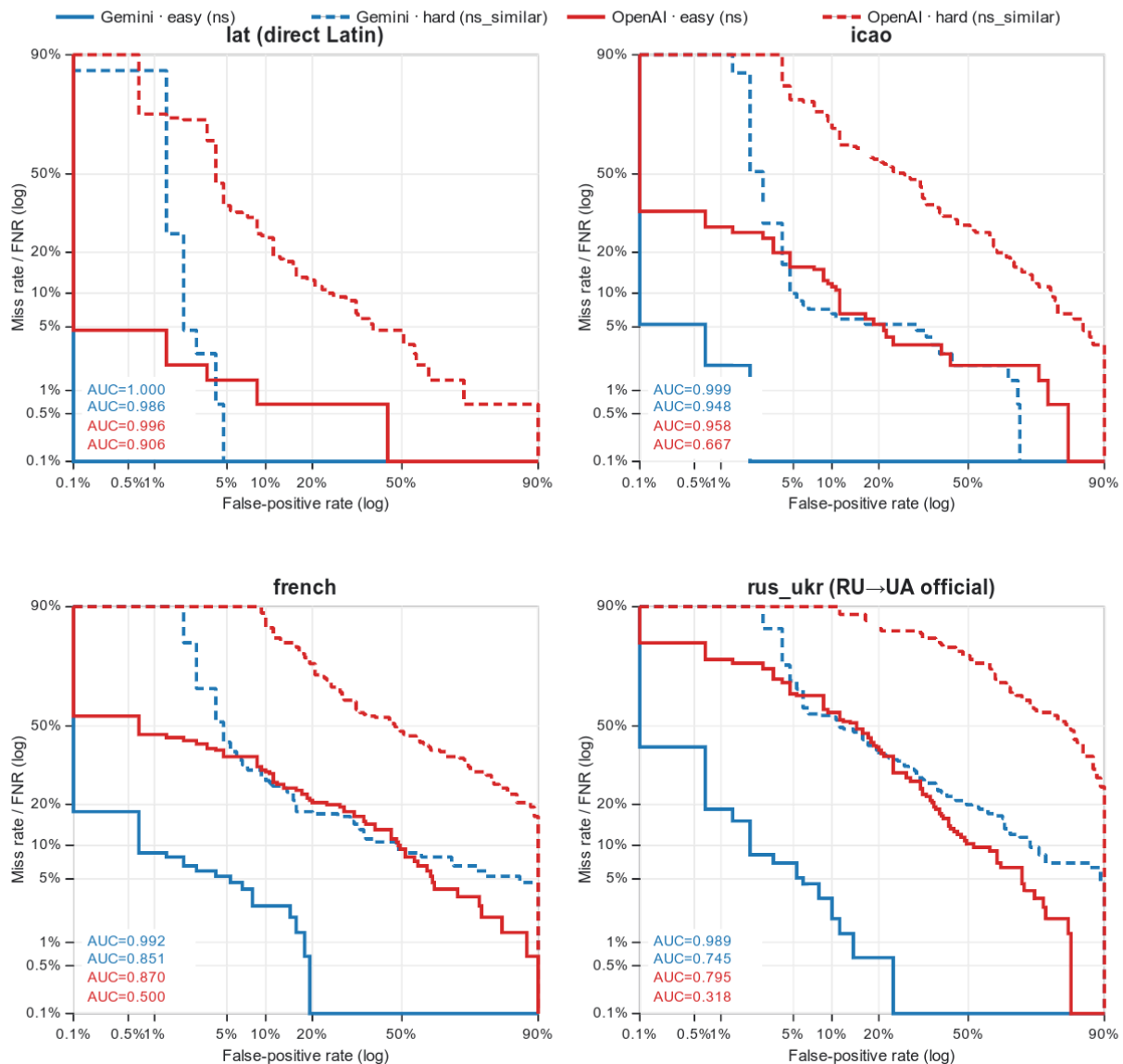


Fig. 2. DET curves (miss rate against false-positive rate on normal-deviate axes) for the same configuration. The low-error corner stays empty for every transliterated-positive/hard-negative pairing: no threshold makes both errors small.

Table 6

Mean top-1 score, positives (+) versus negatives (–), 3072 dimensions, two-part queries

	lat ⁺	icao ⁺	french ⁺	rus_ukr ⁺	ns [–]	ns_similar [–]	ns_similar_nb [–]
Gemini	0.982	0.970	0.958	0.950	0.900	0.938	0.920
OpenAI	0.783	0.720	0.682	0.651	0.583	0.686	0.642

Operating points and realistic prevalence. Holding recall high is cheap against easy negatives and prohibitive against hard look-alikes (Table 8), and those in-sample figures are optimistic because the threshold is swept on the same scores it is scored against. Fixing it on a random calibration half and measuring on the other half (Table 9) preserves that conclusion and adds another: at $n \approx 150$ the recall floor itself does not transfer — Gemini on lat, calibrated to a 95% floor, delivered 89.3% — and the tight-FPR point is pinned by a single high-scoring look-alike. Figure 3 shows the picture is stable across specificity conditions.

Table 7

AUC of top-1 score separability, 3072 dimensions, two-part queries, with 95% entity-cluster bootstrap intervals

Positives	Model	vs ns (easy)	vs ns_similar (hard)	vs ns_similar_nb (hard)
lat	Gemini	1.000 [1.000, 1.000]	0.986 [0.966, 0.999]	0.996 [0.989, 1.000]
lat	OpenAI	0.996 [0.988, 1.000]	0.906 [0.866, 0.941]	0.962 [0.941, 0.980]
icao	Gemini	0.999 [0.998, 1.000]	0.948 [0.918, 0.974]	0.983 [0.968, 0.995]
icao	OpenAI	0.958 [0.934, 0.979]	0.667 [0.606, 0.727]	0.825 [0.777, 0.870]
french	Gemini	0.992 [0.985, 0.997]	0.851 [0.804, 0.897]	0.942 [0.913, 0.968]
french	OpenAI	0.870 [0.827, 0.907]	0.500 [0.432, 0.567]	0.673 [0.614, 0.735]
rus_ukr	Gemini	0.989 [0.981, 0.996]	0.745 [0.687, 0.800]	0.912 [0.880, 0.942]
rus_ukr	OpenAI	0.795 [0.745, 0.843]	0.318 [0.251, 0.385]	0.506 [0.440, 0.571]

ROC stability — two-part vs three-part queries

dim 3072 · hardest negative (ns_similar) · same 150 entities under two-part and three-part query forms

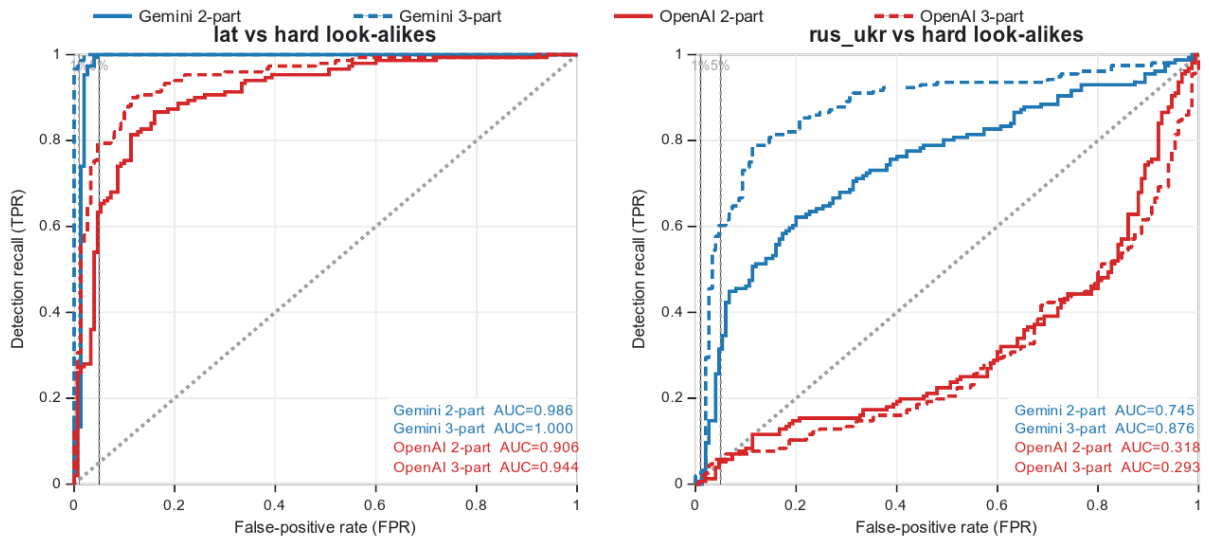


Fig. 3. The same view for lat and rus_ukr against hard look-alikes, with the two-part and three-part conditions overlaid. The near-coincident curves show stability across the specificity factor; because both conditions cover the same 150 entities (see Query design), this is stability, not independent replication.

Table 8

False-positive rate (%) incurred to hold detection recall $\geq 95\%$, 3072 dimensions, threshold swept in-sample (easy = ns, hard = ns_similar)

Positives	Model	easy, 2-part	easy, 3-part	hard, 2-part	hard, 3-part
lat	Gemini	0.0	0.0	2.0	0.0
lat	OpenAI	0.0	0.0	39.3	22.0
icao	Gemini	0.7	0.0	31.3	3.3
icao	OpenAI	21.3	20.0	86.0	86.0
french	Gemini	5.3	0.7	86.7	39.3
french	OpenAI	61.3	74.7	96.0	100.0
rus_ukr	Gemini	6.0	2.7	89.3	72.7
rus_ukr	OpenAI	69.3	70.0	96.7	98.7

Both sets are balanced by construction, whereas designees are rare in real traffic: for Gemini on rus_ukr at a 95% recall floor, one million screened names at a prevalence of one designee per 10,000 give $\approx 60,090$ alerts at $\approx 0.16\%$ precision under the easy anchor (FPR 6.0%) and $\approx 893,010$ at $\approx 0.01\%$ under the hard one (FPR 89.3%). Neither is a forecast; the width between them is the point, and even the easy anchor sends about 6% of screened names to review, which is why Implications for architecture and model choice treats the score as a queue-ordering signal.

Table 9

Held-out operating points, 3072 dimensions, two-part queries: threshold fixed on a random calibration half at a 95% recall floor, recall and FPR measured on the other half ($n = 75-78$).

Positives	Model	Threshold	Recall, easy	FPR, easy	Recall, hard	FPR, hard
lat	Gemini	0.974	89.3	0.0	90.7	2.7
lat	OpenAI	0.686	97.3	0.0	98.7	42.7
icao	Gemini	0.942	94.7	0.0	97.4	38.7
icao	OpenAI	0.620	97.4	26.7	97.4	94.7
french	Gemini	0.915	100.0	17.3	96.0	88.0
french	OpenAI	0.548	98.7	78.7	92.1	96.0
rus_ukr	Gemini	0.925	96.2	5.3	98.7	94.7
rus_ukr	OpenAI	0.559	94.9	62.7	96.2	98.7

Ensembling at an equal budget, and the rank tail. One would naturally *union* the two models' candidate lists: on rus_ukr two-part that union reaches 99.36% (155/156) against Gemini's 98.72% and OpenAI's 73.08%, but at a mean of 7.86 candidates per query, so the comparison charges the ensemble nothing for the extra review load. At an equal budget the gain disappears — Gemini alone at $k = 10$ also recovers 155/156 (Table 10) — and both models miss the same entities, HALAVANAU (9761) and SHOIGU (35173).

Table 10

Recall@5 (%) at 3072 dimensions, two-part queries: per model, unioned, and the equal-budget alternative

Regime	Gemini @5	OpenAI @5	Union of both @5	Missed by both	Gemini @10
lat	100.00	100.00	100.00	0	—
icao	98.68	97.35	99.34	1	—
french	98.68	92.72	98.68	2	—
rus_ukr	98.72	73.08	99.36	1	99.36

The worst rank matters more than the average. The rus_ukr two-part Gemini distribution is extremely concentrated (median 1, $p_{95} = 1$, $p_{99} = 4$, maximum among retrieved = 6), yet one entity sits far outside it (Table 11): going from depth 10 to 20 recovers nobody, and the last entity is absent from the top 40 and first appears at rank 60 in the top 80. The tail is disconnected from the body of the distribution, and no k below 60 would have found it.

That entity is Sergei Shoigu, queried as Shoihu Serhii (entity 35173). The depth-80 score band is only 0.049 wide (0.9214 ... 0.8728) with SHOIGU mid-pack at 0.8753; 45 of the 59 entities ranked above him (76%) share a Sergei/Serhii/Siarhei given name, while Shoihu $\approx$ SHOIGU — one Ukrainian-*h* / Russian-*g* difference — is the only SHOIGU-like surname in the top 80. This single case shows the failure is possible in this configuration, not how often it occurs (see Conclusions).

Table 11

rus_ukr two-part, Gemini 3072: recall against depth ($n = 156$; full runs at $k = 10$ and $k = 20$, the last row combining the $k = 20$ run with a single-query probe).

k	Recall@ k (%)	New catches versus previous k
1	94.23	—
2	96.15	+3
3	97.44	+2
5	98.72	+2
10	99.36	+1
20	99.36	+0
60	(100.00)	SHOIGU first appears at rank 60 (single-query probe)

Discussion

Measuring the right thing. Recall@1 and MRR reward shallow ranking; under a do-not-miss objective the safety-relevant quantities are candidate recall at the operational depth and the rank tail, especially its p_{99} and maximum. "Recall@5 = 98.7%" and "true-entity rank = 60 for one query" describe the same system, and only the second reveals that a name can be missed — so seeing what production would miss requires retrieving deeper than the production k . No finite benchmark can establish a universal absence of misses: where a guarantee is achievable it is architectural, such as a deterministic normalised-exact match never dropped downstream, rather than statistical.

A plausible mechanism. The failure is not a large character edit — Shoihu is one character from Shoigu — but is consistent with a short name string weighting the frequent given name too heavily and the distinctive surname too lightly. Because the case study does not expose the model's token weights, surname dilution is a hypothesis, not a demonstrated mechanism; cross-script normalisation [20] and structure-guided matching [29] motivate the representation-level experiment in Conclusions. Depth is no remedy: between $k = 10$ and $k = 60$ the model returns six times as many candidates and recovers nobody until the last position.

Implications for architecture and model choice. The results support a hybrid, recall-additive architecture in which embedding similarity is one channel among several and never the accept/reject decision:

transliteration normalisation upstream of indexing and querying (ICAO/BGN/ISO-9 cross-walks, Cyrillic-origin phonetic folding), so that comparable orthographies meet;

a deterministic surname/phonetic channel (normalised-exact, edit distance, Slavic-aware phonetic keys) unioned with the semantic candidates, plus a sparse blocker of the Sparkly class [8];

alias/AKA expansion on the list side, absent here by design (see Data and Experimental Setup), and multi-transliteration query expansion on the input side;

scores that order the review queue, never auto-clear; a later stage may add or re-rank candidates, never silently drop one.

That architecture is proposed, not evaluated here; quantifying how much of the rank tail a surname/phonetic channel recovers, and at what false-alert cost, is the main item of future work. Within the embedding channel we retain Gemini at 3072 dimensions, conditioned on this snapshot, these queries and these API versions and on an index built without aliases — not a general ranking of the two services [14]. OpenAI's faster median latency and cleaner negative-score separation are of limited value under a recall-oriented objective, and although the two models retrieve substantially different candidates (top-1 agreement on negatives is only 22.7–27.3% at 3072), Ensembling at an equal budget, and the rank tail shows the union adds nothing that depth in the stronger model does not.

Threats to Validity

Construct and scope. We score the candidate-generation channel in isolation, and the alert-score analysis measures alert-triggering rather than identification correctness (see Metrics and Statistical Treatment); only primary name fields were indexed, so this is not an estimate of production recall for an alias-aware index.

Synthetic queries, clustered observations. The query sets are generated renderings, so difficulty is chosen rather than encountered; one entity moves Recall@1 by ≈ 0.67 pp, observations are clustered by entity and reused across regimes and conditions, and the two specificity conditions share all 150 entities, so nothing here demonstrates replication on a disjoint set.

Negative construction and prevalence. ns_similar(nb) are adversarial constructs and the operational population is not known to lie between them and ns; 150 negatives cannot estimate an operational false-positive rate, in-sample sweeps flatter both recall and FPR, and the held-out split shows the recall floor itself does not transfer at this sample size (see Operating points and realistic prevalence).

Single, partly characterised backend. One HNSW store with recorded build parameters but no measurement of the index's own recall against exact search; absolute scores are model-specific (see Dataset audit and remaining controls; Score distributions and separability), and both services are hosted and mutable under a stable marketing name, tying the results to runs executed 30 May – 27 June 2026.

Censoring, baselines and coverage. Ranks above 5 are censored except in the rus_ukr two-part probes at $k \leq 80$; no string, phonetic, sparse or PLM-based matcher was evaluated alongside the two services; and a single OFAC snapshot and four Slavic schemes leave other scripts and name structures out of scope.

Conclusions

For individual-name sanctions screening, semantic embedding similarity is a strong candidate-generation channel for in-script and typographic variation, but under the tested cross-transliteration conditions it does not support a no-miss objective on its own. Gemini gemini-embedding-001 was the stronger service — by 31–35 pp of paired Recall@1 on Russian-to-Ukrainian forms, with no systematic loss down to 768 dimensions — whereas OpenAI text-embedding-3-large degraded as transliteration distance grew and as dimension fell. Yet no model, dimension, ensemble or feasible depth made the top-1 score a safe accept/reject rule: hard look-alikes are inseparable from transliterated positives by score, ensembling adds nothing at an equal candidate budget, and a high-profile designee sat at rank 60 with a one-character surname difference. The consequence is architectural — embeddings generate candidates, transliteration-normalised and surname-aware deterministic matching carries the coverage argument, and scores order the review queue rather than close it.

Future work, in rough priority order:

1. Build and evaluate the deterministic surname/phonetic and transliteration-normalisation channel, and measure how much rank tail it recovers at what false-alert cost (target: SHOIGU at rank 1).
2. Add strong baselines at an equal candidate budget: a TF/IDF blocker [8], BGE-M3 [16], and a PLM matcher of the Ditto class [30, 31].
3. Characterise the rank tail at large k across all regimes and both models, recording the search-time

ANN parameter and ANN-versus-exact recall.

4. Index aliases, adopt set-valued relevance and re-measure, adding open-set identification metrics (detection-and-identification rate at a false-positive identification rate) and negatives scaled by orders of magnitude so that alerts per million can be reported at stated prevalence.

5. Test transfer and representation: a disjoint entity set, a later OFAC snapshot, OpenSanctions Pairs [26], non-Slavic scripts, and embeddings of normalised or separately embedded name components to test the surname-dilution hypothesis (see A plausible mechanism).

Data Availability

The artefacts accompany this article as a permanently archived deposit [32], developed at <https://github.com/YegorPavlenko/article-2026-embed-ofac>, directories query-sets/ (14 query files and sdn_xml.zip, the OFAC snapshot); embeddings/: results/ (96 raw result files, 13,932 query records) and deep-k-probes/ (the $k = 10$ and $k = 20$ runs behind Table 11 and the probes Shoihu_UKR_{40,80}.json); metrics_recomputed.csv (authoritative point estimates); metrics_entity_ci.csv (Wilson and entity-cluster intervals, Table 4); paired_tests.csv (Tables 4 and 5); auc_ci.csv (Table 7); threshold_roc.csv and operating_points.csv (Table 8) with operating_points_heldout.csv (Table 9); rank_tail.csv (Table 11); latency_percentiles.csv; dataset_audit.csv; and plots/.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, I.P. and V.H.; methodology, I.P.; software, I.P.; validation, I.P.; formal analysis, I.P.; investigation, I.P.; data curation, I.P.; writing—original draft preparation, I.P.; writing—review and editing, V.H.; visualization, I.P.; supervision, V.H. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the authors used Anthropic Claude and OpenAI GPT for translating and data processing code writing. After using these tools, the authors reviewed and edited the content and take full responsibility for the content of this publication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- Kim S., Yang S. Accuracy improvement in financial sanction screening: is natural language processing the solution? *Frontiers in Artificial Intelligence*. 2024. Vol. 7. Art. 1374323. DOI: <https://doi.org/10.3389/frai.2024.1374323>
- Oztas B., Cetinkaya D., Adedoyin F. F., Budka M., Aksu G., Dogan H. Transaction monitoring in anti-money laundering: a qualitative analysis and points of view from industry. *Future Generation Computer Systems*. 2024. Vol. 159. P. 161–171. DOI: <https://doi.org/10.1016/j.future.2024.05.027>
- The Wolfsberg Group. Wolfsberg Guidance on Sanctions Screening. 2019. URL: <https://wolfsberg-group.org/resources/legacy/53>
- Office of Foreign Assets Control, U.S. Department of the Treasury. FAQ 5: How do I determine if I have a valid OFAC match? URL: <https://ofac.treasury.gov/faqs/5>
- Office of Foreign Assets Control, U.S. Department of the Treasury. Sanctions List Search. URL: <https://ofac.treasury.gov/sanctions-list-search-tool>
- Binette O., Steorts R. C. (Almost) all of entity resolution. *Science Advances*. 2022. Vol. 8, No. 12. Art. eabi8021. DOI: <https://doi.org/10.1126/sciadv.abi8021>
- Papadakis G., Skoutas D., Thanos E., Palpanas T. Blocking and filtering techniques for entity resolution: a survey. *ACM Computing Surveys*. 2021. Vol. 53, No. 2. P. 1–42. DOI: <https://doi.org/10.1145/3377455>
- Paulsen D., Govind Y., Doan A. Sparkly: a simple yet surprisingly strong TF/IDF blocker for entity matching. *Proceedings of the VLDB Endowment*. 2023. Vol. 16, No. 6. P. 1507–1519. DOI: <https://doi.org/10.14778/3583140.3583163>
- Zeakis A., Papadakis G., Skoutas D., Koubarakis M. Pre-trained embeddings for entity resolution: an experimental analysis. *Proceedings of the VLDB Endowment*. 2023. Vol. 16, No. 9. P. 2225–2238. DOI: <https://doi.org/10.14778/3598581.3598594>
- Wang R., Zhang Y. Pre-trained language models for entity blocking: a reproducibility study. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2024. P. 8720–8730. DOI: <https://doi.org/10.18653/v1/2024.naacl-long.483>
- Papadakis G., Efthymiou V., Thanos E. et al. An analysis of one-to-one matching algorithms for entity resolution. *The VLDB Journal*. 2023. Vol. 32. P. 1369–1400. DOI: <https://doi.org/10.1007/s00778-023-00791-3>

12. Reimers N., Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019. P. 3982–3992. DOI: <https://doi.org/10.18653/v1/D19-1410>
13. Karpukhin V., Oğuz B., Min S. et al. Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. P. 6769–6781. DOI: <https://doi.org/10.18653/v1/2020.emnlp-main.550>
14. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: massive text embedding benchmark. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023. P. 2014–2037. DOI: <https://doi.org/10.18653/v1/2023.eacl-main.148>
15. Kamaloo E., Zhang X., Ogundepo O., Thakur N., Alfonso-Hermelo D., Rezagholizadeh M., Lin J. Evaluating embedding APIs for information retrieval. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Industry Track. 2023. P. 518–526. DOI: <https://doi.org/10.18653/v1/2023.acl-industry.50>
16. Chen J., Xiao S., Zhang P., Luo K., Lian D., Liu Z. M3-Embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. Findings of the Association for Computational Linguistics: ACL 2024. 2024. P. 2318–2335. DOI: <https://doi.org/10.18653/v1/2024.findings-acl.137>
17. Kusupati A., Bhatt G., Rege A. et al. Matryoshka representation learning. Advances in Neural Information Processing Systems 35. 2022. P. 30233–30249. DOI: <https://doi.org/10.52202/068431-2192>
18. OpenAI Developers. Vector embeddings. URL: <https://developers.openai.com/api/docs/guides/embeddings>
19. Google AI for Developers. Embeddings: Gemini API documentation. URL: <https://ai.google.dev/gemini-api/docs/embeddings>
20. Poulighen B., Steinberger R., Ignat C., Temnikova I., Widiger A. Multilingual person name recognition and transliteration. Corela. 2005. HS-2. DOI: <https://doi.org/10.4000/corela.1219>
21. Kirov C., Johny C., Katanova A., Gutkin A., Roark B. Context-aware transliteration of romanized South Asian languages. Computational Linguistics. 2024. Vol. 50, No. 2. P. 475–534. DOI: https://doi.org/10.1162/coli_a_00510
22. Piskorski J., Marcińczuk M., Yangarber R. Cross-lingual named entity corpus for Slavic languages. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024. P. 4143–4157. URL: <https://aclanthology.org/2024.lrec-main.369/>
23. International Civil Aviation Organization. Doc 9303. Machine readable travel documents. Part 3: Specifications common to all MRTDs. 8th ed. Montréal, 2021. URL: https://www.icao.int/publications/Documents/9303_p3_cons_en.pdf
24. Про впорядкування транслітерації українського алфавіту латиницею: постанова Кабінету Міністрів України від 27.01.2010 № 55. URL: <https://zakon.rada.gov.ua/laws/show/55-2010-%D0%BF>
25. U.S. Board on Geographic Names, Permanent Committee on Geographical Names. Romanization of Russian: BGN/PCGN 1947 system. 2022. URL: https://assets.publishing.service.gov.uk/media/636cd2828fa8f5357a0c652c/ROMANIZATION_OF_RUSSIAN_2022_final.pdf
26. Smith C., Sesodia M., Lindenberg F., Schroeder de Witt C. OpenSanctions Pairs: large-scale entity matching with LLMs: preprint. arXiv. 2026. arXiv:2603.11051. URL: <https://arxiv.org/abs/2603.11051>
27. Office of Foreign Assets Control, U.S. Department of the Treasury. Specially Designated Nationals and Blocked Persons List (SDN): snapshot SDN_XML.ZIP of 14.05.2025, 7,426 individual entries. URL: https://sanctionslistservice.ofac.treas.gov/api/PublicationPreview/exports/SDN_XML.ZIP
28. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947. Vol. 12, No. 2. P. 153–157. DOI: <https://doi.org/10.1007/BF02295996>
29. Chourasia S., Kapoor H., Patil N. Structure-guided entity resolution: fine-tuning LLMs for robust name matching in complex linguistic contexts. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics: Industry Track. 2026. P. 1461–1468. DOI: <https://doi.org/10.18653/v1/2026.acl-industry.101>
30. Li Y., Li J., Suhara Y., Doan A., Tan W.-C. Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment. 2020. Vol. 14, No. 1. P. 50–60. DOI: <https://doi.org/10.14778/3421424.3421431>
31. Li Y., Li J., Suhara Y., Doan A., Tan W.-C. Effective entity matching with transformers. The VLDB Journal. 2023. Vol. 32, No. 6. P. 1215–1235. DOI: <https://doi.org/10.1007/s00778-023-00779-z>
32. Pavlenko I. V., Hnatushenko V. V. Query datasets and run results related to scientific articles about semantic embeddings and RAG for sanctions name screening: dataset. Zenodo, 2026. DOI: <https://doi.org/10.5281/zenodo.22106482>

Єгор ПАВЛЕНКО, Володимир ГНАТУШЕНКО
Національного технічного університету «Дніпровська політехніка»

СЕМАНТИЧНІ ВЕКТОРНІ ПОДАННЯ ЯК КАНАЛ ГЕНЕРАЦІЇ КАНДИДАТІВ ДЛЯ САНКЦІЙНОГО СКРИНІНГУ ІМЕН

Санкційний скринінг імен є задачею ідентифікації у відкритій множині з асиметричною ціною помилок: пропуск особи, унесеної до переліку, має значно тяжчі наслідки, ніж додаткове сповіщення, яке буде перевірено вручну. У роботі оцінено дві сучасні комерційні моделі векторного подання тексту — OpenAI text-embedding-3-large та Google gemini-embedding-001 — як канал генерації кандидатів за фіксованим зрізом переліку фізичних осіб OFAC SDN (7 426 записів; індексовано лише основні поля імені та прізвища, псевдоніми не індексовано). Побудовано контрольований стрес-тест, у якому ті самі 150 підсанкційних осіб подано в чотирьох режимах транслітерації слов'янських імен — пряме латинське написання, романізація за стандартом ICAO, транслітерація за французькою традицією та перетворення російської форми на українську з подальшою офіційною українською романізацією — на двох рівнях повноти імені, проти простих і навмисно складних несанкційних схожих імен, за трьох розмірностей векторів (768–3072). Модель Gemini переважає в кожному нетривіальному режимі, і її перевага зростає зі збільшенням транслітераційної відстані, сягаючи 34,6 відсоткового пункту за показником повноти Recall@1 на двокомпонентних російсько-українських формах (95 % довірчий інтервал 27,1–42,1; точний критерій Мак-Немара $p \approx 1 \times 10^{-16}$); систематичних втрат аж до 768 вимірів не виявлено, тоді як якість моделі OpenAI знижувалася і зі зростанням транслітераційної відстані, і зі зменшенням розмірності вектора. Максимальна косинусна подібність не відокремлює транслітеровані істинні відповідники від складних схожих імен: для OpenAI в найважчому режимі істинні відповідники отримували нижчі оцінки, ніж хибні (площа під ROC-кривою 0,318; довірчий інтервал 0,251–0,385). За однакового бюджету кандидатів об'єднання переліків двох моделей не дало виразу порівняно зі збільшенням глибини пошуку однією сильнішою моделлю (99,36 %), а одну з підсанкційних осіб знайдено лише на 60-й позиції попри відмінність прізвища в один символ. Отже, семантична подібність є корисним каналом генерації кандидатів, але не є надійним правилом прийняття рішення про збіг; одержані висновки стосуються цього зрізу переліку, цих запитів і цих версій програмних інтерфейсів і не є гарантією відсутності пропусків.

Ключові слова: санкційний скринінг; ототожнення сутностей; зіставлення імен; генерація кандидатів; семантичний пошук; векторний пошук; векторні подання тексту; повнота пошуку; фінансовий моніторинг.