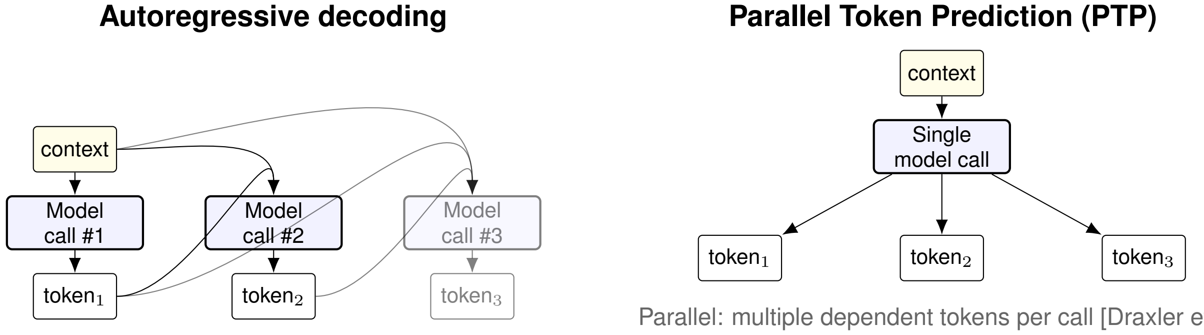
Scaling Parallel Token Prediction for Fast Large Language Model Inference

Scientific Lead: Julia Vogt (ETH Zurich)

Co-Applicants:
Felix Draxler (UC Irvine)
Stephan Mandt (UC Irvine)
Robin C. Geyer (ETH Zurich)

Abstract

Parallel Token Prediction (PTP) is a recently proposed framework for parallel sequence generation in language models that predicts multiple *dependent* tokens in a single transformer call by incorporating the sampling procedure into the model [Draxler et al., 2024]. In prior work, PTP improves decoding efficiency as measured by accepted tokens per verification step (e.g., 4.18 task-average on SpecBench for a Vicuna-7B setting; 7.0 on code under teacher verification). This proposal targets the next step: scaling PTP training and evaluation to much larger foundation models on CSCS Alps, and releasing open science artifacts (code, evaluations, and model releases where licensing permits). The core hypothesis is that scaling PTP can materially reduce end-to-end latency of long-form reasoning and agentic workflows; healthcare decision support serves as an exemplary application domain.



Sequential: context + each token feeds next call.

Parallel: multiple dependent tokens per call [Draxler et al., 2024].

Figure 1: Conceptual illustration: PTP targets the sequential decoding bottleneck by predicting multiple dependent tokens per model call [Draxler et al., 2024].

1 Objectives and Alignment with Swiss AI Initiative

Goal: Scale Parallel Token Prediction (PTP) to large foundation models, enabling significantly faster LLM inference for time-critical applications.

Alignment with Swiss AI Initiative:

- **Advance core AI capabilities:** Efficiency and scalability of LLM inference
- **Open science:** Open-source release of scaled PTP models and training code
- **Cross-institutional collaboration:** ETH Zurich + UC Irvine

- **Application domains:** Enables faster inference for agentic AI systems, with healthcare as exemplary use case

Primary selling point: PTP reduces the *sequential* bottleneck of autoregressive decoding by predicting multiple dependent tokens per model call. In [Draxler et al., 2024], this is quantified via accepted tokens per verification step (higher implies fewer sequential steps). Our goal is to scale this approach to substantially larger models and provide reproducible evidence of latency reductions on long-form generation and agentic workflows.

2 State of the Art

Autoregressive LLMs generate one token per forward pass, creating a latency bottleneck for long outputs. Existing acceleration approaches:

- **Speculative decoding:** Small draft model proposes tokens, large model verifies. Still sequential at the draft level.
- **Multi-token prediction:** Predicts multiple tokens independently. Independence assumption limits quality—produces incoherent sequences (e.g., "`def numpy`", "`import find`").
- **Discrete diffusion:** Iterative refinement, but also assumes token independence per step.

PTP [Draxler et al., 2024]**:** Predicts multiple *dependent* tokens in a single forward pass by feeding auxiliary random variables into the model. The paper proves expressivity matching autoregressive models (no independence assumption) and reports strong empirical decoding efficiency (e.g., 4.18 accepted tokens/step task-average on SpecBench in a Vicuna-7B setting; and 7.0 accepted tokens/step on code in a distilled setup).

Gap: PTP has only been demonstrated at 1–7B scale. Scaling to frontier models (70B+) is unexplored.

3 Activities and Deliverables

3.1 Work Packages

WP1: Scaling PTP Training (Months 1–6)

- Scale PTP distillation and/or fine-tuning to larger foundation models on CSCS Alps
- Validate training stability, throughput, and scaling behavior (feasibility focus)
- Compute plan: use ≈2/3 of the allocated compute in the first 6 months (Swiss AI guideline)

WP2: Inference Optimization (Months 4–9)

- Efficient PTP inference implementation with verification-based decoding
- Systems work (serving and evaluation harness) with Swiss AI engineering support

WP3: Evaluation and Benchmarking (Months 6–12)

- Evaluation on long-form generation and reasoning-oriented benchmarks (latency-focused)
- Evaluation in agentic workflows (multi-step tool use) as a latency stress test
- Domain-specific pilot evaluation in healthcare-related settings (as application exemplar; method-first)

3.2 Deliverables

- Open-source training and inference code for scaled PTP (including reproducible evaluation)
- Model releases where third-party licensing permits (preference for open-weight bases; otherwise adapters/checkpoints)
- Benchmark results and a short technical report summarizing scaling and latency findings

3.3 Distribution and Licensing

Code and evaluation artifacts released under a permissive open-source license (Apache 2.0 or similar). Model releases will follow the most permissive option compatible with third-party licensing (preference for open-weight foundations).

3.4 Compute Estimate

Assumptions:

- Target model size: 70B parameters
- Training tokens: T tokens (distillation or fine-tuning corpus)
- Measured throughput on Alps: X tokens/GPU-hour (to be established in feasibility runs)

Rough calculation (plug-in):

$$ext GPU-hours = \frac{T \times \text{epochs} \times \text{overhead factor}}{X}$$

Preliminary allocation request:

- WP1 (Training/finetuning): [TBD] GPU-hours
- WP2 (Systems/inference): [TBD] GPU-hours
- WP3 (Evaluation): [TBD] GPU-hours
- **Total:** [TBD] GPU-hours (expectation: can exceed 500K GPU-hours)

Note: Precise estimates require feasibility experiments on Alps. We request a small preparatory allocation to validate throughput assumptions.

3.5 Budget Summary

Personnel (ETH Zurich, requiring 1:1 matching):

- 1 Postdoc, 12 months: ∼CHF 130k total (CHF 65k requested + CHF 65k matching from ETH)

Research expenses: ∼CHF 40k (workshops, travel, outreach)
Engineering support requested: 12 FTE-months (from Swiss AI core team, no matching required)
Total requested funding: ∼CHF 105k + compute allocation + engineering support
Note: Exact salary figures to be confirmed by ETH administration.

4 Team and Collaboration Model

Name	Affiliation	Role
Julia Vogt	ETH Zurich	Scientific Lead
Robin C. Geyer	ETH Zurich	Co-Applicant, Project Coordination
Stephan Mandt	UC Irvine	Co-Applicant, PTP Method Expert
Felix Draxler	UC Irvine	Co-Applicant, PTP Method Expert

Expertise:

- **Mandt & Draxler:** Original PTP authors; deep expertise in normalizing flows, variational inference, and probabilistic ML. Stephan Mandt leads the AI in Science Institute at UC Irvine.
- **Vogt:** Associate Professor at ETH Zurich, leading the Medical Data Science Group. Expert in interpretable ML, generative models (VAEs, diffusion), and clinical AI applications. Extensive publication record at NeurIPS, ICML, ICLR on concept bottleneck models, multimodal learning, and healthcare ML.
- **Geyer:** Postdoctoral researcher at ETH Zurich (Vogt group) and Charité Berlin. PhD on representation learning for critical care. Expertise in domain adaptation, clinical time series, and privacy-preserving ML. Prior collaboration with Mandt.

Collaboration model:

- ETH team: Large-scale training on CSCS, project coordination, downstream evaluation
- UCI team: Methodological expertise, algorithm development (no personnel funding requested, external collaborators)

Requested Personnel (ETH Zurich):

- 1 Postdoc (12 months)

Requested Engineering Support (Swiss AI):

- 12 FTE-months engineering support for scaling infrastructure, training pipelines, and inference optimization

5 Novelty and Impact

Novelty:

- First scaling of PTP to frontier model sizes
- Novel training strategies for large-scale PTP

Impact:

- **Inference cost reduction:** 4–7× fewer forward passes for same output
- **Time-critical applications:** Enables real-time LLM inference where latency matters (clinical decision support, autonomous agents)
- **Swiss/European ecosystem:** Open models reduce dependence on proprietary US/China systems

Potential upside (speculative): At scale, PTP models trained from scratch (not distilled) may develop novel reasoning capabilities by "thinking in longer sequences." Early evidence from multi-token prediction literature suggests potential quality improvements, not just speed. We consider this a secondary objective pending empirical validation at scale.

6 Importance for Stakeholders

Swiss stakeholders:

- **Industry/SMEs:** Reduced inference costs for LLM deployment; faster response times for customer-facing applications
- **Healthcare sector:** Faster clinical decision support systems; enables real-time agentic workflows in time-critical settings
- **Academia:** Open foundation models and code for further research

European/Global:

- Contribution to open AI ecosystem, reducing dependence on proprietary systems
- State-of-the-art in efficient LLM inference, applicable across domains

7 Data Availability and Ethical Considerations

Data:

- Training/evaluation: public datasets and standard benchmarks used in the PTP line of work (no new data collection in this proposal)
- No patient-level personal data is required for the core method development

Ethics:

- Faster inference may amplify both beneficial and harmful uses
- Mitigation: Focus on open release enabling community oversight
- Clinical angle: evaluation/benchmarking only; no patient interaction in this project

Legal and regulatory compliance (as required by the call):

- Ensure compliance with applicable data protection laws and regulations (e.g., Swiss FADP; GDPR where applicable)
- Document provenance and licensing of any datasets and model components used; prefer permissively licensed resources
- Where model weights are released, provide accompanying documentation on training data sources and licensing constraints to support due diligence

References

Felix Draxler, Justus Will, Farrin Marouf Sofian, Theofanis Karaletsos, Sameer Singh, and Stephan Mandt. Parallel token prediction for language models. *arXiv preprint arXiv:2512.21323*, 2024.