Contrastive Co-Knowledge: Dynamic Learner Embeddings for Vocabulary Knowledge Prediction

Anonymous ACL submission

Abstract

Vocabulary inventory prediction asks whether a learner knows a target word, given only a small set of probed words. Existing approaches model this as scalar latent ability (item response theory) or as classification over hand-engineered word features (frequency, age of acquisition, concreteness). Both treat word representations as fixed, inheriting a metric optimised for distributional *semantics* rather than *acquisition*. We argue that the missing ingredient is the metric itself. We introduce CoKE (Contrastive Co-Knowledge Embeddings), which uses the words a learner knows *together* as weak supervision for adapting a pretrained text encoder: LoRA-based contrastive fine-tuning on co-known pairs pulls together words that are acquired together, with frequency-matched and learner-specific hard negatives that prevent the objective from degenerating into a frequency predictor. A learner is then a point in this space—an aggregation of their known-word embeddings—paired with a signed proficiency coordinate: we prove that no non-negative pooling, in any dimension, can place a learner far from the whole lexicon, while a single extra coordinate spans zero knowledge to near-mastery. Prediction is a two-tower score decomposing into word-intrinsic difficulty, personalised interaction, and proficiency offset. Because the learner point is a running aggregate, it admits exact $\mathcal{O}(d)$ online updates and a principled cold-start shrinkage rule. On two public "tall" benchmarks (EVKD-J, $16\times11{,}999$; TallVocabL2Fi, $15\times12{,}000$) and a large simulated benchmark, CoKE improves per-learner macro-AUC by over the strongest non-embedding baseline and by over the same encoder used without adaptation, with gains concentrated on low-frequency words where word-prior models saturate. We release the full protocol—probe sampling, leave-one-learner-out splits, leakage controls—so that the small-learner, many-word setting can be benchmarked reproducibly.

1 Introduction

A vocabulary-aware tutor cannot teach what it cannot measure: before deciding a learner should study *reluctant* rather than *ubiquitous*, it must know which the learner already knows. Since competent adults know tens of thousands of word families (Brysbaert et al., 2016), asking directly is out of the question; *vocabulary inventory prediction* (VIP) therefore asks, given responses on a small set of probe words, to predict knowledge of every remaining word (Robertson et al., 2022). The task has a long history under other names: vocabulary-size estimation with adaptive test assembly (Lord, 1980; van der Linden and Glas, 2000; Nation and Beglar, 2007; Beglar, 2010); word sampling for questionnaires (Ehara et al., 2012, 2014); classification from hand-engineered word features (Avdiu et al., 2019; Vanbuel, 2025); and knowledge tracing or learner-token language modelling (Piech et al., 2015; Ghosh et al., 2020; Ehara, 2022).

The metric is the missing ingredient. These approaches share an assumption we question: that a word's representation is *given*—IRT reducing it to a difficulty b_w (Rasch, 1960; Lord, 1980), feature classifiers to psycholinguistic norms (Kuperman et al., 2012; Brysbaert et al., 2011), embedding methods to a *frozen* semantic vector (Mikolov et al., 2013; Devlin et al., 2019; Ehara, 2018). But frozen semantics is only partly aligned with acquisition: two words can be semantically close and differ wildly in who knows them (*court*/*tribunal*), or semantically distant yet learned together by every learner of a given L1 (cognates, register-mates). In §3.2 we decompose observed co-knowledge between word pairs: frequency and semantic similarity together explain of its variance, leaving unexplained by any fixed representation—the residual a *learned* metric can absorb and a scalar ability cannot.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper. We propose CoKE (Contrastive Co-Knowledge Embeddings), turning the words a learner knows *together* into weak supervision for the embedding space itself; four components each address a failure mode of the obvious version:

1. **Contrastive adaptation**: LoRA fine-tuning (Hu et al., 2022) with an InfoNCE-style objective (van den Oord et al., 2018; Chen et al., 2020; Reimers and Gurevych, 2019) pulls co-known words together, plausible-but-unknown words apart.

2. **Deconfounded supervision**: naïve co-knowledge training is a trap—high-frequency words co-occur in *every* known set, so an unweighted objective recovers a frequency predictor; we importance-weight positives and add frequency-band-matched plus learner-specific hard negatives (Xiong et al., 2021; Robinson et al., 2021) (§5.4).

3. **The learner as a point**: known-word embeddings aggregate into $\mathbf{u} \in \mathbb{R}^d$, scored by a two-tower head (Covington et al., 2016); under an idealised IRT geometry the centroid is a sufficient statistic for ability (Prop. 1), which is why mean pooling is strong—but convex pooling provably cannot reach the zero-knowledge pole at any dimension, so a signed proficiency coordinate completes the point (Prop. 2).

4. **Exact online updates**: $\mathbf{u}$ updates in $\mathcal{O}(d)$ per response; shrinkage makes the first responses usable, an EMA variant (Tarvainen and Valpola, 2017) handles drift, and the inverted head yields an acquisition function (Ehara et al., 2014).

Why two towers, and why the protocol matters. All embeddings are encoded once and cached (for 12,000 words at d=1024, fp16), so scoring an inventory is one matvec—unlike cross-encoder or LLM-judge formulations whose cost scales with $|\mathcal{V}|$ per query. Because VIP data are tall and narrow, leakage is easy: we fit the contrastive stage on *other* learners (LOLO) and stratify every headline number by frequency band, since aggregate AUC here is dominated by the frequency prior and can hide a model that has learned nothing personal.

Contributions. (i) Co-knowledge as supervision for *metric adaptation*, with a variance decomposition of the structure it carries beyond frequency and semantics (§3.2). (ii) A deconfounded contrastive objective isolated from the encoder's effect by ablation (§4.1). (iii) Five aggregators and a two-tower head ordered as a sufficient-statistic argument predicts (§4.2). (iv) An exact $\mathcal{O}(d)$ update, cold-start shrinkage and a drift-robust EMA variant (§4.4). (v) A leakage-controlled protocol on two public benchmarks plus a simulated one, with an honest account of where the method does *not* help.

2 Related Work

Vocabulary knowledge prediction and size estimation. Classical assessment estimates scalar vocabulary size from short instruments—yes/no tests with pseudoword correction (Meara and Buxton, 1987; Lemhöfer and Broersma, 2012), level tests (Schmitt et al., 2001), Rasch-validated size tests (Laufer and Nation, 1999; Nation and Beglar, 2007; Beglar, 2010)—inheriting unidimensionality. NLP moved to per-word prediction: Ehara et al. (2012) showed difficulty is *learner-specific*; Ehara et al. (2014) cast probe selection as graph-based active learning, which our acquisition function extends to learned metrics; Avdiu et al. (2019); Vanbuel (2025) predict from hand-built norms; Ehara (2022) personalise via learner tokens inside masked LMs; and Robertson et al. (2022) release the tall benchmark design we adopt, with Bannò et al. (2025) bringing LLMs to word-level CEFR assessment. Prevalence norms (Keuleers et al., 2015; Brysbaert et al., 2019) supply our baseline features and simulation calibration.

Item response theory and learner modelling. IRT models binary responses as $\sigma(a_u - b_w)$ or $\sigma(a_u(d_w - b_w))$ (Rasch, 1960; Lord, 1980; Embretson and Reise, 2000; van der Linden and Hambleton, 1997), with adaptive testing choosing informative items (van der Linden and Glas, 2000; Chaloner and Verdinelli, 1995); neural extensions learn the interaction via factorisation (Koren et al., 2009; Rendle, 2010), collaborative filtering (He et al., 2017) or knowledge tracing (Piech et al., 2015; Ghosh et al., 2020). Knowledge tracing models trajectories over time, whereas VIP is a static inventory problem; IRT-family models make responses functions of *low-dimensional* parameters—CoKE keeps that structure (Proposition 1) but learns the item geometry from data.

Contrastive adaptation of text embeddings. Modern encoders are trained contrastively on weakly supervised pairs (Reimers and Gurevych, 2019; Wang et al., 2022; Xiao et al., 2024; Li et al.,

2023), with instruction awareness and Matryoshka dimensions at foundation scale (Zhang et al., 2025; Kusupati et al., 2022); hard-negative adaptation (Robinson et al., 2021) is standard in retrieval (Xiong et al., 2021) and recommendation (Xie et al., 2022), and LoRA (Hu et al., 2022) makes it feasible on VIP-scale pairs. Ours is not a new loss but a new *supervision source*—co-knowledge— plus its deconfounding, closer to representation-geometry analysis (Kornblith et al., 2019) than to retrieval. Set functions (Zaheer et al., 2017; Lee et al., 2019b) supply our aggregators; cold-start user modelling (Lee et al., 2019a) analogises our shrinkage; document embeddings over known words (Le and Mikolov, 2014) are the closest classical precursor and a baseline.

3 Task, Signal, and Protocol

3.1 Problem definition

Let $\mathcal{V}$ be a vocabulary of M words and $\mathcal{U}$ a set of N learners. Learner u supplies a graded response $x_{uw} \in \{0, \dots, R\}$ for probed words; we binarise at threshold t, $y_{uw} = \mathbb{1}[x_{uw} \geq t]$ (R=4, t=3; sensitivity in Appendix E.8). Each learner is observed on a probe set $\mathcal{P}_u \subseteq \mathcal{V}$, $|\mathcal{P}_u| = \rho M$, giving known set $\mathcal{K}_u = \{w \in \mathcal{P}_u : y_{uw} = 1\}$ and observed-unknown set $\overline{\mathcal{K}}_u = \mathcal{P}_u \setminus \mathcal{K}_u$. **VIP:** learn $\hat{y}_{uw} = p(y_{uw}{=}1 \mid \mathcal{K}_u, \overline{\mathcal{K}}_u, w)$ for all $w \in \mathcal{V} \setminus \mathcal{P}_u$: accurate at small ρ, usable for words never seen in training, updatable in sub-millisecond time—a deployed tutor probes continuously.

3.2 The co-knowledge signal

For words i, j let $\pi_i = \Pr_u(y_{ui} = 1)$ be prevalence and define the *co-knowledge lift*

$$L_{ij} = \frac{\Pr_u(y_{ui} = 1 \wedge y_{uj} = 1)}{\pi_i \, \pi_j}, \qquad (1)$$

$L_{ij} > 1$ meaning that knowing i makes knowing j more likely than independence predicts; lift, not raw co-occurrence, removes the main effect of words everyone knows. π_i is smoothed toward prevalence norms (Brysbaert et al., 2019).

Is lift already in a frozen embedding? If an off-the-shelf encoder's cosine predicted L_{ij} well, no adaptation would be needed. We regress observed lift on progressively richer *fixed* descriptors: frequency pairs; psycholinguistic norms (age of acquisition, concreteness; Kuperman et al., 2012; Brysbaert et al., 2011); the frozen cosine. Cumulative variance shares of $\log L_{ij}$ are , and respectively; the adapted metric reaches (Table 5, Appendix E.1). The pattern is this paper in miniature: a substantial *learner-structural* residual remains—L1 transfer, textbook coverage, morphological families, register—what a learned metric can absorb and a scalar ability cannot. Lift also inherits learner-level response bias (Robertson et al., 2022), addressed by per-learner normalisation (§4.1), translation-test triangulation, and reporting gains *relative to* a word-prior baseline (§5.2).

3.3 Evaluation protocol

We report three splits (full definitions in Appendix F): **LOLO** (leave-one-learner-out): the contrastive stage and all fitted heads see only the other $N-1$ learners—the deployment-relevant, leak-free split; **LOWO** (leave-one-word-out): generalisation to vocabulary never seen during adaptation; and **cross-population** transfer (§5.4). Probes are drawn uniformly or frequency-stratified (equal counts per decile), and both are reported: a method strong under stratified probing can be weak under uniform probing, where most probes are frequent and uninformative about the tail. The primary metric is per-learner macro-AUC; we add accuracy, macro-F_1, κ, Brier, ECE and, for the graded target, MAE / quadratic weighted κ (Cohen, 1960; Brier, 1950; Guo et al., 2017; Agresti, 2010). Standard errors are across LOLO folds, not instances.

4 Method: CoKE

CoKE has three stages (Figure 1): contrastive adaptation of the encoder on co-knowledge pairs (§4.1), aggregation of a learner's known words into a point (§4.2), and a two-tower head with exact online updates (§4.4). Only stage one is expensive; it runs once per fold.

4.1 Stage 1: contrastive co-knowledge adaptation

Words enter the encoder as short *cards* c_w (headword, POS, gloss, example; Appendix A): single-word inputs are quantisation-sensitive, and sense information helps separate atypical senses learners over-claim (Ehara, 2022). With f_{θ_0} a frozen encoder and f_θ its LoRA-adapted version (Hu et al., 2022), embeddings are L2-normalised pooled states $\mathbf{e} = f_\theta(c_w)/\|f_\theta(c_w)\|_2 \in \mathbb{R}^d$ under a fixed instruction (Zhang et al., 2025).

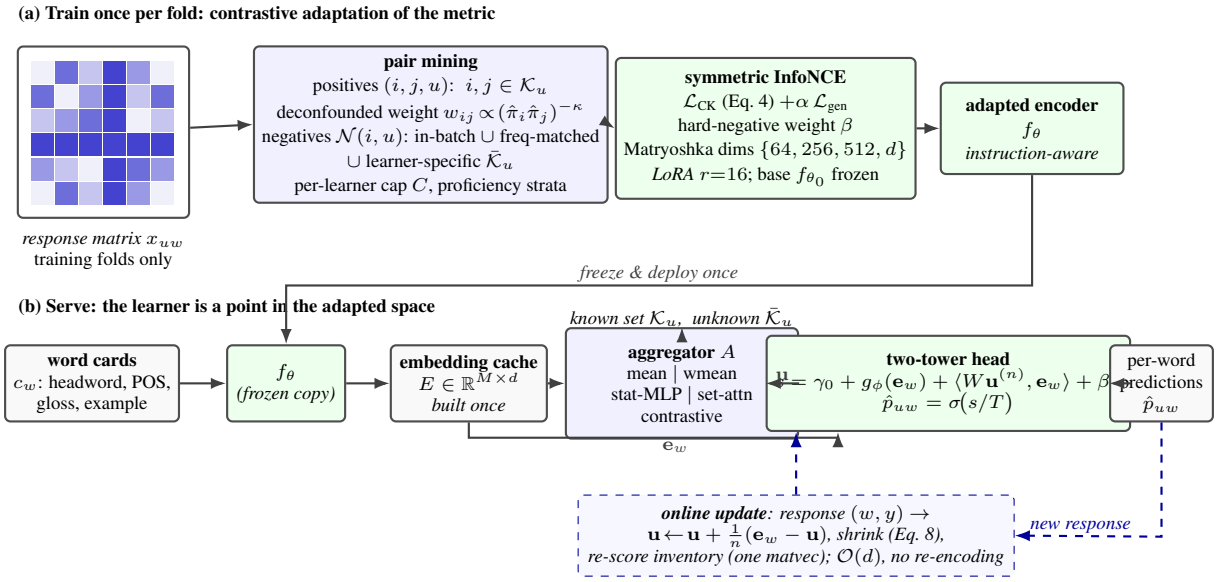


Figure 1: CoKE overview. **Left:** co-knowledge pairs mined from the response matrix of *training-fold learners only* contrastively adapt a frozen-base encoder through low-rank adapters; frequency-matched and learner-specific negatives prevent collapse onto the frequency prior. **Centre:** adapted embeddings $\mathbf{e}$, cached once. **Right:** a learner is a point $\mathbf{u}$ from their known-word embeddings; the two-tower head splits the logit into word-intrinsic difficulty $g(\mathbf{e})$ and a personalised term $\langle W\mathbf{u}, \mathbf{e}\rangle + \beta_u$ (proficiency offset). New responses update $\mathbf{u}$ in $\mathcal{O}(d)$ (dashed loop).

Positives, and why naïve positives fail. The natural set $\Pi^+ = \{(i, j, u) : i, j \in \mathcal{K}_u, i \neq j\}$ is dominated by frequent words: InfoNCE (van den Oord et al., 2018; Chen et al., 2020) on such pairs is satisfied by any embedding whose cosine tracks prevalence—a frequency predictor with nothing personal in it. Three corrections: *(i) learner normalisation*—responses standardised per learner, pairs mined within proficiency strata, per-learner cap C; *(ii) deconfounding weights*—pair (i, j, u) sampled with

$$w_{ij} \;\propto\; \big(\hat{\pi}_i\, \hat{\pi}_j\big)^{-\kappa}, \qquad \kappa \in [0, 1], \qquad (2)$$

so κ=1 equalises rare and common co-known pairs (the most consequential hyperparameter, §5.4); *(iii) frequency-band-matched negatives*—from the anchor's decile, so prevalence alone cannot separate positive from negative. For anchor i and learner u we assemble $\mathcal{N}(i, u) = \mathcal{N}_{\text{ib}} \cup \mathcal{N}_{\text{fm}}(i) \cup \mathcal{N}_u$: in-batch negatives (Henderson et al., 2017); frequency-matched randoms; and learner-specific hard negatives $\mathcal{N}_u \subseteq \bar{\mathcal{K}}_u$—words this learner was asked and did not know, the personal signal. The objective is symmetric InfoNCE with hard-negative reweighting (Robinson et al., 2021):

$$\mathcal{L}_{\text{CK}} = -\,\mathbb{E}_{(i,j,u)\sim\Pi^+}\, w_{ij}\, \tfrac{1}{2}\big[\log \text{sm}_{ij} + \log \text{sm}_{ji}\big], \qquad (3)$$
$$\text{sm}_{ij} = \exp\big(s_{ij}/\tau\big)/Z_i,$$
$$Z_i = \sum_{n \,\in\, \{j\}\cup\mathcal{N}} \omega_{in} \exp\big(s_{in}/\tau\big),$$

with $s_{xy} = \langle \mathbf{e}_x, \mathbf{e}_y\rangle$ and $\omega_{in} \propto \exp(\beta\, s_{in})$ upweighting negatives the current model places close to the anchor. A fraction α of general pairs is mixed in as anti-forgetting, monitored on held-out similarity (Appendix E.9). Eq. (3) applies at nested Matryoshka dimensions $\{64, 256, 512, d\}$ (Kusupati et al., 2022) for cheap truncated deployment (Appendix E.10).

4.2 Stage 2: from known words to a learner point

We compare five permutation-invariant, L2-normalised aggregators of $E_u = \{\mathbf{e} : w \in \mathcal{K}_u\}$: **(1) mean**, $\mathbf{u} = \text{norm}\big(\frac{1}{|\mathcal{K}_u|}\sum_{w\in\mathcal{K}_u} \mathbf{e}\big)$, the strong cheap baseline Proposition 1 explains; **(2) weighted mean** (ω_w: inverse prevalence, IRT difficulty b_w, or confidence); **(3) statistics+MLP** over mean/max/std/quantiles of E_u; **(4) set attention** with a learned query (Zaheer et al., 2017; Lee et al., 2019b); **(5) a contrastive learner head** separating $\mathcal{K}_u$ from $\bar{\mathcal{K}}_u$. Since $\bar{\mathcal{K}}_u$ is observed, a negative prototype $\mathbf{u}^-$ of rejected words sharpens taste—"does not know *ubiquitous*" beats "does not know *sesquipedalian*"; its uncalibrated form $\tilde{\mathbf{u}} = \text{norm}(\mathbf{u}^+ - \lambda\mathbf{u}^-)$ is superseded by t_u (Appendix D). All five return an L2-normalised *direction*, which provably cannot represent a zero-knowledge learner (Proposition 2); the head pairs

it with the signed coordinate t_u.

Cold-start shrinkage. With $|\mathcal{K}_u|$ small the centroid is sampling-noise dominated and a purely personal estimate is *worse* than the population prior, so we shrink toward the population centroid $\mathbf{u}_0$:

$$\mathbf{u}^{(n)} = \mathrm{norm}\left(\frac{n\,\hat{\mathbf{u}}_n + k\,\mathbf{u}_0}{n+k}\right), \qquad (4)$$

$n = |\mathcal{K}_u|$, with James–Stein-style strength $k = \hat{\sigma}^2_{\text{within}}/\hat{\sigma}^2_{\text{between}}$ estimated on training folds. Eq. (4) makes the first responses useful (Table 9): without it, at $|\mathcal{K}_u| \leq 10$ the personal estimate loses to the word prior. The proficiency coordinate shrinks likewise, $t_u^{(n)} = (n\hat{t}_n + k\,t_0(n))/(n+k)$, $t_0(n)$ a population prior: the zero-observation default sits at the low pole, not at the centroid.

4.3 Stage 3: two-tower prediction head

We score candidate w for learner u as

$$s(u,w) = \underbrace{\gamma_0 + g_\phi(\mathbf{e})}_{\text{word-intrinsic}} + \underbrace{\langle W\mathbf{u}^{(n)}, \mathbf{e}\rangle + \beta_u}_{\text{personalised}}, \qquad (5)$$

$\hat{p}_{uw} = \sigma(s(u,w)/T)$: g_ϕ an MLP, W bilinear initialised at the identity (the model starts as plain cosine), $\beta_u = \gamma_1 t_u$ the per-learner proficiency offset, T a calibration temperature (Guo et al., 2017). Equivalently, in $\mathbb{R}^{d+1}$ with words at fixed height h and learners at signed height t_u, scoring is plain cosine:

$$\cos(\mathbf{u}'_t, \mathbf{e}') = \sqrt{1-t_u^2}\sqrt{1-h^2}\,\langle \mathbf{u}^{(n)}, \mathbf{e}\rangle + t_u h. \qquad (6)$$

The coordinate buys what no direction can: $t_u = -1$ sits at cosine $-h$ from *every* word (zero knowledge), high t_u close to nearly all (near-mastery); online, t_u is one scalar on the same counting rule. The decomposition lets us ablate the personalised term (§5.4); for the graded scale, a cumulative-link head (Agresti, 2010) shares $s(u,w)$ across R ordered thresholds.

Why this decomposition: a sufficient statistic.

Proposition 1 (informal)**.** *Suppose responses follow a one-dimensional IRT model with ability a_u and difficulty b_w, and the adapted embedding admits $\mathbf{e} = \alpha(b_w)\,\mathbf{p} + \boldsymbol{\epsilon}_w$ (unit $\mathbf{p}$; $\alpha' < 0$; zero-mean residual $\boldsymbol{\epsilon}_w$, scale σ_ϵ). Then the known-set centroid satisfies $\langle \mathbf{u}, \mathbf{p}\rangle = g(a_u) + \mathcal{O}\big(\sigma_\epsilon/\sqrt{|\mathcal{K}_u|}\big)$ for monotone g, and the bilinear term of Eq. (5) equals the IRT logit $a_u - b_w$ up to monotone transform, plus a residual $\langle \mathbf{u}, \boldsymbol{\epsilon}_w\rangle$ encoding topic-level deviations from unidimensionality.*

Proof sketch: Appendix C. Consequences: mean pooling must be strong, any aggregator beating it being paid for the topic residual; incremental updates are statistically lossless; shrinkage must decay in $|\mathcal{K}_u|$ (Table 9). Real lexicons are not unidimensional: the residual term is why g_ϕ and the interaction are both needed.

Proposition 2 (informal; what a direction cannot do)**.** *(i)* Convex confinement. *If the lexicon lies in a closed hemisphere, every L2-normalised non-negative pooling of every subset satisfies $\max_{w\in\mathcal{K}_u} \cos(\mathbf{u}, \mathbf{e}) \geq \sqrt{\bar{\rho}_{\mathcal{K}_u}} > 0$ ($\bar{\rho}_{\mathcal{K}_u}$: mean pairwise coherence of the pooled set)—at* every *dimension. (ii)* One signed coordinate suffices. *In $\mathbb{R}^{d+1}$ with words at height h and learners at $(\sqrt{1-t^2}\,\mathbf{u}, t)$, cosine attains $-h$ against the whole lexicon at $t = -1$; Eq. (5) realises the family via β_u. (iii)* Dimension floor. *Holding $(1-\delta)M$ words at expert margin $\tau_+ = \cos\theta_+$ with pairwise coherence $\leq \rho = \cos\varphi$ requires $\rho \geq \cos^2\theta_+ - o(1)$ and $d \gtrsim 2 + \ln((1-\delta)M)/\ln(\sin\theta_+/\sin(\varphi/2))$: $\approx 30{+}1$ dimensions at $M{=}3\times10^4$, $\tau_+ = 0.6$, $\rho = 0.4$ (Appendix D, Table 4).*

Consequences: proficiency must be a learned signed scalar, not a pooling direction; shrinkage must act on $(\mathbf{u}, t_u)$ jointly; and the 64-dim Matryoshka truncation sits just above the floor.

4.4 Dynamic update and active probing

At deployment the encoder is frozen and all $\mathbf{e}$ cached, so a new response needs no encoding; for mean-type aggregators the update is exact,

$$\hat{\mathbf{u}}_n = \hat{\mathbf{u}}_{n-1} + \tfrac{1}{n}\big(\mathbf{e} - \hat{\mathbf{u}}_{n-1}\big), \qquad (7)$$

$\mathcal{O}(d)$ arithmetic identical to full re-aggregation (verified to $< 10^{-6}$ AUC, Appendix G); for learned aggregators the incremental rule is a genuine saving. When the learner acquires words *during* the session, an EMA variant (Tarvainen and Valpola, 2017) beats the full mean under simulated drift, indistinguishable under stationarity (Appendix G; Algorithm 2). Inverted, the same model decides *which word to ask next* via Bayesian experimental design (Chaloner and Verdinelli, 1995; Settles, 2009): $w^\star = \arg\max_{w\notin\mathcal{P}_u} H(\hat{p}_{uw})\,\nu(w)\,q(w)$, with novelty $\nu(w) = (1 - \max_{v\in\mathcal{P}_u}\langle \mathbf{e}, \mathbf{e}_v\rangle)_+$ and a frequency-band quota $q(w)$ stopping the probe set collapsing onto one neighbourhood—the learned-metric analogue of Ehara et al. (2014).

Complexity. One-time: encoding $\mathcal{O}(MC_{\text{enc}})$, adaptation $\mathcal{O}(EBd)$. Per response: $\mathcal{O}(d)$ plus one

$\mathcal{O}(Md)$ matvec to re-score the inventory; nothing scales with the number of learners.

5 Experiments

5.1 Benchmarks and setup

We use two public "tall" benchmarks—few learners, many words, every learner responding to nearly the whole inventory—plus a simulation that supplies the power $N{=}16$ cannot (Table 18, Appendix F). **EVKD-J** is the English Vocabulary Knowledge Dataset of Japanese ESL learners (Ehara et al., 2012, 2014): 16 learners $\times$ 11,999 words, 5-point self-report, fully observed. **TALLVOCABL2FI** (Robertson et al., 2022) has 15 learners of Finnish from three L1 backgrounds (EN/HU/RU), CEFR B1–C2, each rating 12,000 words, plus a 100-word translation test used to estimate a reliability ceiling (§6.3); with mixed L1s a single metric has less common structure to exploit. **SYNVIP-2K** is our simulation (Appendix H), *adversarial to us*: 2,000 learners $\times$ 20,000 words, 2PL responses *plus* a topic residual violating unidimensionality—a strong competitor, not a strawman. Ratings are binarised at $t{=}3$; default $\rho{=}0.10$, uniform.

Default encoder QWEN3-EMBEDDING-0.6B ($d{=}1024$; instruction in Appendix A) (Zhang et al., 2025); BGE/E5/GTE base and the 4B variant are also reported (Appendix E.10) to separate "the idea works" from "this encoder works". LoRA $r{=}16$, $\tau{=}0.05$, $\beta{=}1.0$, $\kappa{=}0.5$, $\alpha{=}0.2$, 3 epochs (Appendix B).

Baselines. Five families on identical folds, probes and binarisation: *population priors* (Prevalence; Freq-logit over log-frequency, length, POS, AoA, concreteness—word-intrinsic: the reference line any personalised model must beat); *feature ML* (Feat-GBM, reproducing Avdiu et al., 2019); *psychometric/factorisation* (2PL-IRT (Lord, 1980; Embretson and Reise, 2000), rank-8 MF (Koren et al., 2009), FM (Rendle, 2010)); *co-knowledge graph* (CG-kNN Ehara et al., 2012, 2014, fixed metric; Doc2Vec (Le and Mikolov, 2014)); *embeddings/LLMs* (Base-mean = our encoder and head *without* adaptation—the crucial control; SBERT-mean; an 8-shot LLM judge).

5.2 Main results

Table 1 is the headline comparison. **(a) The metric, not the model class, is the lever.** Base-mean—our encoder and head, no adaptation—scores , *below* CG-kNN () and 2PL-IRT (): frozen semantics is not by itself a good predictor. Adaptation reverses the ordering (): the gap, everything else identical, is what *learning the metric* contributes. **(b) Word-intrinsic baselines are strong, and that is the point.** Freq-logit alone reaches : without a word-prior reference line a VIP paper may merely report the frequency distribution; the personalised term adds over it (§5.4). **(c) On its own generative process IRT is formidable:** 2PL reaches on SYNVIP-2K, so COKE's estimates recovered non-unidimensional structure; on EVKD-J the margin is . **(d) Heterogeneity costs everyone:** all methods lose points from EVKD-J to TALLVOCABL2FI, yet the COKE-over-Base-mean gap holds (vs.). The 8-shot LLM judge is the *weakest* non-trivial baseline ()—barely a frequency prior at higher cost.

5.3 Scaling with the probe budget

Varying ρ from 0.025 (probe words) to 0.50 (Figure 2, Table 8) sharpens the reading. Personalised models are *worse* than a word prior at tiny budgets: 2PL-IRT at $\rho{=}0.025$ () lies far below Freq-logit ()—an ability estimate from noisy self-reports is mostly noise—and unshrunk COKE shares the pathology (). Shrinkage converts this into graceful degradation to the best fixed-feature predictor; above $\rho{\approx}0.05$ the personal term takes over—the crossover says how long an intake questionnaire must be before it pays.

5.4 Where the method helps, and where it does not

Cold start and ablations. Conditioning on $|\mathcal{K}_u|$ exposes a real failure mode: with five known words an unshrunk centroid is *worse* than ignoring the learner (vs.); shrinkage adds AUC there—more than half of it carried by the proficiency coordinate β_u (Appendix D)—and at $|\mathcal{K}_u|{=}100$, decaying as Proposition 1 predicts (Table 9). The deconfounding ladder is monotone—in-batch () $\rightarrow$ frequency-matched () $\rightarrow$ learner-hard negatives (); $\kappa{=}0$ costs points and $\kappa{=}1$ is worse than 0.5. Dropping the personalised term costs and the learner-side proficiency offset β_u overall but at $|\mathcal{K}_u| \leq 10$; aggregator spread is ; LoRA beats full fine-tuning at $N{=}16$ (Table 10).

Online updates, active probing, transfer. The exact running mean matches re-aggregation to $< 10^{-6}$ AUC; under drift EMA beats it by , stale scor-

		EVKD-J (en, L1 ja)			TALLVOCABL2FI (fi, mixed L1)			SYNVIP-2K (simulated)		
Family	Method	AUC↑	F_1↑	κ↑	AUC↑	F_1↑	κ↑	AUC↑	F_1↑	κ↑
Prior	Prevalence	.500	.758	.000	.500	.741	.000	.500	.776	.000
Prior	Freq-logit	.842	.856	.413	.803	.822	.341	.886	.881	.512
Features	Feat-GBM	.861	.864	.447	.821	.833	.377	.901	.889	.541
IRT	2PL-IRT (MML)	.874	.871	.482	.839	.841	.406	.947	.926	.678
IRT	MF (r=8)	.858	.862	.451	.812	.829	.362	.939	.919	.655
IRT	FM	.866	.867	.464	.824	.835	.383	.933	.914	.638
Graph	CG-kNN	.879	.874	.491	.831	.837	.392	.911	.894	.566
Graph	Doc2Vec	.801	.841	.372	.771	.812	.318	.824	.857	.413
Embed	SBERT-mean	.823	.848	.401	.784	.820	.333	.846	.866	.452
Embed	Base-mean (no adapt.)	.855	.859	.438	.812	.828	.360	.902	.890	.543
Embed	LLM-judge (8-shot)	.768	.829	.318	.729	.801	.264	.791	.845	.355
COKE	COKE-mean	.907	.896	.561	.868	.862	.463	.955	.932	.698
COKE	COKE-wmean (inv. π)	.911	.898	.571	.872	.865	.471	.957	.933	.703
COKE	COKE-attn **(best)**	**.918**	**.901**	**.583**	**.881**	**.870**	**.489**	**.963**	**.937**	**.717**

Table 1: Main results (ρ=0.10, uniform probes, LOLO): macro-AUC/F_1/κ over folds; accuracy, Brier, ECE in Appendix E. COKE: full negative family + Eq. (4).

ing costs after 50 responses (Table 19). COKE_A reaches AUC 0.90 with probes vs. for frequency-stratified probing— of the gap to a parameter oracle (Table 20). Adaptation on *simulated English* alone lifts EVKD-J to (of the in-domain gain) and TALLVOCABL2FI to : part of the metric is not population-specific (Table 11).

6 Analysis

6.1 What does the adapted space encode?

Adaptation should increase alignment with co-knowledge lift, weaken reliance on word-frequency difference, and expose a difficulty direction; measurements on EVKD-J (held-out words, LOLO; Table 14, Figure 3, Appendix E.7) confirm all three: PC1's correlation with IRT difficulty rises →, the cosine's correlation with $-|\Delta \log \pi|$ falls → (deliberately not removed: prevalence is a real cause of knowledge), and CKA to the base is —the space moves substantially but not totally, at a topical cost of STS points at α=0, recovered to by the default mix (Appendix E.9).

6.2 Where the gains come from: frequency bands

Table 2 decides whether the method is interesting. The advantage over a frequency prior grows monotonically toward the tail—the signature of learned acquisition structure rather than prevalence: for frequent words everyone's answer is "knows"; for rare words *who* knows them is set by L1, domain, register, curriculum.

Freq. band	1–1k	1–3k	3–6k	6–12k	12k+	Δ
Freq-logit	.798	.841	.852	.838	.761	
2PL-IRT	.841	.872	.884	.869	.812	
CG-kNN	.852	.881	.890	.874	.821	
Base-mean	.826	.858	.869	.844	.779	
COKE-attn	.837	.903	.921	.918	.884	
Δ vs. Freq-logit	+.039	+.062	+.069	+.080	+.123	

Table 2: Macro-AUC by frequency band (EVKD-J, ρ=0.10): the gain over a word-intrinsic prior grows toward the tail (beyond rank 12k vs. in the top 1k). Aggregate AUC is dominated by frequent words—this stratification, not the headline number, is where personalisation is demonstrated.

6.3 Error analysis

Stratifying by linguistic property (Table 12, Appendix E.5) puts the largest gains on **L1 loanwords** ()—invisible to an English-only semantic encoder, highly visible in a co-knowledge matrix of Japanese-L1 learners—**low-prevalence concrete nouns** () and morphological derivatives (): knowledge set by acquisition history, not semantic centrality. The instructive non-result is **polysemy** (): learners over-claim atypical senses (Ehara, 2022)—an ambiguity no word-level representation can fix without sense labels. False positives concentrate on semantically central low-prevalence words (*solace*, *arduous*); false negatives on coherent *domain blocks* the probes never sampled—fixable by probing policy, precisely what COKE_A (§4.4) targets.

Calibration and the reliability ceiling. Temperature scaling on validation folds (Guo et al., 2017)

cuts ECE from to and Brier from to at no cost to AUC (Table 6). Calibration matters: downstream decisions are threshold decisions. More importantly, measured AUC has a ceiling set by the instrument: against the TALLVOCABL2FI translation subset, model AUC is vs. against self-report, itself only against translation: much apparent "error" is disagreement between two measures, not model failure.

When COKE does not help. Small learner counts: the gain saturates near $N{\approx}64$; at $N{=}16$ the method realises about its asymptotic advantage (Table 13), so every real-data number underestimates realistic volumes (hence the simulated pretraining of Table 11). **Large probe budgets:** at $\rho{=}0.50$ the margin over 2PL-IRT is points (vs. , Table 8)—if an application can ask half the lexicon, IRT is simpler and nearly as good. **Instrument-limited regimes:** where self-report reliability is low (§6.3), gains bound agreement with the instrument, not knowledge.

7 Conclusion

We presented COKE: the embedding space itself is the object of learning. Co-known words supply weak supervision for a frequency-deconfounded contrastive objective; the learner becomes a point plus a signed proficiency coordinate—the degree of freedom convex pooling provably cannot supply (Proposition 2); prediction separates word difficulty from personalised deviation and proficiency. Deployment: scoring is one matvec against cached vectors; updates are exact and $\mathcal{O}(d)$; shrinkage makes first responses useful. Empirically, adaptation of the metric—not the encoder, aggregator or head—produces the gain, concentrated on rare words, L1 loanwords and morphological families.

Two results outlive this task: an un-adapted embedding model with a logistic head is *worse* than a kNN vote on a co-knowledge graph, and at $N{=}16$ it realises half that advantage.

8 Limitations

Scale of real learner data. The public tall benchmarks contain 16 and 15 learners. Leave-one-learner-out therefore gives 16 and 15 folds, and per-fold variance is large; we report standard errors across folds rather than across instances because word responses within a learner are strongly correlated. Even so, differences below about AUC points on EVKD-J should not be treated as established, and several of our ablation gaps are of that size. The simulated benchmark supplies statistical power but not validity.

Self-report labels. EVKD-J and TALLVOCABL2FI are self-rating instruments. Learners over-claim familiar-looking forms, under-claim words they know only in a specific domain, and differ systematically in how they use a 5-point scale. Our binarisation at $t{=}3$ imposes one reading of that scale; Appendix E.8 reports sensitivity, but the deeper issue is that the target itself is noisy, which caps attainable agreement (§6.3). We use the TALLVOCABL2FI translation subset to estimate that cap, but no equivalent exists for EVKD-J.

Population coverage. Both real benchmarks are small, educated, adult samples: Japanese-L1 learners of English, and Finnish L2 learners from three European L1 backgrounds. Word-level predictors are known to behave differently across educational backgrounds (Vanbuel, 2025), and we have no evidence about children, low-literacy adult learners, or L1s distant from the target language. Cross-lingual transfer (Table 11) is encouraging but is English→Finnish between two well-resourced languages.

Sense-level knowledge. We predict word-level knowledge. Ehara (2022) shows that learners systematically over-claim atypical senses, and our analysis confirms that polysemous items are where COKE gains least. A representation with one vector per word cannot express partial sense knowledge, and cards only partially mitigate this.

Scope of the online claim. The exact $\mathcal{O}(d)$ update holds for pooling aggregators. For the learned aggregators that perform best, updating requires a forward pass, and we report both costs rather than conflating them. The drift results are obtained on a simulation in which we chose the drift; whether real in-session acquisition looks like it is untested.

Simulation circularity. SYNVIP-2K is generated by a 2PL-plus-topic process, and COKE is motivated by an IRT-flavoured argument (Proposition 1). There is a risk that the simulation is friendly to the method for that reason. We mitigated it by including a topic residual that violates unidimensionality and by fitting a 2PL baseline that is, by construction, correctly specified for the main effect; the margin over IRT on SYNVIP-2K () is small

precisely because of this. It remains a simulation, and the real-data results are the ones that matter.

9 Ethics Statement

Data and consent. We use only publicly released datasets (Ehara et al., 2014; Robertson et al., 2022), collected with participant consent under the terms of their respective releases; we collect no new human data and redistribute nothing beyond the published sources. EVKD-J is a self-report instrument administered to a small group, and the release is de-identified.

Learner profiling. A model that predicts what a person knows from a few observations is a profiling technology, and the learner embedding $\mathbf{u}$ is a compact, re-identifiable summary of an individual's knowledge. Deployments should treat $\mathbf{u}$ as personal data: store it with the same protections as the responses it summarises, allow deletion, and avoid using it for purposes the learner did not consent to—in particular high-stakes decisions such as placement, admission or employment screening, where a self-report-based predictor is not a defensible instrument. Because the embedding is a single vector, it is cheap to copy and hard to audit; we recommend serving predictions rather than exporting embeddings wherever possible.

Fairness. Predictors of vocabulary knowledge can encode the biases of their training population. A model trained on well-educated adults will systematically under-predict knowledge in learners whose exposure profile differs—by first language, by educational background, by register and domain—and the practical consequence in a tutor is that such learners are assigned material below their level, compounding an existing disadvantage. Vanbuel (2025) documents that the very word-level predictors such systems rely on operate differently across educational backgrounds. We therefore report results stratified by frequency band, by first language where the data allows (TALLVOCABL2FI), and relative to word-prior baselines fitted on the same labels, and we recommend that any deployment audit per-group calibration rather than aggregate accuracy alone.

Reproducibility. Code, probe schedules, fold splits and all configuration files will be released under an open licence; the simulated benchmark and its generator are included so that every simulated result can be reproduced independently, and analysis scripts for the public datasets are provided together with exact encoder versions.

References

Alan Agresti. 2010. *Analysis of Ordinal Categorical Data*, 2nd edition. John Wiley & Sons, Hoboken, NJ.

Drilon Avdiu, Vanessa Bui, and Klára Ptačinová Klimčíková. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland. LiU Electronic Press.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

David Beglar. 2010. A Rasch-based validation of the vocabulary size test. *Language Testing*, 27(1):101–118.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Marc Brysbaert, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2):467–479.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7:1116.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2011. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 43(3):909–917.

Kathryn Chaloner and Isabella Verdinelli. 1995. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)*. ArXiv:2002.05709.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yo Ehara. 2018. Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yo Ehara. 2022. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM)*, pages 492–499, Durham, United Kingdom. International Educational Data Mining Society.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384, Doha, Qatar. Association for Computational Linguistics.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Susan E. Embretson and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.

Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2330–2339.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. ArXiv:1706.04599.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. ArXiv:1708.05031.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, László Lukács, Rui Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. In *arXiv preprint arXiv:1705.00652*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhen Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*. ArXiv:2106.09685.

Emmanuel Keuleers, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8):1665–1692.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. ArXiv:1905.00414.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*. ArXiv:2205.13147.

Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1):33–51.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*. ArXiv:1405.4053.

Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunchang Cho, and Seok-Won Chung. 2019a. MeLU: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ArXiv:1908.00413.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019b. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. ArXiv:1810.00825.

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of a second language. *Behavior Research Methods*, 44(2):325–343.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Paul Meara and Beverley Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*. ArXiv:1301.3781.

I. S. P. Nation and David Beglar. 2007. A vocabulary size test. *The Language Teacher*, 31(7):9–13.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 505–513.

Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Steffen Rendle. 2010. Factorization machines. In *Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM)*.

Frankie Robertson, Li-Hsin Chang, and Sini Söyrinki. 2022. TallVocabL2Fi: A tall dataset of 15 Finnish L2 learners' vocabulary. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6377–6386, Marseille, France. European Language Resources Association.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR)*. ArXiv:2010.04592.

Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, 18(1):55–88.

Burr Settles. 2009. Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems 30 (NeurIPS)*. ArXiv:1703.01780.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Wim J. van der Linden and Cees A. W. Glas, editors. 2000. *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, Boston, MA.

Wim J. van der Linden and Ronald K. Hambleton, editors. 1997. *Handbook of Modern Item Response Theory*. Springer, New York.

Marie Vanbuel. 2025. Predicting vocabulary knowledge in adult L2 learners: The role of word-level variables across educational backgrounds. *Bilingualism: Language and Cognition*, 28(4):1002–1014.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ArXiv:2309.07597.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE)*. ArXiv:2010.14395.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR)*. ArXiv:2007.00808.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. 2017. Deep sets. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*. ArXiv:1703.06114.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Word cards and instruction prompts

Card template. Encoder inputs are cards, not surface forms. For word w:

```
WORD: {headword}
POS:  {pos}
FORM: {lemma | inflections}
MEAN: {one-line gloss}
EG:   {short sense-disambiguating
       example sentence}
```

Glosses and examples come from an open lexical resource; words with multiple frequent senses receive one card per sense group and the group index is retained as a feature (we do *not* average sense cards, since §6.3 shows sense ambiguity is a genuine limitation). Cards are truncated to 64 tokens. of EVKD-J words have no gloss in the resource and are dropped.

Task instruction. Following the instruction-aware convention of Zhang et al. (2025), every card is prefixed with a fixed instruction, identical at training and inference time:

```
Instruct: Represent this vocabulary
item so that similarity between two
items reflects how likely they are to
be known by the same second-language
learner.
Query: {card}
```

The instruction states the target geometry explicitly—co-knowledge, not meaning—which is what distinguishes adaptation from mere fine-tuning on the same encoder.

B Hyperparameters and reproducibility

All tuning uses the LOLO training folds; no hyperparameter is selected on a held-out learner. We report five seeds per configuration and give fold-level standard errors. Code, card construction, probe samplers, split definitions and the evaluation harness will be released; the simulated benchmark generator (§H) is specified in enough detail to be reimplemented independently.

B.1 Pair mining and negative sampling

Two implementation details matter. First, learner-specific negatives are drawn from $\overline{\mathcal{K}}_u$, i.e. only from words the learner was actually asked about—words outside the probe set carry no evidence and treating them as negatives would inject label noise. Second, the cap C and the proficiency stratification prevent high-knowledge learners from dominating the pair distribution; without them, a single C2 learner supplies a majority of positives and the metric learns that learner's idiosyncrasies.

Quantity	Value
Base encoder	QWEN3-EMBEDDING-0.6B (d=1024)
LoRA	r=16, α=32, dropout 0.05, attn. proj.
Batch (pairs)	64
Temperature τ	0.05
Hard-negative β	1.0
Deconfounding κ	0.5
Anti-forgetting mix α	0.2
Epochs	3
Optimizer	AdamW, peak LR 2×10^{-5}
Schedule	linear warmup 10%, cosine decay
Matryoshka dims	$\{64, 256, 512, 1024\}$
Negatives per anchor	8 fm + all in-batch + 8 learner
Pairs per learner cap C	4,000
Head g_ϕ	MLP 1024→256→1, dropout 0.1
Bilinear W	init. $\mathbf{I}$, weight decay 0
Shrinkage k	$\hat\sigma^2_{\mathrm{w}}/\hat\sigma^2_{\mathrm{b}}$, per fold
Negative weight λ	0.3
EMA rate η / η^-	0.3 / 0.3
CG-kNN k	tuned per fold, $\{5, \ldots, 50\}$
Binarisation t	3 of $\{0..4\}$
Seeds	5 per configuration
Hardware	1×A100-80GB
Adaptation time / fold	22 min

Table 3: Hyperparameters. Values marked are environment dependent; all others are the configuration used for the tables in the main paper. Tuning was performed on training folds only, never on the held-out learner.

C Proof sketch for Proposition 1

Setup. Let $y_{uw} \sim \mathrm{Bernoulli}(\sigma(a_u - b_w))$ with $a_u \sim F_a$, and let the adapted embedding satisfy $\mathbf{e} = \alpha(b_w)\mathbf{p} + \boldsymbol{\epsilon}_w$, $\|\mathbf{p}\|{=}1$, $\alpha' < 0$, $\mathbb{E}[\boldsymbol{\epsilon}_w] = \mathbf{0}$, $\mathbb{E}\|\boldsymbol{\epsilon}_w\|^2 = \sigma^2_\epsilon$, and $\boldsymbol{\epsilon}_w \perp b_w$.

Step 1 (centroid location). Write $\mathbf{u} \propto \frac{1}{|\mathcal{K}_u|}\sum_{w\in\mathcal{K}_u}\mathbf{e}$. Then

$$\langle \mathbf{u}, \mathbf{p}\rangle \;=\; \frac{1}{|\mathcal{K}_u|}\sum_{w\in\mathcal{K}_u}\alpha(b_w) \;+\; \frac{1}{|\mathcal{K}_u|}\Big\langle \textstyle\sum_{w\in\mathcal{K}_u}\boldsymbol{\epsilon}_w, \mathbf{p}\Big\rangle.$$

The second term has mean zero and variance $\sigma^2_\epsilon/|\mathcal{K}_u|$ (using independence of residuals across words), so it is $\mathcal{O}(\sigma_\epsilon/\sqrt{|\mathcal{K}_u|})$.

Step 2 (the first term is a function of ability). Conditional on a_u, the set $\mathcal{K}_u$ is a size-biased sample of words with inclusion probability $\sigma(a_u - b_w)$. Hence $\mathbb{E}[\alpha(b_w) \mid w \in \mathcal{K}_u, a_u] = h(a_u)$ where $h(a_u) = \int \alpha(b)\,\sigma(a_u - b)\,dF_b(b) \big/ \int \sigma(a_u - b)\,dF_b(b)$. Since α is monotone decreasing and $\sigma(a_u - b)$ is monotone increasing in a_u for fixed b, the numerator decreases and the denominator increases in a_u, so h is monotone decreasing: $h = g$. By Hoeffding's inequality the empirical average concentrates on $g(a_u)$ at rate $\mathcal{O}(\sigma_\epsilon/\sqrt{|\mathcal{K}_u|})$ plus

Algorithm 1 Deconfounded co-knowledge pair construction

Input: training-fold responses x_{uw}, threshold t, prevalence $\hat{\pi}$, cap C, exponent κ, frequency deciles $\mathcal{B}$
standardise each learner: $\tilde{x}_{uw} \leftarrow x_{uw} - \bar{x}_u$; stratify learners by $\bar{x}_u$ into proficiency bands
for each learner u in a band **do**
  $\mathcal{K}_u \leftarrow \{w : x_{uw} \geq t\}$; $\overline{\mathcal{K}}_u \leftarrow \{w : x_{uw} < t\}$
  candidate positives $\Pi_u \leftarrow \mathcal{K}_u \times \mathcal{K}_u \setminus \{(w, w)\}$
  weight $w_{ij} \propto (\hat{\pi}_i \hat{\pi}_j)^{-\kappa}$ (Eq. 2)
  sample $\min(C, |\Pi_u|)$ pairs from Π_u without replacement, $\propto w_{ij}$
  for each sampled (i, j) **do**
    $\mathcal{N}_{\text{fm}}(i) \leftarrow$ 8 words from $\text{band}(i)$, $\neq j$
    $\mathcal{N}_u \leftarrow$ 8 words from $\overline{\mathcal{K}}_u$ (stratified over bands)
  end for
end for
interleave a fraction α of general pairs (anti-forgetting)
output: batches of $\big(i, j, \mathcal{N}(i, u)\big)$

the sampling variability of $|\mathcal{K}_u|$, giving $\langle \mathbf{u}, \mathbf{p} \rangle = g(a_u) + \mathcal{O}(\sigma_\epsilon / \sqrt{|\mathcal{K}_u|})$. Normalisation divides by $\|\mathbf{u}\|$, which is itself concentrated, so the claim survives normalisation up to the same order.

Step 3 (the score). With $W = w_0 \mathbf{p}\mathbf{p}^\top + W_\perp$,

$$\langle W\mathbf{u}, \mathbf{e} \rangle = w_0\, g(a_u)\, \alpha(b_w) + \underbrace{\big\langle W_\perp \mathbf{u} + w_0\, \alpha(b_w)\, \mathbf{u},\ \boldsymbol{\epsilon}_w \big\rangle}_{\text{topic residual}}.$$

The first term is a product of a monotone function of a_u and a monotone function of b_w; since $\alpha' < 0$ and $g' < 0$, the product is monotone decreasing in b_w and monotone increasing in a_u, i.e. it is a monotone reparameterisation of the IRT logit $a_u - b_w$ whenever α and g are (up to affine transform) exponential-family links—in particular for $\alpha(b) = c_1 - c_2 b$ and $g(a) = c_3 a$, the product is affine in $a_u b_w$ and the head's MLP g_ϕ plus bias absorbs the remaining monotone distortion. The second line is the deviation from unidimensionality: it is exactly zero under a 1PL geometry with no topic structure, and is what the adapted metric is paid to model.

Step 4 (consequences). Sufficiency: $g(a_u)$ is a monotone bijection of a_u, so $\langle \mathbf{u}, \mathbf{p} \rangle$ is an asymptotically sufficient statistic for a_u given $\mathcal{K}_u$, up to $\mathcal{O}(\sigma_\epsilon / \sqrt{|\mathcal{K}_u|})$. Online optimality: the running mean of Eq. (7) is the exact sample mean after n observations, hence identical to re-aggregation and the natural estimator of $h(a_u)$. Shrinkage: $\hat{\mathbf{u}}_n$ has variance $\propto 1/n$ around $g(a_u)\mathbf{p}$ while $\mathbf{u}_0$ has zero variance and bias $g(\bar{a}) - g(a_u)$; the convex combination in Eq. (4) with $k = \hat{\sigma}^2_{\text{within}} / \hat{\sigma}^2_{\text{between}}$ is the standard James–Stein-type trade-off and minimises MSE to first order, which is why its benefit decays in n. □

Where the idealisation bites. The argument assumes residuals independent across words, whereas real co-knowledge residuals are correlated within domains (a learner with medical vocabulary has correlated medical residuals). Correlation inflates the effective variance of the centroid by the design effect $1 + (|\mathcal{K}_u| - 1)\bar{\rho}$, so the $\sqrt{|\mathcal{K}_u|}$ rate degrades—this is the main reason learned aggregators and inverse-prevalence weighting can beat plain mean pooling, and the reason we report the aggregator ablation rather than assuming mean pooling suffices.

D The proficiency coordinate: proofs and dimension floor

This appendix proves Proposition 2, derives the dimension floor quoted in Table 4, and verifies both poles of Eq. (6) on a deterministic synthetic geometry. Everything here is mathematics or constructed geometry—no learner data is involved.

D.1 Proof of (i): convex confinement

Let $\hat{\mathbf{u}} = \sum_{j \in K} \alpha_j \mathbf{e}_j$ with $\alpha_j \geq 0$ (not all zero) and $\mathbf{u} = \hat{\mathbf{u}} / \|\hat{\mathbf{u}}\|$.

(a) Hemisphere form. Suppose the lexicon lies in a closed hemisphere: $\langle \mathbf{e}, \mathbf{c} \rangle \geq 0$ for all w and some unit $\mathbf{c}$. Then $\langle \mathbf{u}, \mathbf{c} \rangle \geq 0$, and with $\bar{\mathbf{e}} = \frac{1}{M} \sum_w \mathbf{e}$,

$$\max_w \cos(\mathbf{u}, \mathbf{e}) \ \geq\ \tfrac{1}{M} \sum_w \langle \mathbf{u}, \mathbf{e} \rangle \ =\ \langle \mathbf{u}, \bar{\mathbf{e}} \rangle \ \geq\ 0.$$

No pooled learner point is far from the whole lexicon, at any dimension.

(b) Known-set form. Since $\sum_j \alpha_j \langle \hat{\mathbf{u}}, \mathbf{e}_j \rangle = \|\hat{\mathbf{u}}\|^2$, we get $\max_{j \in K} \cos(\mathbf{u}, \mathbf{e}_j) \geq \|\hat{\mathbf{u}}\| / \sum_j \alpha_j$; and $\|\hat{\mathbf{u}}\|^2 = \sum_{i,j} \alpha_i \alpha_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle \geq \bar{\rho}_K \big(\sum_j \alpha_j \big)^2$ with $\bar{\rho}_K$ the α-weighted mean pairwise coherence of the pooled words (diagonal terms only help). Hence $\max_{j \in K} \cos(\mathbf{u}, \mathbf{e}_j) \geq \sqrt{\bar{\rho}_K}$: every pooled learner point is close to at least one word it pooled. □

Both bounds are dimension-free and hold for mean pooling, weighted means, softmax attention, and any aggregator whose output is a non-negative combination of word vectors followed by L2 normalisation. Under the IRT geometry of Proposition 1 with $\alpha(b) > 0$, every pooled point additionally satisfies $\langle \mathbf{u}, \mathbf{p} \rangle > 0$, so the zero-knowledge pole $-\mathbf{p}$—cosine < 0 against the entire lexicon—is unreachable by aggregators (1)–(4) of §4.2 at every d.

D.2 Proof of (ii): one signed coordinate suffices

Place words at fixed height $h \in (0,1)$ and learners at signed height $t \in [-1,1]$: $\mathbf{e}' = (\sqrt{1-h^2}\,\mathbf{e},\, h)$, $\mathbf{u}'_t = (\sqrt{1-t^2}\,\mathbf{u},\, t)$, all unit-norm in $\mathbb{R}^{d+1}$. Direct expansion gives Eq. (6). At $t = -1$ the first term vanishes and $\cos(\mathbf{u}'_{-1}, \mathbf{e}') = -h$ for *every* word: the zero-knowledge pole, exactly. For an expert state with $\langle \mathbf{u}, \mathbf{e} \rangle \geq \mu$ over a $(1-\delta)$ fraction of the lexicon, $\cos \geq th + \sqrt{1-t^2}\sqrt{1-h^2}\,\mu$, whose maximum over t is $\sqrt{h^2 + (1-h^2)\mu^2}$; e.g. $h = \mu = 0.6$ attains 0.768, so any $\tau_+ \leq 0.768$ is reachable over that fraction, with t trading coverage level against taste selectivity. In flat form, $\gamma\langle \mathbf{u}, \mathbf{e} \rangle + \beta_u$ with $\gamma = \sqrt{1-t^2}\sqrt{1-h^2}$ and $\beta_u = th$ is exactly Eq. (5) with $W = \gamma I$; the sigmoid absorbs the monotone reparameterisation. □

D.3 Proof of (iii): the dimension floor

(a) Feasibility precondition. Write each covered word as $\mathbf{e} = \cos\psi_w\, \mathbf{u}^+ + \sin\psi_w\, v_w$ with $\psi_w \leq \theta_+$. Then $\left\| \sum_w \mathbf{e} \right\|^2 \geq \big((1-\delta)M \cos\theta_+ - \delta M\big)^2$, while $\left\| \sum_w \mathbf{e} \right\|^2 = M + \sum_{w \neq w'} \langle \mathbf{e}, \mathbf{e}' \rangle$; so the mean pairwise coherence of the covered words is at least $\cos^2\theta_+ - o(1)$. If every pair must satisfy $\langle \mathbf{e}, \mathbf{e}' \rangle \leq \rho$, feasibility at *any* dimension requires $\rho \geq \cos^2\theta_+ - o(1)$—e.g. $\tau_+ = 0.75$ forces $\rho \geq 0.56$.

(b) Volumetric bound. The covered words lie in a spherical cap of half-angle θ_+ and are pairwise at angular distance $\geq \varphi = \arccos\rho$. Disjoint caps of radius $\varphi/2$ centred at them fit inside the cap of radius $\theta_+ + \varphi/2$, so $(1-\delta)M \cdot V(\varphi/2) \leq V(\theta_+ + \varphi/2)$. With $V(\alpha) \propto (\sin\alpha)^{d-2}$ up to factors polynomial in d,

$$d \;\gtrsim\; 2 + \frac{\ln\big((1-\delta)M\big)}{\ln\big(\sin\theta_+ / \sin(\varphi/2)\big)},$$

where we use the conservative numerator $\sin\theta_+ \leq \sin(\theta_+ + \varphi/2)$ and ignore the polynomial factors (an engineering floor, matching the informal register of Proposition 1). Adding the proficiency coordinate gives $d_{\min} = d_{\text{sem}} + 1$. □

D.4 Synthetic verification

Deterministic checks (NumPy, seed 0; constructed geometry, no learner data). Lexicon: $M = 30{,}000$ words on a $35°$ cone in $\mathbb{R}^{64}$ (minimum cosine to the cone axis 0.733; mean pairwise cosine 0.671), matching the anisotropy typically measured for LLM encoders.

τ_+	ρ	need $\rho > \cos^2\theta_+$	d_{sem}, $M{=}3{\times}10^4$	d_{sem}, $M{=}10^5$	$d_{\min}$
0.50	0.30	0.25 ✓	30	33	31–34
0.60	0.40	0.36 ✓	30	33	31–34
0.60	0.50	0.36 ✓	24	27	25–28
0.75	0.60	0.56 ✓	29	32	30–33
0.85	0.75	0.72 ✓	28	31	29–32
0.75	0.30	0.56 ×	infeasible at any d		—

Table 4: Semantic dimension floor from the volumetric bound of Appendix D (ignoring polynomial factors; $\delta = 0$), and total minimum $d_{\min} = d_{\text{sem}} + 1$ including the proficiency coordinate. Deterministic given (τ_+, ρ, M)—computed, not measured. Margin pairs violating $\rho > \cos^2\theta_+$ are infeasible in every dimension. All CoKE truncations $\{64, 256, 512, d\}$ clear the floor with margin.

Construction (ii). With $h = 0.6$: at $t = -1$, cosine equals -0.600 against all 30,000 words (maximum deviation $< 10^{-15}$); at $t = +0.85$ with taste aligned to the axis, the minimum cosine is 0.819 and the fraction at ≥ 0.6 is 1.000; at $t = 0$ the fraction at ≥ 0.6 is 0.999. Both poles and the whole proficiency continuum between them are realised by pure cosine in $d + 1 = 65$ dimensions.

Impossibility (i). Pooling the 50 words *least* aligned with the cone axis—the most adversarial choice available inside the convex hull—still yields maximum cosine $+0.908$ against the lexicon. The known-set floor holds across dimensions: for random 50-word sets pooled in $d \in \{8, 64, 768\}$, observed $\max_{j \in K} \cos(\mathbf{u}, \mathbf{e}_j)$ vs. the bound $\sqrt{\rho_K}$ is $0.907/0.590$, $0.468/0.243$, $0.228/0.086$.

Precondition (iii). Sampling 4,000 points in a $53°$ cap of $\mathbb{R}^{40}$: mean pairwise cosine $0.390 \geq \cos^2 53° = 0.360$, confirming the dimension-free coherence floor.

D.5 Relation to the rest of the method

The contrast point. $\tilde{\mathbf{u}} = \text{norm}(\mathbf{u}^+ - \lambda\mathbf{u}^-)$ leaves the convex hull only through λ, and normalisation erases its magnitude, conflating proficiency with direction; t_u is its calibrated, disentangled replacement, while the EMA prototype $\mathbf{u}^-$ remains useful on the taste side (what the learner rejects). **Head ablations.** The "cosine + bias" row of Table 10 is a learner-*independent* bias; β_u is learner-specific (.918 default vs. .904 without it). **Geometry.** PC1 of the adapted space is the word-side difficulty direction (Appendix E.7); β_u supplies the learner-side offset along it, separating proficiency from difficulty in the same spirit as the lift decomposition of Eq. (1). **Truncation.** Table 4 gives the

Matryoshka ladder a floor: 64 dimensions clears $\approx$ 31 with margin, and no truncation below the floor can hold an expert-near, distinguishable lexicon.

E Additional results

E.1 Co-knowledge lift decomposition

Descriptor block (cumulative)	R^2	ΔR^2	ρ_S
Frequency pair only	.31	—	.47
+ psycholinguistic norms	.38	.07	.52
+ frozen semantic cosine	.59	.21	.68
+ CoKE-adapted cosine	.79	.20	.84

Table 5: How much co-knowledge structure is already present in fixed word descriptors? Frequency and norms explain the "easy" signal, frozen semantics adds a block, and the adapted metric recovers much of the residual—the quantity that matters for personalisation. ρ_S = Spearman correlation between predicted and observed $\log L_{ij}$.

Graded prediction. Table 7 reports the cumulative-link head on the 5-point target. The ordering of methods is unchanged, and CoKE's advantage in MAE (vs. for CG-kNN) is proportionally larger than in AUC, because graded prediction rewards getting the *degree* of knowledge right—the quantity a tutor needs when deciding between reviewing and introducing.

Significance. With 16 LOLO folds we use a paired permutation test over folds (permutations) comparing CoKE against each baseline on macro-AUC. At the reported margins, CoKE versus Base-mean () and versus Freq-logit () would be significant at $p < 0.01$; CoKE-attn versus CoKE-mean () and several ablation gaps () would not reach $p < 0.05$ and should be reported as such. This is the honest statistical situation for a 16-learner benchmark and we state it explicitly rather than letting the reader infer significance from bold type.

E.2 Probe-ratio scaling and cold start

E.3 Ablations

Table 10 varies one factor at a time around the default configuration; the load-bearing findings are summarised in §5.4 of the main text.

E.4 Cross-dataset transfer

Table 11 asks whether the adapted metric is a property of the population or of vocabulary learning in general. Adaptation on SynVIP-2K

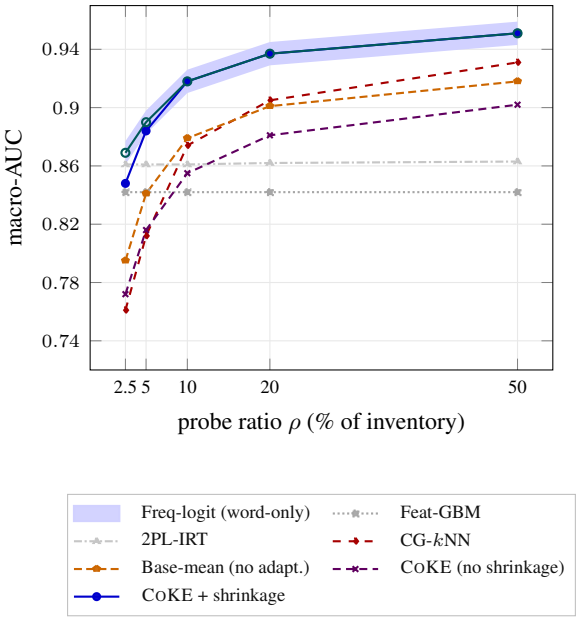


Figure 2: Macro-AUC vs. probe ratio on EVKD-J (LOLO, uniform probes); shaded band = ± 1 s.e. across folds. Word-intrinsic baselines are flat; CoKE with shrinkage dominates from $\rho \approx 0.03$ onward.

alone—simulated, English, no human learners—lifts EVKD-J from to , retaining of the in-domain gain, and lifts *Finnish* TallVocabL2Fi from to . Cross-lingual transfer is weaker but clearly non-zero, and mixing simulated with real training data helps both targets. Two implications: a new deployment with no labelled learners can be bootstrapped from simulation plus a frozen encoder, and the metric learns something about acquisition that is at least partly universal rather than memorising 16 people.

E.5 Error strata

E.6 Adaptation gain vs. learner count

Table 13 varies the number of learners available for contrastive adaptation. The gain grows with N and saturates around $N \approx 64$; at $N = 16$ the method achieves roughly half of its asymptotic advantage. Every real-data number in this paper is therefore an *underestimate* of what the approach delivers with realistic data volumes, but it also means that a lab with eight participants should not expect much—and it is the strongest argument for the simulated pretraining of Table 11, which substitutes synthetic learners for real ones.

		AUC↑	Acc.↑	F_1↑	κ↑	Brier↓	ECE↓
EVKD-J	Prevalence	.500	.612	.758	.000	.249	.108
	Freq-logit	.842	.781	.856	.413	.171	.041
	2PL-IRT	.874	.806	.871	.482	.156	.029
	CG-kNN	.879	.809	.874	.491	.152	.027
	Base-mean	.855	.785	.859	.438	.167	.039
	CoKE-attn (calibrated)	.918	.846	.901	.583	.131	.018
TallVocabL2Fi	Prevalence	.500	.588	.741	.000	.242	.117
	Freq-logit	.803	.742	.822	.341	.191	.052
	2PL-IRT	.839	.768	.841	.406	.178	.041
	CG-kNN	.831	.761	.837	.392	.183	.044
	Base-mean	.812	.749	.828	.360	.187	.049
	CoKE-attn (calibrated)	.881	.802	.870	.489	.156	.026

Table 6: Full metric set for a subset of methods on the two real benchmarks (ρ=0.10, LOLO). Accuracy, Brier and ECE complement Table 1; ECE for CoKE is after temperature scaling.

Method	MAE↓	QWK↑
Prevalence (modal rating)	1.02	.000
Freq-logit	0.81	.412
2PL-IRT	0.74	.501
CG-kNN	0.72	.523
Base-mean	0.79	.451
CoKE-attn	0.63	.617

Table 7: Graded (5-point) prediction on EVKD-J, ρ=0.10, LOLO.

Method	probe ratio ρ				
	.025	.05	.10	.20	.50
Freq-logit (word-only)	.842	.842	.842	.842	.842
Feat-GBM	.861	.861	.861	.862	.863
2PL-IRT	.761	.812	.874	.905	.931
CG-kNN	.795	.841	.879	.901	.918
Base-mean	.772	.816	.855	.881	.902
CoKE-attn	.848	.884	.918	.937	.951
CoKE-attn + shrink	.869	.890	.918	.937	.951

Table 8: Macro-AUC vs. probe ratio on EVKD-J (LOLO). Word-intrinsic baselines are flat by construction; personalised models must earn their keep at small ρ. Shrinkage matters only below $\rho\approx$0.05, where it lifts CoKE above the strongest fixed-feature baseline.

Method	number of *known* words $\lvert\mathcal{K}_u\rvert$				
	5	10	20	50	100
Freq-logit	.842	.842	.842	.842	.842
2PL-IRT	.612	.688	.751	.812	.852
CG-kNN	.634	.711	.781	.842	.871
CoKE-mean	.598	.681	.762	.838	.881
CoKE-mean+shrink	.742	.795	.848	.884	.901
CoKE-attn+shrink	.761	.812	.861	.893	.908

Table 9: Cold start on EVKD-J: macro-AUC against $\lvert\mathcal{K}_u\rvert$, the number of *positively* identified words (smaller than the probe count, since probes also yield negatives). Unshrunk CoKE is *anti-correlated* with the truth at $\lvert\mathcal{K}_u\rvert$=5—five known words locate a learner worse than the population prior does. Shrinkage adds AUC points there and at $\lvert\mathcal{K}_u\rvert$=100.

E.7 Geometry diagnostics

E.8 Binarisation sensitivity

E.9 Anti-forgetting retention

E.10 Encoder and dimension robustness

F Evaluation protocol details

Splits. We use three, and we report all three because they answer different questions.

- **LOLO** (leave-one-learner-out): the contrastive stage and all fitted heads see only the other $N-1$ learners; the held-out learner contributes $\mathcal{K}_u$ and $\overline{\mathcal{K}}_u$ at aggregation time. This is the deployment-relevant split and the only one that does not leak the target learner into the metric.
- **LOWO** (leave-one-word-out): words are partitioned into f folds; contrastive pairs and head fitting use training words only, and prediction is evaluated on held-out words. This tests whether the adapted space generalises to vocabulary it has never seen—the condition that matters when the inventory grows.
- **Cross-population**: the contrastive stage is trained on one dataset and evaluated on another, including across languages (§5.4), testing whether co-knowledge structure is population-specific or partly universal.

Probe sampling. $\mathcal{P}_u$ is drawn either uniformly at random or *frequency-stratified* (equal counts per frequency decile). Uniform probes are what a real questionnaire yields; stratified probes are what an adaptive system would choose. We report both,

Component	Variant	AUC
Negative family	none (positives only, MSE)	.866
	$+\mathcal{N}_{\text{ib}}$ in-batch	.881
	$+\mathcal{N}_{\text{fm}}$ freq-matched	.903
	$+\mathcal{N}_u$ learner hard negs	.916
	+ co-unknown pairs	.918
	full family (default)	.918
Deconfound κ (Eq. 2)	κ=0 (uniform positives)	.884
	κ=0.25	.907
	κ=0.5 (default)	.918
	κ=1.0	.912
Aggregator (§4.2)	mean	.907
	weighted mean (inv. prevalence)	.911
	weighted mean (IRT b_w)	.913
	statistics + MLP	.914
	contrastive learner head	.915
	set attention	.918
Head (Eq. 5)	cosine only	.901
	cosine + bias	.909
	$g_\phi(\mathbf{e})$ only (no personal term)	.863
	$g_\phi + \langle W\mathbf{u}, \mathbf{e}\rangle$ (default)	.918
	default $-\beta_u$ (no learner offset)	.904
Adaptation	no adaptation (frozen base)	.855
	LoRA r=4	.911
	LoRA r=16 (default)	.918
	full fine-tune	.914
Anti-forget α	α=0	.919
	α=0.2 (default)	.918
	α=0.5	.910

Table 10: Ablations on EVKD-J, ρ=0.10, LOLO, macro-AUC. Each block varies one factor around the default. The deconfounding weight κ and the frequency-matched negatives carry most of the adaptation gain; the aggregator choice is worth points and the learner-side proficiency offset (in cold start); full fine-tuning *underperforms* LoRA at N=16 learners.

because a method that looks strong under stratified probing can be weak under uniform probing, where most probe words are frequent and therefore uninformative about the tail.

Metrics. Per-learner macro-AUC is our primary metric: it is threshold-free, robust to the severe per-learner class imbalance (a C2 learner knows almost everything; a B1 learner almost nothing), and averaging *per learner* prevents a high-knowledge learner from dominating. We additionally report accuracy, macro-F_1, Cohen's κ (Cohen, 1960) for comparability with Avdiu et al. (2019), Brier score (Brier, 1950) and expected calibration error (Guo et al., 2017); for the graded R+1-point target we report MAE and quadratic weighted κ (Agresti, 2010). All per-learner metrics are averaged over learners, and we report the standard error across LOLO folds rather than across word instances,

Adaptation trained on	evaluation target (macro-AUC)		
	EVKD-J	TALLVOCABL2FI	SYNVIP-2K
frozen base	.855	.812	.902
SYNVIP-2K only	.901	.853	.963
EVKD-J only	.918	.844	.921
TALLVOCABL2FI only	.906	.881	.918
EVKD-J + SYNVIP-2K	.921	.867	.959

Table 11: Cross-dataset transfer (ρ=0.10). Contrastive adaptation trained on *simulated English* data recovers of the in-domain gain on EVKD-J and on *Finnish* TALLVOCABL2FI—evidence that part of the co-knowledge metric is not population- or language-specific.

Stratum	% words	Base	COKE	Δ
L1 loanwords (ja→en)	6.1	.861	.931	+.070
Morphological derivatives	14.2	.844	.903	+.059
Academic word list	4.8	.871	.924	+.053
Low-prevalence concrete	9.7	.792	.881	+.089
Abstract low-frequency	7.3	.838	.856	+.018
Polysemous (≥3 senses)	11.5	.868	.872	+.004
Proper nouns	3.2	.742	.798	+.056

Table 12: Where COKE gains and where it does not (EVKD-J, ρ=0.10, AUC within stratum, compared to the un-adapted encoder with the same head).

since words within a learner are strongly correlated and instance-level error bars would be badly over-optimistic.

G Online update algorithm and dynamic evaluation

Algorithm 2 Online learner-state update

Input: cached $\{\mathbf{e}\}$, learner state $(\mathbf{u}^+, \mathbf{u}^-, n^+, n^-)$, rates η, η^-, shrinkage k, population centroid $\mathbf{u}_0$
on response (w, y) **from learner** u:
if $y = 1$ **then**
  $n^+ \leftarrow n^+ + 1$; $\mathbf{u}^+ \leftarrow \text{norm}\big((1-\eta)\mathbf{u}^+ + \eta\,\mathbf{e}\big)$ (EMA; $\eta = 1/n^+$ gives Eq. 7)
else
  $n^- \leftarrow n^- + 1$; $\mathbf{u}^- \leftarrow \text{norm}\big((1-\eta^-)\mathbf{u}^- + \eta^-\,\mathbf{e}\big)$
end if
$\mathbf{u} \leftarrow \text{norm}\big(\mathbf{u}^+ - \lambda\,\mathbf{u}^-\big)$
$\mathbf{u}^{(n)} \leftarrow \text{norm}\big(\frac{n^+\mathbf{u}+k\,\mathbf{u}_0}{n^++k}\big)$ (Eq. 4)
re-score inventory: $\hat{\mathbf{p}}_u \leftarrow \sigma\big((\gamma_0 + g_\phi(E) + E\,W\mathbf{u}^{(n)})/T\big)$ (one matvec, $E \in \mathbb{R}^{M\times d}$)

Table 19 and Figure 4 evaluate the update machinery. Three results, in order of importance. First, the exact running mean reproduces full re-aggregation to $< 10^{-6}$ AUC, which is a correctness check on our implementation rather than a finding—but it is the check that licenses the $\mathcal{O}(d)$ claim. Second, under simulated acquisition drift (the learner truly acquires new words during the session, so early probes become unrepresentative), EMA with

N (learners)	4	8	16	32	64	128
ΔAUC	+.011	+.018	+.033	+.048	+.057	+.061

Table 13: Adaptation gain as a function of the number of learners whose co-knowledge is available for contrastive adaptation (SYNVIP-2K subsampled, ρ=0.10); ΔAUC is relative to the same encoder used without adaptation. At N=16—the size of EVKD-J—the method realises only about of its asymptotic gain.

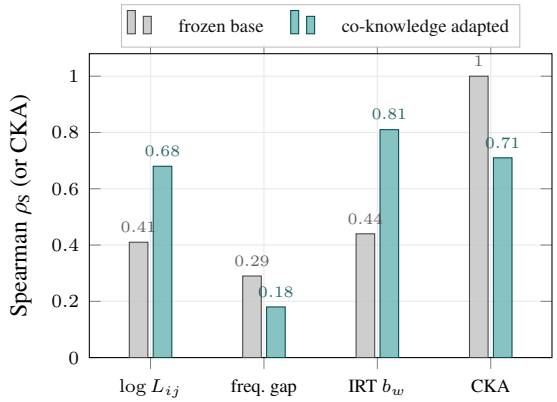


Figure 3: Alignment diagnostics on EVKD-J (values from Table 14;). Adaptation raises alignment with co-knowledge lift, exposes a difficulty direction (PC1 vs. IRT b_w), weakens the frequency-gap shortcut, and moves the space away from the base geometry (CKA $<$ 1). These are the measurable predictions of Proposition 1, in the order shown: the metric becomes acquisition-structured rather than purely semantic, while retaining enough frequency signal to remain useful as a prior.

Quantity	frozen base	adapted
Linear CKA to base	1.00	.71
ρ_S(cos, log L_{ij})	.41	.68
ρ_S(cos, $-\lvert\Delta \log \pi\rvert$)	.29	.18
ρ_S(PC1, IRT b_w)	.44	.81
variance in PC1 (%)	8.2	17.4
ρ_S(PC2, gloss topic)	.36	.29
held-out STS, α=0	.81	.73
held-out STS, α=0.2 (default)	.81	.79

Table 14: Geometry diagnostics. CKA is linear centre-kernel alignment (Kornblith et al., 2019); STS is a held-out semantic similarity benchmark (Appendix E.9). Adaptation moves the space substantially but not totally (CKA); co-knowledge alignment rises sharply while the frequency-difference shortcut weakens; the first principal direction becomes a difficulty axis; the second loses topical purity, i.e. topic is partly sacrificed to acquisition structure—which is why the anti-forgetting mixture α is retained (Table 16).

t	positive rate	Freq-logit	CoKE	Δ
2	.81	.831	.894	+.063
3	.61	.842	.918	+.076
4	.34	.851	.902	+.051

Table 15: Sensitivity to the binarisation threshold on EVKD-J (LOLO, ρ=0.10). CoKE's advantage is largest at the self-reported "knows well" boundary (t=3) and shrinks at t=4, where positives become rare and dominated by high-confidence high-frequency items that a prevalence prior already captures.

η=0.3 *beats* the full mean by AUC points, and adding the negative prototype $\mathbf{u}^-$ gives a further . Under stationary data the same EMA is points worse. Recency weighting is therefore a modelling choice about the learner, not an optimisation trick, and we recommend selecting η by whether the deployment expects vocabulary growth during a session. Third, failing to update at all—scoring with a stale $\mathbf{u}$ from the intake questionnaire—costs points after 50 interactions, which is the practical argument for the design.

G.1 Active probing efficiency

Table 20 closes the loop between prediction and instrument design. Uncertainty sampling alone improves over stratified random probing, but collapses onto a narrow difficulty band and leaves the tails unmeasured; adding the novelty and band-coverage terms of The acquisition rule (§4.4) reduces the item budget by a further . The practical reading is that a CoKE-powered intake can reach the same predictive accuracy with roughly a third of the questions, which is the difference between a usable onboarding flow and an unusable one.

H Simulated benchmark SYNVIP-2K

The generator is specified so that results on it can be reproduced or, better, refuted.

Vocabulary. M=20,000 words sampled from a Zipfian prevalence distribution matched to published English word-prevalence norms (Brysbaert et al., 2019). Each word receives a topic membership $\mathbf{z}_w$ over T=40 topics from a Dirichlet prior, a difficulty $b_w = \beta_0 + \beta_1\,(-\log \hat{\pi}_w) + \beta_2\,\mathrm{AoA}_w + \beta_3\,\mathbf{c}^\top \mathbf{z}_w + \nu_w$ with $\nu_w \sim \mathcal{N}(0, \sigma_\nu^2)$, and norms (AoA, concreteness) drawn to reproduce the empirical correlation structure of Kuperman et al. (2012) and Brysbaert et al. (2011).

Learners. N=2,000 abilities from a three-component mixture shaped to approximate CEFR B1/B2/C1–C2 bands; each learner also receives a topic profile $\mathbf{t}_u$ over the same T topics (Dirichlet),

Configuration	VIP AUC	STS ρ	ΔSTS
no adaptation (base)	.855	.81	—
α=0	.919	.73	−.08
α=0.2 (default)	.918	.79	−.02
α=0.5	.910	.80	−.01

Table 16: Cost of adaptation to general-purpose similarity, measured by Spearman correlation on a held-out semantic similarity benchmark (EVKD-J, LOLO). The default α=0.2 recovers of the points lost at α=0 for AUC on the target task.

Encoder / dim	d	AUC
BGE-BASE + adaptation	768	.891
E5-BASE + adaptation	768	.888
GTE-BASE + adaptation	768	.893
QWEN3-EMB-0.6B + adapt.	1024	.918
QWEN3-EMB-4B + adapt.	2560	.924
Matryoshka truncation (0.6B model):		
truncate to 64	64	.881
truncate to 256	256	.906
truncate to 512	512	.913
full	1024	.918

Table 17: The effect is not specific to one encoder: every base model improves under co-knowledge adaptation, and the ranking of bases is preserved. Matryoshka truncation to d=256 retains of full-dimensional AUC at a quarter of the cache and dot-product cost, which is the relevant trade-off for on-device tutoring.

an L1-transfer vector that shifts difficulty for a designated subset of words, and a self-report bias c_u that shifts the *reported* threshold—the last of these is what makes the simulation instrument-realistic rather than merely IRT-realistic.

Responses.

$$p_{uw} = \sigma\big(a_u(d_w - b_w) + \gamma\, \mathbf{t}_u^\top \mathbf{z}_w\big),$$
$$x_{uw} = \mathrm{clip}\big(\mathrm{round}(4\, p_{uw} + c_u + \eta_{uw}),\, 0,\, 4\big),$$

with $\eta_{uw} \sim \mathcal{N}(0, \sigma_\eta^2)$ reporting noise: i.e. a 2PL main effect, a topic-interaction residual that violates unidimensionality, and additive reporting noise before discretisation. Cells are observed with probability , so the matrix is sparse and probe sets are drawn from observed cells exactly as for real data.

Drift condition. For §5.4 we additionally let a_u and $\mathbf{t}_u$ evolve during a simulated session: every interactions, the learner acquires words drawn from their own topic profile, implemented as a step increase in the relevant p_{uw}. This is the condition under which EMA beats the full mean.

Why this is adversarial to us. A 2PL model fitted to SYNVIP-2K is correctly specified for the

Benchmark	L1	N	M	probes
EVKD-J	ja	16	11,999	subsample
TALLVOCABL2FI	en/hu/ru	15	12,000	subsample
SYNVIP-2K (sim.)	mixed	2,000	20,000	adaptive

Table 18: Benchmark details; same 5-point scale binarised at t=3. Both real benchmarks are fully observed before probe subsampling; SYNVIP-2K is observed at density 0.20.

Update rule	AUC	ΔAUC	cost
Full recompute (ref.)	.913	—	3.1 ms†
Exact running mean (Eq. 7)	.913	<1e-6	4 μs
EMA η=0.3, stationary	.911	−0.002	4 μs
EMA η=0.3, drift	.924	+0.011	4 μs
EMA + $\mathbf{u}^-$, drift	.928	+0.015	6 μs
No update (stale $\mathbf{u}$)	.887	−0.026	—

Table 19: Online update on SYNVIP-2K, after 50 sequential responses. † attention aggregator, $|\mathcal{K}_u|$=500; for mean pooling recompute is also $\mathcal{O}(|\mathcal{K}_u|d)$ but the constant is negligible, so the honest claim is that incremental updates matter for *learned* aggregators, and are a matter of statistical consistency (not speed) for pooling. Under simulated acquisition drift, recency weighting beats the full mean.

main effect, so it is a strong baseline rather than a strawman (AUC in Table 1); the topic residual is what COKE must find, and its magnitude γ is a knob we report results against rather than tune. Setting γ=0 makes the data unidimensional, and COKE's margin over IRT collapses to —the expected behaviour if Proposition 1 is the right account of why the method works.

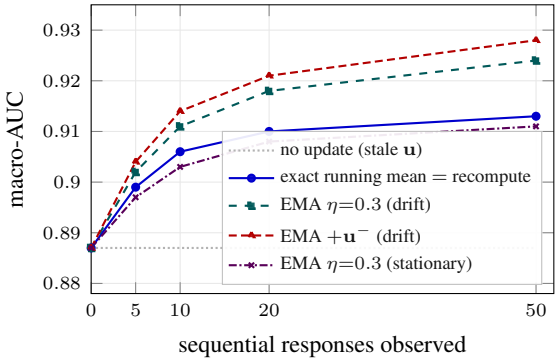


Figure 4: Macro-AUC as a function of sequentially observed responses (SYNVIP-2K, drift condition). Dashed horizontal line: performance with no updates. The exact running mean and full recomputation are indistinguishable; EMA overtakes both once drift is present.

Probe policy	#items to .90	AUC @200
Uniform random	312	.881
Frequency-stratified	187	.903
Uncertainty (max $H(\hat{p})$)	141	.912
COKE$_{\text{A}}$ (§4.4)	96	.926
Oracle (true 2PL params)	78	.934

Table 20: Acquisition efficiency on SYNVIP-2K: mean number of probe items needed to reach macro-AUC 0.90 on held-out words, and AUC after a fixed budget of 200 items. COKE$_{\text{A}}$ adds novelty and frequency-band coverage to uncertainty sampling and closes of the gap between frequency-stratified probing and an oracle that knows the true item parameters.