PTPP-Aware Adaptation Scaling Laws: Predicting Domain-Adaptation Performance at Unseen Pre-Training Budgets

Etienne Goffinet
Cerebras Systems
etienne.goffinet@cerebras.net

Shane Bergsma
Cerebras Systems
shane.bergsma@cerebras.net

Avraham Sheinin
Cerebras Systems

Natalia Vassilieva
Cerebras Systems

Shaheer Muhammad
Cerebras Systems

Preslav Nakov
MBZUAI, Abu Dhabi
Abu Dhabi, UAE

Gurpreet Gosal
Cerebras Systems
gurpreet.gosal@cerebras.net

Abstract

Continual pre-training (CPT) for domain adaptation must balance target-domain gains with stability on the base domain. Existing CPT scaling laws typically assume a fixed pre-training budget, which limits their ability to forecast adaptation outcomes for models trained at different tokens-per-parameter (PTPP). We present *PTPP-aware* adaptation scaling laws that make the pre-training budget an explicit variable, enabling accurate *prediction* of adaptation loss at unseen P_{TPP}. On a multilingual setup (English/Arabic → French), PTPP-aware formulations trained on early stages (P_{TPP}={15,31}) predict target loss at P_{TPP}=279 and outperform a PTPP-agnostic D-CPT transfer baseline on metrics (Huber-on-log, MAE_{rel}, calibration slope); full diagnostics (RMSE, MAPE) are in the appendix. Beyond forecasting, we show a practical use case: planning replay ratios and adaptation token budgets that satisfy target and forgetting constraints under compute limits.

1 Introduction

Capabilities of LLMs (large language models) continue to scale with model size, data size, and thus the total compute used for pre-training. Language models trained on a mixture of domains, dominated by web-scale corpora, yield general LLMs [Biderman et al., 2023, Dey et al., 2023, Team et al., 2025, OLMo et al., 2025, Yang et al., 2025]. These generalist models may not perform well in tasks requiring specialized knowledge (e.g., in fields such as medicine, law, finance) or those requiring language capabilities beyond the dominant pre-training language. We must therefore *adapt* these models to new, domain-specific, or target-language-specific data. This adaptation process presents a fundamental challenge: achieving strong performance in the target domain while preserving general capabilities (avoiding catastrophic forgetting [Kirkpatrick et al., 2017]). Various strategies have been proposed to minimize forgetting [Chen et al., 2025, Ostapenko et al., 2022, Biderman et al., 2024, Ibrahim et al., 2024, Gupta et al., 2023].

Pre-training scaling laws are well established—e.g., relations between model/data size and performance in Kaplan et al. [2020], Hoffmann et al. [2022]—whereas CPT-specific laws are comparatively underexplored. D-CPT extends Chinchilla with *replay* to study compute-optimal CPT at a fixed pre-training stage [Que et al., 2024], and forgetting laws quantify degradation on the pre-training domain at that stage [Bethune et al., 2025].

However, most CPT scaling laws—D-CPT and forgetting laws included—assume a *fixed* pre-training budget (a single PTPP stage), which limits *forecasting* across budgets. Prior work indicates that PTPP modulates learning dynamics and downstream adaptation [Springer et al., 2025, Ash and Adams, 2020, Lyle et al., 2023, Kumar et al., 2024]. We therefore condition explicitly on PTPP, yielding *PTPP-aware* adaptation laws that predict target-domain loss at unseen PTPP and clarify replay–stage interactions. Our central question: can a law fit at early stages (PTPP={15,31}) *forecast* target validation loss at PTPP=279?

Although our experiments focus on language adaptation, treating PTPP as an explicit driver of adaptation dynamics is broadly applicable. Prior multilingual adaptation work underscores the need to mitigate and estimate forgetting while acquiring target competence [de Vries and Nissim, 2021, Fujii et al., 2024, Huang et al., 2024, Gosal et al., 2024, Zhao et al., 2024]. Our contributions include:

1. **PTPP-aware adaptation laws.** We extend CPT scaling laws by integrating the pre-training budget (PTPP) as an explicit variable in the functional form.
2. **Forecasting at unseen PTPP.** Fits at PTPP=15,31 predict French loss at PTPP=279 and outperform a PTPP-agnostic D-CPT baseline on all metrics; a handful of 241M-scale "anchor" points at PTPP=279 (20 calibration measurements at the evaluation stage) further improve accuracy at low-cost.
3. **Planning under constraints.** Using the fitted law, we find an optimal replay ratio and adaptation token budget that satisfy target and forgetting constraints under compute limits.

2 Methodology and Experiments

Setup. We study loss $\hat{L}$ as a function of model size N, adaptation tokens D, replay ratio $r \in (0,1]$ (s.t. $1-r$ is the target domain fraction), and pre-training PTPP. We use GPT-2–style decoder-only models pre-trained on a mixed English–Arabic corpus; the *adaptation domain* is French. Fits use PTPP={15,31} and are *evaluated* on PTPP=279 (unseen), across $r \in \{0.10, 0.25, 0.50\}$ and $N \in$ {241M, 517M, 1.4B, 8.1B}.

PTPP-Aware candidate formulations (1–3). All laws share an N-term and an r-barrier; they differ in how PTPP affects the data-efficiency term. Let $\varepsilon = 10^{-5}$.

(1) Additive PTPP prior (floor).

$$\hat{L} \;=\; E \;+\; \frac{A}{N^{\alpha}} \;+\; \frac{B\,r^{\nu}}{D^{\beta}} \;+\; \frac{C}{(r+\varepsilon_r)^{\gamma}} \;+\; \frac{F}{P\mathrm{TPP}^{\eta}}.$$

PTPP *lowers the* floor *of $\hat{L}$ via an additive term.*

(2) PTPP-gated data exponent (no floor).

$$\hat{L} \;=\; E \;+\; \frac{A}{N^{\alpha}} \;+\; \frac{B\,r^{\nu}}{D^{\beta_{\text{eff}}}} \;+\; \frac{C}{(r+\varepsilon_r)^{\gamma}}, \qquad \beta_{\text{eff}} = \beta\left(1-\lambda\frac{P\mathrm{TPP}^{\zeta}}{1+P\mathrm{TPP}^{\zeta}}\right), \;\; \beta_{\text{eff}} \geq 10^{-6}.$$

PTPP *controls the* shape *of the data law via a bounded gate $\lambda \frac{P\mathrm{TPP}^{\zeta}}{1+P\mathrm{TPP}^{\zeta}} \in [0,\lambda)$, so $\beta_{\text{eff}} = \beta\,(1-g(P\mathrm{TPP}))$, representing the impact of pre-training budget on adaptation efficiency.*

(3) PTPP-gated data exponent + floor.

$$\hat{L} \;=\; E \;+\; \frac{A}{N^{\alpha}} \;+\; \frac{B\,r^{\nu}}{D^{\beta_{\text{eff}}}} \;+\; \frac{C}{(r+\varepsilon_r)^{\gamma}} \;+\; \frac{F}{P\mathrm{TPP}^{\eta}}, \qquad \beta_{\text{eff}} = \beta\left(1-\lambda\frac{P\mathrm{TPP}^{\zeta}}{1+P\mathrm{TPP}^{\zeta}}\right).$$

PTPP *acts* twice*: (i) a bounded gate reshapes the D-response (as in Form 2), and (ii) an additive prior $F/P\mathrm{TPP}^{\eta}$ lowers the loss floor. This captures both shape and offset effects of pre-training.*

Anchors. We also report a few-shot variant that augments the fit with 20 small-scale (241M) *anchors* collected at the evaluation stage (PTPP=279) across the (r, D) grid; all other PTPP=279 points remain held out (unlike the oracle, which fits on the full PTPP=279 set). These anchors tighten calibration (slope $\to 1$) and error metrics at low-cost.

Data & models. GPT-2–style decoders are pre-trained on English/Arabic (source) and adapted to French (target). We consider PTPP$\in \{15, 31, 279\}$; replay $r \in \{0.10, 0.25, 0.50\}$; and model sizes $\{$241M, 517M, 1.4B, 8.1B$\}$. We focus on the French target; source-domain (English/Arabic) results are deferred to the appendix.

Fitting constraints. We minimize Huber loss on log residuals (δ=0.02) with L–BFGS–B under positivity constraint for all parameters except $\zeta \in \mathbb{R}$. We clip $r \in [10^{-9}, 1 - 10^{-9}]$.

Metrics. We assess predictions at PTPP=279 with three metrics. *Huber-on-log* is the Huber loss to residuals $r = \log \hat{y} - \log y$ with $\delta = 0.02$. *MAE$_{rel}$* is the mean absolute relative error $\frac{1}{n}\sum_i \frac{|\hat{y}_i - y_i|}{y_i}$, i.e., the typical percentage miss (lower is better). *Calibration (intercept/slope)*: parameters (a, b) from an Ordinary Least Squares (OLS) fit $\log y = a + b \log \hat{y}$; ideal is $a \approx 0$, $b \approx 1$.

3 Results: Forecasting at Unseen PTPP

Formulation	Huber$_{\log}\downarrow$	MAE$_{\text{rel}}\downarrow$	Calib. slope ≈ 1
Form 1 (Additive Prior)	2.34×10^{-4}	2.08×10^{-2}	0.991
Form 2 (Gated Exponent)	1.99×10^{-4}	1.83×10^{-2}	0.970
Form 3 (Gated+Floor)	4.43×10^{-5}	6.70×10^{-3}	0.991
D-CPT (no PTPP, transfer)	4.74×10^{-4}	3.43×10^{-2}	0.961

Table 1: French prediction at unseen PTPP=279 (no anchors; trained on PTPP=$\{$15,31$\}$). Full metrics (including calibration intercepts and RMSE) appear in Appendix 5.

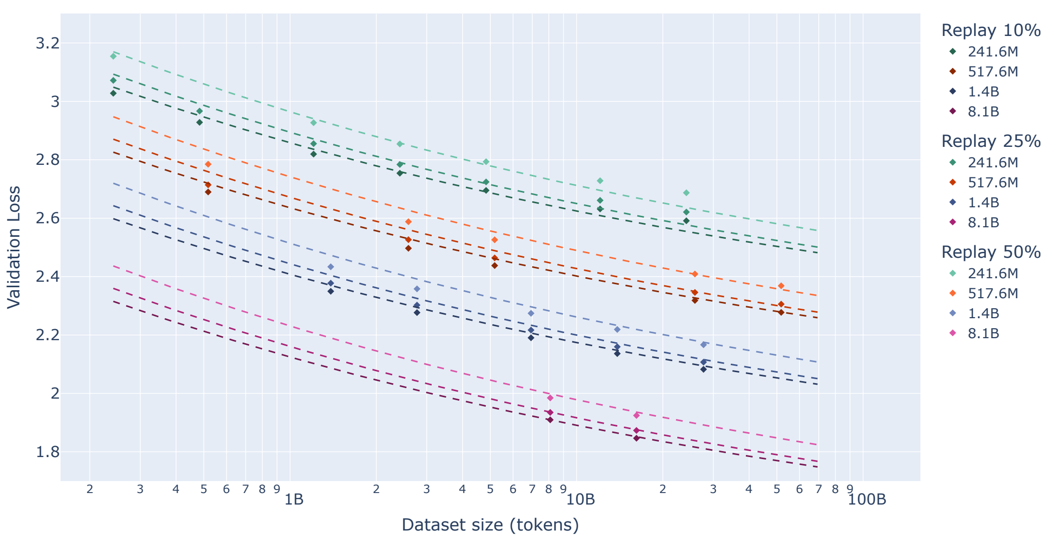

Figure 1: PTPP=279 *predictions* of the *gated+floor* model (dashed) vs. *observations* (markers) of validation loss for $r \in \{0.10, 0.25, 0.50\}$ and $\{$241M, 517M, 1.4B, 8.1B$\}$.

Formulation	$\text{Huber}_{\log}\downarrow$	$\text{MAE}_{\text{rel}}\downarrow$	Calib. slope ≈ 1
Form 1 (Additive Prior)	5.22×10^{-5}	8.56×10^{-3}	0.956
Form 2 (Gated Exponent)	4.23×10^{-5}	8.17×10^{-3}	0.992
Form 3 (Gated+Floor)	3.54×10^{-5}	7.39×10^{-3}	0.992

Table 2: French prediction at unseen $P\text{TPP}$=279 with 20 anchors at 241M-scale.

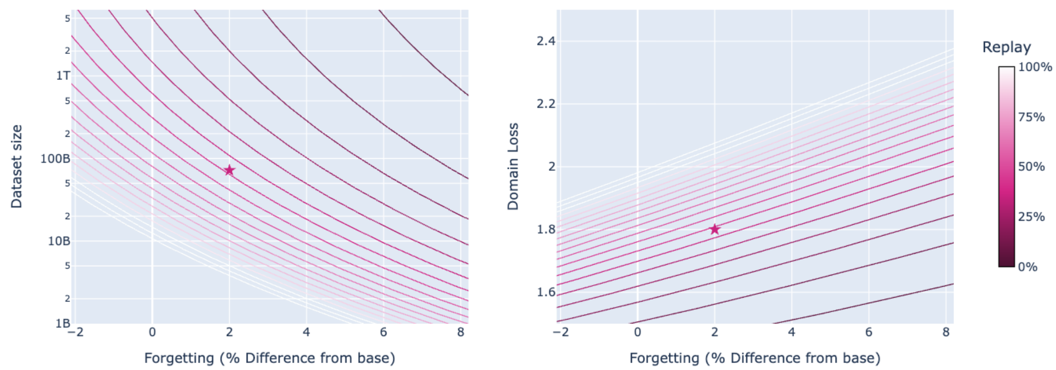

Figure 2: Replay (0–100%) determines the trade-off between forgetting and domain performance. Left: Forgetting / Dataset size landscape. Right: resulting French loss. The star highlights the solution (8.9 $A\text{TPP}$, 34% replay), minimizing FLOPs s.t. forgetting is $\leq$+2% and French loss $\leq$ 1.8.

Takeaways. Across French at unseen $P\text{TPP}$=279, the *gated+floor* variant (Form 3) is consistently best, with low errors and near-ideal calibration (slope $\approx$0.99) both without anchors and with 20 small-scale anchors; the *gated-only* variant (Form 2) is reliably second and ahead of D-CPT transfer. Anchors uniformly tighten Huber/RMSE and calibration without changing the methods' rankings.

On the English/Arabic source domain (Appendix 5), the picture depends on supervision at the evaluation stage: without anchors, *floor-only* (Form 1) suffices—suggesting $P\text{TPP}$ mainly shifts the baseline—whereas with anchors the *gated-only* form (Form 2) becomes best, revealing a data-efficiency (shape) effect once lightly calibrated at $P\text{TPP}$=279. Overall, results support that a) the preferred functional form can be domain and/or supervision-dependent and b) a direct link exists between pre-training compute and adaptation efficiency that manifests as both a floor shift and a learning-curve shape change; few-shot anchors prove to be a low-cost way to calibrate the latter.

4 Use Case: Joint Compute and Replay Optimization

In domain adaptation, one must balance *forgetting* of the source domain with improvements in the target domain, under strict compute budgets. A key feature of our method is that ptpp-aware scaling-law fits allow prediction of both losses at an unseen $P\text{TPP}$ (279), making it possible to solve this trade-off analytically rather than through brute-force sweeps. We consider a target model scale of $N = 8.1$B, pretrained at $P\text{TPP} = 279$, and seek the smallest adaptation tokens-per-parameter ($A\text{TPP}$) that meets the forgetting and target-performance constraints, under the Form 1 hypothesis for English/Arabic loss and Form 3 for French. Let the adaptation budget $A\text{TPP} = D/N$. We solve:

$$\min_{A\text{TPP}\geq 0,\ r\in[0,1]} \quad A\text{TPP} \quad \text{s.t.} \quad \Delta L_{\text{src}}\big(N,\ D,\ r,\ 279\big) \leq \delta,\ \ L_{\text{tgt}}\big(N,\ D,\ r,\ 279\big) \leq \tau.$$

where $\Delta L_{\text{src}} = L_{\text{src}}(N, D, r, P\text{TPP}) - L_{\text{src}}(N, 0, 1, P\text{TPP})$, N is model size, and $r \in [0, 1]$ the replay ratio. Constraints are given by tolerated forgetting δ (e.g. +2%) and target French loss threshold τ=1.8. The optimal solution, displayed on Fig. 2, is $A\text{TPP} = 8.9$ and replay 34%.

5 Conclusion

We proposed *PTPP-aware* adaptation scaling laws that condition on the pre-training budget and *predict* target performance at unseen P_{TPP}. On French at P_{TPP}=279, laws fit at early stages ($P_{\text{TPP}} \in \{15, 31\}$) generalize well and outperform a P_{TPP}-agnostic D-CPT transfer baseline; a small set of 241M-scale anchors further improves accuracy. Empirically, pre-training progress modulates both the *loss floor* and the adaptation efficiency, and a few low-cost anchors further enhance the prediction performances; on the source domain, floor shifts explain most gains without anchors, while the light anchoring reveals a data-dependent effect at P_{TPP}=279. These fits enable the optimization of replay and adaptation tokens under compute constraints. Promising directions include investigating how language transfer shapes the P_{TPP} effect, extending to additional P_{TPP} stages and domains, assessing task-level metrics, and adding uncertainty quantification with cost-aware anchor selection.

References

Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in Neural Information Processing Systems*, 33:3884–3894, 2020.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection, 2025. URL `https://arxiv.org/abs/2502.06042`.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less, 2024. URL `https://arxiv.org/abs/2405.09673`.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL `https://arxiv.org/abs/2304.01373`.

Yupeng Chen, Senmiao Wang, Yushun Zhang, Zhihang Lin, Haozhe Zhang, Weijian Sun, Tian Ding, and Ruoyu Sun. MoFO: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning, 2025. URL `https://arxiv.org/abs/2407.20999`.

Wietse de Vries and Malvina Nissim. As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, page 836–846. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.74. URL `http://dx.doi.org/10.18653/v1/2021.findings-acl.74`.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. Cerebras-GPT: Open compute-optimal language models trained on the Cerebras wafer-scale cluster, 2023. URL `https://arxiv.org/abs/2304.03208`.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities, 2024. URL `https://arxiv.org/abs/2404.17790`.

Gurpreet Gosal, Yishi Xu, Gokul Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming, Chen, Biswajit Mishra, Natalia Vassilieva, Joel Hestness, Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Onkar Pandit, Satheesh Katipomu, Samta Kamboj, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Soundar Doraiswamy, Mohamed El Karim Chami, and Preslav Nakov. Bilingual adaptation of monolingual foundation models, 2024. URL `https://arxiv.org/abs/2407.12869`.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pre-training of large language models: How to (re)warm your model?, 2023. URL `https://arxiv.org/abs/2308.04014`.

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, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL `https://arxiv.org/abs/2203.15556`.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. AceGPT, localizing large language models in arabic, 2024. URL `https://arxiv.org/abs/2309.12053`.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models, 2024. URL `https://arxiv.org/abs/2403.08763`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL `https://arxiv.org/abs/2001.08361`.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Ãgnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. *arXiv preprint arXiv:2303.01486*, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2025. URL `https://arxiv.org/abs/2501.00656`.

Oleksiy Ostapenko, Timothee Lesort, Pau Rodríguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay, 2022. URL `https://arxiv.org/abs/2205.00329`.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, et al. D-CPT law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37: 90318–90354, 2024.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune, 2025. URL `https://arxiv.org/abs/2503.19206`.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe

Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. LLaMA beyond English: An empirical study on language capability transfer, 2024. URL https://arxiv.org/abs/2401.01055.

Appendix A: Full metric tables

Metrics. Let $(y_i, \hat{y}_i)$ be observed and predicted *validation losses* at the held-out stage (PTPP=279). We evaluate errors primarily in *log space* to capture multiplicative miss and stabilize heteroscedasticity. Define the log–residuals $r_i = \log \hat{y}_i - \log y_i$.

$\mathrm{Huber}_{\log}$ ($\downarrow$) is the mean Huber loss applied to r_i with threshold δ=0.02:

$$\mathrm{Huber}_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le \delta, \\ \delta\big(|r| - \frac{1}{2}\delta\big), & |r| > \delta, \end{cases} \qquad \mathrm{Huber}_{\log} = \frac{1}{n}\sum_i \mathrm{Huber}_\delta(r_i).$$

It is quadratic near zero (like MSE) but linear for outliers, making it robust.

$\mathrm{RMSE}_{\log}$ ($\downarrow$) is the root-mean-square of the log–residuals,

$$\mathrm{RMSE}_{\log} = \sqrt{\tfrac{1}{n}\sum_i r_i^2},$$

which measures typical multiplicative error (e.g., $\mathrm{RMSE}_{\log} = 0.01$ corresponds to $\approx 1\%$ relative miss under small-error linearization).

$\mathrm{MAE}_{\mathrm{rel}}$ ($\downarrow$) is the mean absolute *relative* error in the original scale,

$$\mathrm{MAE}_{\mathrm{rel}} = \frac{1}{n} \sum_i \frac{|\hat{y}_i - y_i|}{y_i},$$

i.e., the average percentage miss.

$\mathrm{MAPE}_{\mathrm{clip}}$ ($\downarrow$) is a clipped MAPE that avoids division by tiny y_i:

$$\mathrm{MAPE}_{\mathrm{clip}} = \frac{1}{n} \sum_i \frac{|\hat{y}_i - y_i|}{\max(y_i,\, y_{\mathrm{clip}})},$$

with a small $y_{\mathrm{clip}} > 0$; when all $y_i \gg y_{\mathrm{clip}}$, $\mathrm{MAPE}_{\mathrm{clip}}$ equals $\mathrm{MAE}_{\mathrm{rel}}$.

Intercept/Slope report calibration from the OLS fit

$$\log y_i \;=\; a \;+\; b \,\log \hat{y}_i \;+\; \varepsilon_i.$$

Perfect calibration gives $a \approx 0$ (no systematic bias) and $b \approx 1$ (correct sensitivity). We therefore seek small $|a|$ and b close to 1.

French — Unseen PTPP=279, no anchors.

Formulation	$\mathrm{Huber}_{\log}$ ↓	$\mathrm{RMSE}_{\log}$ ↓	$\mathrm{MAE}_{\mathrm{rel}}$ ↓	$\mathrm{MAPE}_{\mathrm{clip}}$ ↓	Interc.	Slope
Form 1 (Additive)	2.34×10^{-4}	2.27×10^{-2}	2.08×10^{-2}	2.08×10^{-2}	-0.01	0.991
Form 2 (Gated)	1.99×10^{-4}	2.12×10^{-2}	1.83×10^{-2}	1.83×10^{-2}	0.05	0.970
Form 3 (G+F)	4.43×10^{-5}	9.53×10^{-3}	6.70×10^{-3}	6.70×10^{-3}	0.01	0.991
D-CPT (transfer)	4.74×10^{-4}	3.47×10^{-2}	3.43×10^{-2}	3.43×10^{-2}	-0.00	0.961

French — Unseen PTPP=279, with 241M anchors.

Formulation	$\mathrm{Huber}_{\log}$ ↓	$\mathrm{RMSE}_{\log}$ ↓	$\mathrm{MAE}_{\mathrm{rel}}$ ↓	$\mathrm{MAPE}_{\mathrm{clip}}$ ↓	Interc.	Slope
Form 1 (Additive)	5.22×10^{-5}	1.02×10^{-2}	8.56×10^{-3}	8.56×10^{-3}	0.03	0.956
Form 2 (Gated)	4.23×10^{-5}	9.20×10^{-3}	8.17×10^{-3}	8.17×10^{-3}	0.00	0.992
Form 3 (G+F)	3.54×10^{-5}	8.42×10^{-3}	7.39×10^{-3}	7.39×10^{-3}	0.00	0.992

English/Arabic source — Unseen PTPP=279, no anchors.

Formulation	$\mathrm{Huber}_{\log}$ ↓	$\mathrm{RMSE}_{\log}$ ↓	$\mathrm{MAE}_{\mathrm{rel}}$ ↓	$\mathrm{MAPE}_{\mathrm{clip}}$ ↓	Interc.	Slope
Form 1 (Additive)	9.89×10^{-5}	1.44×10^{-2}	1.18×10^{-2}	1.18×10^{-2}	-0.05	1.034
Form 2 (Gated)	2.79×10^{-4}	2.75×10^{-2}	2.27×10^{-2}	2.27×10^{-2}	-0.03	1.045
Form 3 (G+F)	7.55×10^{-4}	5.06×10^{-2}	4.65×10^{-2}	4.65×10^{-2}	0.01	1.030
D-CPT (transfer)	5.73×10^{-4}	4.08×10^{-2}	3.91×10^{-2}	3.91×10^{-2}	0.04	0.932

English/Arabic source — Unseen PTPP=279, with 241M anchors.

Formulation	$\mathrm{Huber}_{\log}$ ↓	$\mathrm{RMSE}_{\log}$ ↓	$\mathrm{MAE}_{\mathrm{rel}}$ ↓	$\mathrm{MAPE}_{\mathrm{clip}}$ ↓	Interc.	Slope
Form 1 (Additive)	9.21×10^{-5}	1.39×10^{-2}	1.14×10^{-2}	1.14×10^{-2}	0.01	0.981
Form 2 (Gated)	5.94×10^{-5}	1.10×10^{-2}	9.01×10^{-3}	9.01×10^{-3}	0.04	0.960
Form 3 (G+F)	8.77×10^{-5}	1.36×10^{-2}	1.14×10^{-2}	1.14×10^{-2}	0.00	0.989
D-CPT (transfer)	5.73×10^{-4}	4.08×10^{-2}	3.91×10^{-2}	3.91×10^{-2}	0.04	0.932

Appendix B: Oracle baseline

For reference, a PTPP-*wise oracle* that fits D-CPT directly on PTPP=279 and evaluates on the same:

Formulation	$\mathrm{Huber}_{\log}$	$\mathrm{RMSE}_{\log}$	$\mathrm{MAE}_{\mathrm{rel}}$	$\mathrm{MAPE}_{\mathrm{clip}}$	Interc.	Slope
D-CPT (French)	2.05×10^{-6}	2.03×10^{-3}	1.63×10^{-3}	1.63×10^{-3}	0.00	1.000
D-CPT (English/Arabic)	1.67×10^{-5}	5.77×10^{-3}	4.44×10^{-3}	4.44×10^{-3}	-0.00	1.001

The Oracle uses full PTPP=279 supervision and serves only as an upper bound.

Appendix C: In-sample grid (Form 3)

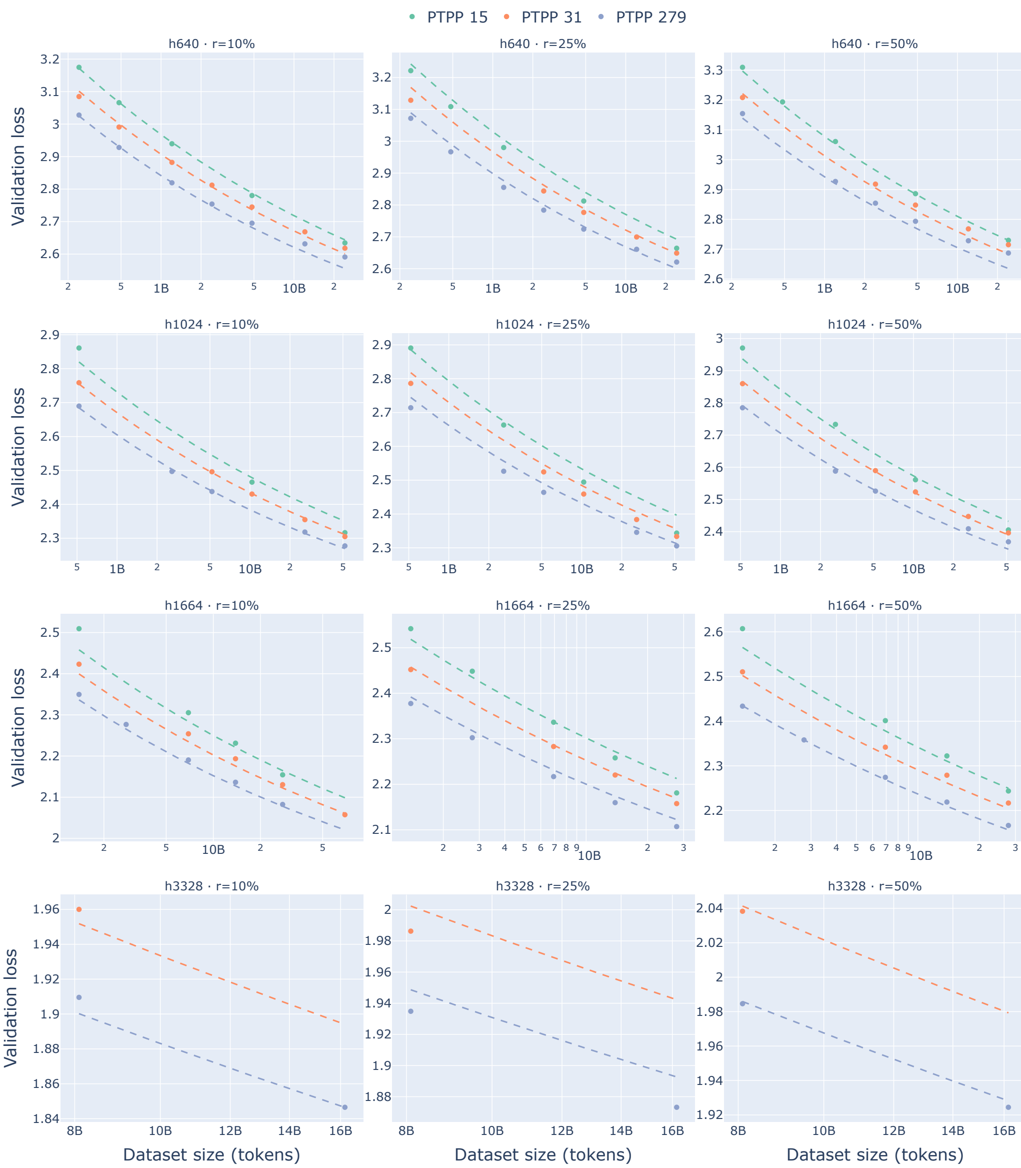

Figure 3: In-sample fits for Form 3 (gated+floor). Rows: $r \in \{0.10, 0.25, 0.50\}$; columns: {241M, 517M, 1.4B, 8.1B}. Dashed: fitted curves; markers: observations. Used only as an auxiliary fit-quality check.