Better Budgeting for Author-Specific Handwritten Text Recognition: Characterizing Vision Language Model Adaptation Curves

Anonymous ACL submission

Abstract

Handwritten text recognition of historical manuscripts is highly accurate, but expensive to adapt to new authors. No principled basis exists for deciding how much annotation per author is needed to achieve a target character error rate (CER), leading to costly over-annotation. We sweep diverse training budgets for a state-of-the-art vision-language model and find that per-author adaptation curves follow a power law in the training budget. Consequently, we show that we can almost perfectly predict a new author's CER at any annotation budget, using only a zero-shot evaluation and a small initial training pilot (leave-one-out mean absolute error around 0.015). With these results, we can predict how much annotation each new author requires. We only need to calibrate on a few authors first, and can then choose the right amount of training budget for any target error rate. Our results could make it much cheaper and faster to digitize the large collections of handwritten documents.

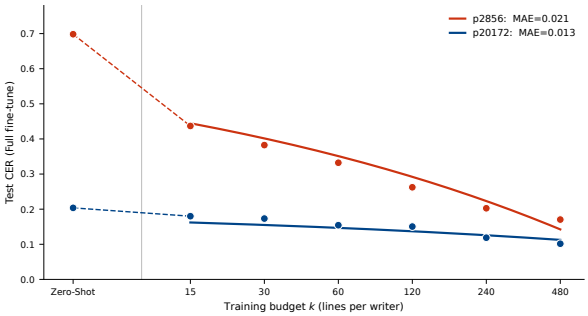

Figure 1: Predicted vs observed CER for two Bullinger authors, illustrating the prediction framework (log-scale): per-budget CER predicted from a 15-line pilot (lines) and the author's observed CER values (points).

1 Introduction

Handwritten manuscripts have been the default text format for most of human history, even after the invention of the printing press. They hold valuable evidence for fields from archaeology to economics and linguistics, and represent an untapped text resource that could expand the empirical basis of language modeling. If, that is, they can be transliterated into digital form.

Transliteration of handwritten text, aka handwritten text recognition (HTR), is fairly accurate, thanks to advances in OCR and language modeling. However, to adapt to new data, they still require costly annotation. Each new corpus, and especially each new author, requires expert paleographers to produce ground truth before a recognizer can be adapted to that author. With the multitude of different writing styles, fonts, and conventions, the per-author annotation cost varies widely: some reach usable recognition with few annotations, while others require annotation of substantially more samples before reaching acceptable error rates. Before committing to a labeling effort, practitioners therefore face a budgeting problem without a principled answer, with the default response – a uniform per-author budget – being wasteful on some authors and insufficient on others.

Vision-language models (VLMs) pretrained on historical text have changed what annotation means in this context. CHURRO-3B (Semnani et al., 2025), a 3-billion-parameter model fine-tuned on 155 historical corpora, transliterates many authors usefully out of the box. The practical question is no longer how to train a *recognizer* from scratch, but how much *additional annotation* each author needs on top of this baseline. The budgeting problem becomes quantitative: which authors are worth annotating heavily, and how much?

We provide a quantitative answer. We study 17 authors from the 16th-century Bullinger correspondence corpus (Scius-Bertrand et al., 2023) at training budgets up to 480 lines, with the IAM corpus (Marti and Bunke, 2002) as a secondary modern-handwriting validation set. We find that VLM adaptation curves follow a stable power law

in the training budget. Consequently, we can predict an author's CER at any candidate budget and at their full annotation pool, using only two measures, namely a zero-shot evaluation of a small set of authors and a small initial training pilot. We turn these findings into a practical workflow for budgeting annotation per author based on a required target CER, with calibration on a small initial set of target-dataset authors.

Contributions

1. We empirically show that on CHURRO-3B, the normalized per-line CER gain $b_w(k)$ decays as a power law in the training budget, $b_w(k) = c_w \cdot k^{-\gamma_w}$. The decay exponent γ_w is structurally consistent within each dataset.

2. We derive a prediction framework. The power law enables us to predict an author's CER at any candidate annotation budget and at their full annotation pool from a zero-shot evaluation and a small initial training pilot.

2 Datasets

Bullinger. The Bullinger correspondence corpus (Scius-Bertrand et al., 2023) contains 16th-century letters in Early New High German and Latin. We include all 20 authors that have at least 480 training lines annotated, with per-author training pools ranging from 485 to 4,051 lines (median 686). Validation and test splits per author follow the partitions defined in the original release. Ground truth comes from an automatic line alignment and carries documented errors in hyphenation, capitalization, abbreviation, and punctuation (Jungo et al., 2023). Three Bullinger authors produce catastrophic per-seed test CER (> 1.0) in over 25% of measured runs, with no other retained author showing a single such occurrence. We trace this to a decoding-time pathology on very short test references (§6). We excluded these authors from all analyses, leaving $n = 17$. Zero-shot CER on this subset ranges from 0.20 to 0.70.

IAM. The IAM database (Marti and Bunke, 2002) contains modern English handwriting from 657 authors. Existing partitions focus on author-independent recognition, so we construct a per-author split instead. Per-author training pools range from 15 to 494 lines (median 26). For each of the 38 authors with at least 51 available lines, we choose 18 lines each for validation and test. We chose this number via bootstrap analysis of zero-shot CER stability: the 18-line threshold lies past the diminishing-returns floor of the bootstrap CI on per-author ZS estimates. We exclude authors with zero-shot CER below 0.03, where per-author power-law fits are dominated by measurement noise, leaving $n = 23$ authors. Zero-shot CER on this subset ranges from 0.03 to 0.18.

We drop individual runs with test CER exceeding $2\times$ the cell median as decoding failures.

3 Methods

3.1 Training protocol

As base model, we use CHURRO-3B (Semnani et al., 2025), a Qwen2.5-VL-3B vision-language model fine-tuned on 155 historical corpora for HTR. We perform **full fine-tuning** of the entire model throughout our experiments, sweeping three learning rates $\{1\times10^{-5}, 2\times10^{-5}, 5\times10^{-5}\}$. Appendix A shows the detailed training setup. For each author, we uniformly sample training sets of size k from that author's available training pool ($k \in \{15, 30, 60, 120, 240, 480\}$ on Bullinger and $k \in \{2, 4, 8, 15\}$ on IAM). We vary data sampling by introducing seed counts per (author, k) cell, decreasing with k due to compute cost: Bullinger uses $\{8, 4, 4, 3, 3, 3\}$ seeds for each respective k; IAM uses 5 seeds per cell uniformly.

Learning rate selection. Rather than selecting learning rates per author, which risks per-author overfitting and conflates its choice with the adaptation behavior we study, we select a single learning rate per k on a held-out set of 6 hyperparameter-tuning authors per dataset, disjoint from the test authors and stratified along the zero-shot CER difficulty axis to ensure coverage of the per-author ZS distribution. Full fine-tuning performs best with a learning rate of 2×10^{-5} in our sweeps, see Appendix B for more information.

Evaluation metric. We report character error rate (CER) as the per-line Levenshtein distance normalized by reference length, averaged across the author's test lines. Cell-level CER is the median across seeds.

3.2 Per-author fits and regression models

Per-author power-law fits. For each author we fit the normalized per-line CER gain

$$b_w(k) = \frac{\mathrm{ZS}_w - \mathrm{CER}_w(k)}{k \cdot \mathrm{ZS}_w}$$

as a power law in the training budget,

$$b_w(k) = c_w \cdot k^{-\gamma_w}, \tag{1}$$

by nonlinear least squares in CER space on the closed-form rearrangement $\mathrm{CER}_w(k) = \mathrm{ZS}_w(1 - c_w k^{1-\gamma_w})$.

Cross-author regression. For prediction, we treat γ as a dataset-level constant, obtained by jointly fitting the closed form across all authors of a dataset with a single shared γ (§4.1). With γ pinned at this pooled value, we refit each author's $\log_{10} c_w$ and regress it across authors on $(\mathrm{ZS}_w, \mathrm{EGF}_w)$ by ordinary least squares:

$$\log_{10} c_w = \beta_0 + \beta_1 \mathrm{ZS}_w + \beta_2 \mathrm{EGF}_w \qquad (2)$$

where

$$\mathrm{EGF}_w = (\mathrm{ZS}_w - \mathrm{CER}_w(k_0))/\mathrm{ZS}_w \qquad (3)$$

is the early-gain fraction at the pilot budget ($k_0 = 15$ on Bullinger, $k_0 = 2$ on IAM).

Predictor. For a new author, Equation 2 predicts $\hat{c}_w = 10^{\hat{\beta}_0 + \hat{\beta}_1 \mathrm{ZS}_w + \hat{\beta}_2 \mathrm{EGF}_w}$ from one zero-shot pass and one small training pilot at k_0. Plugging $\hat{c}_w$ and the pooled γ into the closed form yields

$$\widehat{\mathrm{CER}}_w(k) = \mathrm{ZS}_w \left(1 - \hat{c}_w\, k^{1-\gamma}\right), \qquad (4)$$

predicting per-author CER at any k.

Evaluation. For each held-out author w, we refit Equation 2's coefficients on the remaining $n-1$ authors (with γ held at the full-sample pooled value) and evaluate Equation 4 at each measured budget. Leave-one-out mean absolute error is reported in CER units.

4 Results

4.1 Per-author power-law scaling

Per-author adaptation curves on CHURRO-3B follow the power law of Equation 1. Fitting per-author with γ_w free yields median CER-R^2 of 0.97 on Bullinger and 0.78 on IAM (detailed Figure in Appendix C). IAM's lower fit quality reflects its narrower budget grid (4 measurements per author across $k \in \{2, 4, 8, 15\}$ vs. 6 across $k \in \{15, \ldots, 480\}$ on Bullinger), which leaves per-author fits more sensitive to measurement noise. On Bullinger, $b_w(k)$ falls $\approx 13\times$ between $k = 15$ and $k = 480$, the first measured budget is by far the most cost-efficient per line. Since marginal returns collapse rapidly, per-author annotation must be planned upfront rather than extrapolated naively from early gains.

Bullinger ($n = 17$)	MAE
$k = 30$	0.015
$k = 60$	0.014
$k = 120$	0.019
$k = 240$	0.021
$k = 480$	0.021
Oracle	0.029

IAM ($n = 23$)	MAE
$k = 4$	0.011
$k = 8$	0.008
$k = 15$	0.011
Oracle	0.016

Table 1: Leave-one-out mean absolute error (CER units) of the Power-Law predictor (Equation 4). Pilot budgets ($k_0 = 15$ on Bullinger, $k_0 = 2$ on IAM) are excluded because EGF_w is measured at k_0, making the prediction there not independent of the calibration. Oracle reports MAE at each author's full annotation pool.

Cross-author γ consistency. Per-author γ_w is concentrated within each dataset, more tightly on Bullinger than on IAM: Bullinger median 0.77 (IQR width 0.085), IAM median 0.53 (IQR width 0.68, about $8\times$ Bullinger's). Joint NLS over all authors yields pooled $\gamma_{\text{Bull}} = 0.79$ and $\gamma_{\text{IAM}} = 0.70$, both within the per-author IQR on the respective dataset. We treat γ as a dataset-level constant in what follows; per-author variation is carried by c_w alone (§4.2).

4.2 Per-author CER prediction

With γ pinned at the dataset-pooled value, the per-author trajectory reduces to a single free parameter c_w. The cross-author regression of $\log_{10} c_w$ on $(\mathrm{ZS}_w, \mathrm{EGF}_w)$ (Equation 2) achieves leave-one-out R^2 of 0.82 on Bullinger and 0.53 on IAM. EGF dominates the regression on both datasets: regressing $\log c_w$ on ZS_w alone increases pooled per-budget LOO MAE from 0.018 to 0.029 on Bullinger (+63%) and from 0.010 to 0.019 on IAM (+88%).

End-to-end prediction. Equation 4 achieves pooled leave-one-out MAE of 0.018 on Bullinger and 0.010 on IAM across non-pilot budgets; at each author's full annotation pool, oracle MAE is 0.029 on Bullinger and 0.016 on IAM (Table 1).

Calibration cost. Calibrating the predictor end-to-end – jointly pooling γ across calibration authors and fitting Equation 2 on the resulting $(\mathrm{ZS}_w, \mathrm{EGF}_w, \log_{10} c_w)$ triples – requires $n_{\text{cal}} = 10$ authors to recover $\geq 85\%$ of asymptotic per-budget accuracy on both datasets and of asymptotic oracle accuracy on Bullinger (Appendix D).

4.3 Breakdown and remedy

The closed-form predictor has a theoretical limit: because $1 - \gamma > 0$, the factor $k^{1-\gamma}$ in Equation 4 grows unboundedly and $\widehat{\mathrm{CER}}_w(k)$ eventually crosses zero past a per-author breakdown threshold k_w^*. On Bullinger, k_w^* under the calibrated regression has median $\approx$ 2500 lines; three of seventeen authors have annotation pools beyond their k_w^*, where the closed-form oracle prediction is no longer valid. For these regimes we substitute a direct OLS regression

$$\begin{aligned}\mathrm{oracle}_w = \alpha_0 + \alpha_1\,\mathrm{ZS}_w + \alpha_2\,\mathrm{EGF}_w \\ +\alpha_3 \log_{10} n_{\mathrm{pool},w}\end{aligned} \tag{5}$$

which avoids k-extrapolation and achieves LOO MAE 0.019 (vs. 0.029 for the closed form) on Bullinger. Within the calibration window the closed form remains preferred for its continuous k-dependence; Equation 5 is the recommended substitute when the planned annotation pool sits near or past k_w^*.

5 Related Work

Per-author adaptation has been studied via several routes. Closest to our setup are supervised fine-tuning approaches: Aradillas et al. (2021) boost recognizers with few labeled lines; Kohút and Hradiš (2025) show that fine-tuning is a surprisingly effective HTR adaptation baseline; Pippi et al. (2023) study how to select a pretrained recognizer for single-author fine-tuning. None of these characterize the *form* of the per-author trajectory or provide a budgeting protocol.

Vision-language models have extended HTR beyond corpus-specific recognizers. TrOCR (Li et al., 2022) combines a vision Transformer encoder with a pre-trained text decoder. CHURRO (Semnani et al., 2025) is a 3B-parameter VLM, based on Qwen 2.5 VL (Bai et al., 2025), fine-tuned on 155 historical corpora spanning 22 centuries and 14 script families – we use it as our base model.

Neural scaling laws describe how model performance scales with parameters, data, and compute. Hestness et al. (2017) documented power-law learning curves across modalities; Kaplan et al. (2020) and Hoffmann et al. (2022) (Chinchilla) established compute-optimal scaling for language model pretraining. Hernandez et al. (2021) extended scaling-law analysis to transfer learning, showing that pretraining effectively multiplies fine-tuning dataset size in a power-law form.

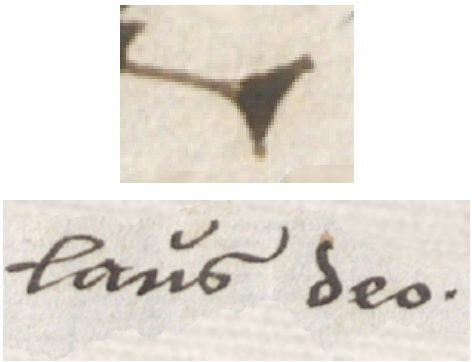

Figure 2: Examples of test images for p1548. The top line causes catastrophic generation with CER far above 1.0 (top), while CHURRO decodes the bottom one correctly.

6 Conclusion

The empirical regularities of §4 combine into a practical workflow for budgeting per-author annotation. Once the predictor is calibrated on $n_{\mathrm{cal}} \approx 10$ authors (§4.2), deployment to each new author involves:

1. Run zero-shot evaluation to obtain ZS_w.
2. Annotate k_0 lines, train, record the $\mathrm{CER}_w(k_0)$, and compute EGF_w (Equation 3).
3. Predict the per-author scale $\hat{c}_w$ via Equation 2.
4. For a target CER T, use Equation 4 to find the smallest sufficient training budget k_w. If k_w exceeds the writer's annotation pool, the target is unreachable from the writer's data; deprioritize. When the pool itself exceeds k_w^* (§4.3), use Equation 5 to predict the achievable pool-size CER.

On the test images causing the exclusions of Bullinger authors – small images with minimal character content (canonical references "S", "S.", "."; Figure 2) – CHURRO falls back to long autoregressive completions that inflate per-line CER far above 1.0. Practitioners should pre-screen test sets for very short or low-content references and treat per-line CER above 1.0 as a signal of decoding failure rather than a legitimate measurement.

The workflow applies wherever per-author adaptation curves follow the power law characterized in §4.1. We validate on 16th-century Bullinger correspondence and modern IAM handwriting; broader validation across script families and base models is left to future work.

7 Limitations

Scope of evaluation. Our findings are scoped to a single base model (CHURRO-3B) and two datasets: the 16th-century Bullinger correspondence (Early New High German and Latin) and modern English IAM. We have not validated on other base models, scripts (e.g. Greek, Hebrew, Arabic), languages, or historical periods. Generalization beyond these settings and cross-dataset transfer of power law coefficients is an open question.

Pathology mechanism. We do not isolate the mechanism behind the short-reference failure characterized in §6. We attempted several characterization experiments without isolating a clear cause. The mechanism's opacity means we cannot predict in advance which corpora will exhibit this failure mode – pre-screening test sets for short low-content references is currently the only available safeguard.

Workflow validation. The practitioner workflow proposed in §6 is derived from the regressions characterized in §4, but has not been validated end-to-end in a deployment setting. Reported MAE values are leave-one-out estimates of prediction quality; deployment-grade validation would require running the workflow on a held-out corpus and using its predictions for actual budgeting decisions.

Sample sizes and empirical asymptote. Sample sizes are modest: 17 authors on Bullinger, 23 on IAM. Replication on larger author panels would strengthen the headline findings. The "oracle" CER throughout is the empirical CER at the author's full available pool, not a true mathematical asymptote; authors with substantially larger annotation budgets could reveal further descent.

References

Jose Carlos Aradillas, Juan Jose Murillo-Fuentes, and Pablo M. Olmos. 2021. Boosting offline handwritten text recognition in historical documents with few labeled lines. *IEEE Access*, 9:76674–76688.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *Preprint*, arXiv:2102.01293.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *Preprint*, arXiv:1712.00409.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Michael Jungo, Lars Vögtlin, Atefeh Fakhari, Nathan Wegmann, Rolf Ingold, Andreas Fischer, and Anna Scius-Bertrand. 2023. Impact of the ground truth quality for handwriting recognition. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT 2023, page 135–140. ACM.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Jan Kohút and Michal Hradiš. 2025. Fine-tuning is a surprisingly effective domain adaptation baseline in handwriting recognition. *Preprint*, arXiv:2302.06308.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. Trocr: Transformer-based optical character recognition with pre-trained models. *Preprint*, arXiv:2109.10282.

U.-V. Marti and H. Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Vittorio Pippi, Silvia Cascianelli, Christopher Kermorvant, and Rita Cucchiara. 2023. How to choose pretrained handwriting recognition models for single writer fine-tuning. *Preprint*, arXiv:2305.02593.

Anna Scius-Bertrand, Phillip Ströbel, Martin Volk, Tobias Hodel, and Andreas Fischer. 2023. The bullinger dataset: A writer adaptation challenge. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part I*, page 397–410, Berlin, Heidelberg. Springer-Verlag.

Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 2025. Churro: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34765–34812.

A Detailed training protocol

Setting	Value
Base model	CHURRO-3B
Fine-tuning	Full (all parameters)
Optimizer	AdamW
Learning rates	$\{1, 2, 5\} \times 10^{-5}$
LR schedule	Cosine, 10% warmup
Batch size	4
Epochs	10
`max_new_tokens`	1024
Fixed mechanical seed	42
Bullinger k	$\{15, 30, 60, 120, 240, 480\}$
Bullinger data seeds	$\{8, 4, 4, 3, 3, 3\}$
IAM k	$\{2, 4, 8, 15\}$
IAM data seeds	5 (uniform)
GPU	H200
`transformers`	5.5.0
`torch`	2.11.0

Table 2: Training setup.

B Hyperparameter sweep: median validation CER per learning rate

The learning rate (LR) for each training budget k was chosen by a sweep on a separate set of 6 HP-tuning writers per dataset that does not overlap with the production sets reported in the main text. For each (writer, k, LR) cell we trained with 1–8 seeds; for each seed we took the best CER across epochs, then took the median across seeds. The values reported in Table 3 are the median of those per-writer values across the 6 HP writers (rounded to four decimals). The production LR for each k is the LR that minimises this median; on both datasets this selects $\mathrm{LR} = 2{\times}10^{-5}$ at every k for full-FT.

k	LR=$1{\times}10^{-5}$	LR=$2{\times}10^{-5}$	LR=$5{\times}10^{-5}$
Bullinger, full-FT			
15	0.3052	**0.2785**	0.3438
30	0.2821	**0.2534**	0.2779
60	0.2637	**0.2219**	0.2254
120	0.2259	**0.1781**	0.1813
240	0.1882	**0.1431**	0.1526
480	0.1499	**0.1272**	0.1442
IAM, full-FT			
2	0.0349	**0.0311**	0.0564
4	0.0407	**0.0344**	0.0552
8	0.0395	**0.0262**	0.0393
15	0.0388	**0.0195**	0.0401

Table 3: HP-sweep median validation CER per (k, LR) for full-FT, computed from the raw HP-sweep logs across all 6 HP writers. The production LR for each k is the per-row minimum (in **bold**); on both datasets this yields $\mathrm{LR} = 2{\times}10^{-5}$ uniformly.

C Per-author power law fit

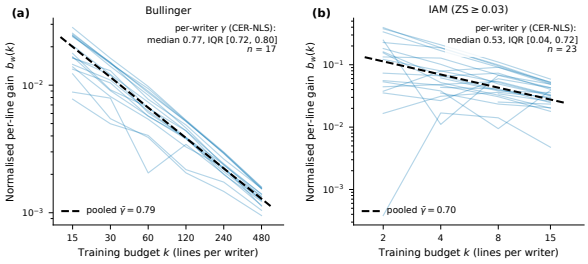

Figure 3: Per-writer power-law collapse for Bullinger (n=17) and IAM (n=23, ZS $\geq$ 0.03) under full fine-tuning. Each line shows the normalized per-line gain $b_w(k) = (\mathrm{ZS} - \mathrm{CER}(k))/(k \cdot \mathrm{ZS})$ for one writer on a log-log scale; k-points where $b_w \leq 0$ are omitted. The dashed reference line is a joint NLS fit of the closed form with a single shared γ across writers. Near-parallel alignment across writers supports a universal slope γ: Bullinger median $\hat{\gamma} = 0.77$ (IQR $[0.72, 0.80]$); IAM median $\hat{\gamma} = 0.53$ (IQR $[0.04, 0.72]$).

D Calibration cost

For each candidate calibration size n_{cal} we sample n_{cal} writers uniformly at random (without replacement) from the dataset, run the full fixed-γ workflow on the sampled subset, and report the leave-one-out (LOO) MAE on the $n_{\mathrm{full}} - n_{\mathrm{cal}}$ held-out writers. Each row averages $B = 200$ bootstrap subsamples (writers re-sampled, held-out writers re-evaluated). The reference row $n_{\mathrm{full}}{-}1$ is the within-dataset LOO baseline (the asymptotic best achievable).

The workflow is the joint-calibration form: from the same n_{cal} writers we (i) estimate the dataset-universal γ_{pooled} by joint NLS minimisation of $\sum_{w,k}\big(\mathrm{CER}_w(k) - \mathrm{ZS}_w(1 - 10^{\log c_w}\, k^{1-\gamma})\big)^2$ in one shared γ and n_{cal} per-writer $\log c_w$; (ii) refit $\log c_w$ per writer with γ held at the pooled value; (iii) fit the OLS regression $\log c_w \sim \beta_0 + \beta_1\, \mathrm{ZS}_w + \beta_2\, \mathrm{EGF}_w$ on the same n_{cal} writers; (iv) for each held-out writer predict $\widehat{\log c}$ from the regression and recover $\widehat{\mathrm{CER}}(k) = \mathrm{ZS}\left(1 - 10^{\widehat{\log c}}\, k^{1-\gamma_{\mathrm{pooled}}}\right)$ at every measured budget and at $k = n_{\mathrm{pool}}$ for the oracle column.

Bullinger (n_{full}=17, γ_{pooled}=0.79)							
n_{cal}	k=15	k=30	k=60	k=120	k=240	k=480	oracle
4	0.0228	0.0274	0.0288	0.0347	0.0392	0.0453	0.0591
5	0.0176	0.0213	0.0215	0.0261	0.0291	0.0332	0.0467
7	0.0137	0.0168	0.0157	0.0208	0.0236	0.0264	0.0367
10	0.0127	0.0154	0.0143	0.0189	0.0217	0.0235	0.0337
13	0.0119	0.0138	0.0135	0.0182	0.0214	0.0232	0.0313
16 (ref.)	0.0117	0.0146	0.0135	0.0192	0.0208	0.0209	0.0292

IAM, ZS ≥ 0.03 (n_{full}=23, γ_{pooled}=0.70)					
n_{cal}	k=2	k=4	k=8	k=15	oracle
4†	—	—	—	—	—
5	0.0197	0.0256	0.0307	0.0437	0.0797
7	0.0111	0.0135	0.0118	0.0154	0.0289
10	0.0104	0.0125	0.0101	0.0132	0.0235
13	0.0096	0.0114	0.0091	0.0126	0.0190
17	0.0096	0.0115	0.0082	0.0115	0.0177
22 (ref.)	0.0089	0.0104	0.0086	0.0122	0.0155

Table 4: Joint-calibration LOO MAE on held-out writers as a function of the calibration sample size n_{cal}. Both per-budget MAE and full-pool oracle MAE are shown; all values in CER units, averaged over 200 bootstrap samples. The reference row ($n_{\text{full}}-1$) is the within-dataset LOO baseline. † On IAM, n_{cal}=4 is rank-deficient for the 3-parameter $\log c \sim \text{ZS} + \text{EGF}$ regression on the narrow IAM ZS range, with held-out MAEs in the 10^4 range; the row is omitted.