Latent & Implicit Thinking – Going Beyond CoT Reasoning

ICLR 2026 Workshop Proposal

Introduction

AI models have demonstrated remarkable reasoning capabilities [16, 42] by explicitly generating intermediate steps in natural language, an approach known as Chain-of-Thought (CoT) reasoning [28, 45]. While CoT reasoning has become the dominant paradigm for reasoning in AI systems, it incurs substantial computational costs due to the increasing length and complexity of the generated reasoning chains [13, 29]. Moreover, although autoregressive natural language reasoning aligns intuitively with human cognition, it may not represent the most efficient or effective internal computation medium for neural networks [6]. Thus, emerging bodies of research are exploring alternative forms of intermediate reasoning, including using continuous hidden representations [6, 5, 14, 48, 10, 34], learning discrete reasoning tokens optimized for efficiency rather than readability [40], and other non-autoregressive thinking approaches [51, 53]. These alternative paradigms present promising opportunities to achieve more efficient and potentially more powerful implicit reasoning capabilities within future AI models, thus motivating this workshop proposal.

One prominent direction involves optimizing implicit reasoning directly through novel training methods, exemplified by implicit CoT approaches [6, 5], which first demonstrated the feasibility of internalizing reasoning steps within continuous hidden states. Subsequent advances, including Coconut [14], Dualformer [39], CODI [37], and the recent mixture of tokens line of work [58, 62, 1] have substantially improved implicit reasoning. Complementing these training strategies, new model architectures have been developed to facilitate latent reasoning. Models using recurrent connections or looping hidden states [48, 10, 34, 43] allow reasoning at arbitrary depths without explicit token generation. Another innovative research direction investigates alternative probabilistic formulations, shifting from traditional autoregressive modeling toward non-autoregressive generative paradigms such as diffusion models [12, 52, 27, 51, 53], search-tree-based methods [21], and parallel CoTs [46, 59]. Additionally, efforts in compressing explicit reasoning traces into compact, reasoning-specific discrete tokens have shown significant promise in improving computational efficiency [40]. Collectively, these empirical developments are further bolstered by theoretical insights, indicating that transformers inherently possess implicit reasoning capabilities closely tied to model depth [49, 22, 34], and latent continuous tokens are capable of encoding superpositions of reasoning states [61, 60]. Other related ideas include modeling thinking as a latent variable process [47], and unifying “system 1” and “system 2” thinking processes [54, 39, 20]. There are also reinforcement learning algorithms to extract and verbalize latent thinking of machines without explicit rewards [33, 7, 15]. Latent reasoning chains in other modalities beyond text has also shown its effectiveness recently [50, 31].

The proposed workshop, **Latent & Implicit Thinking – Going Beyond CoT Reasoning (LIT)**, seeks to unify these diverse yet complementary research directions. By fostering interdisciplinary dialogue among researchers pursuing implicit reasoning strategies, novel model architectures, and non-autoregressive generative frameworks, we aim to deepen our collective understanding of how to harness and extend LLMs’ inherent implicit reasoning capacities. Participants will gain comprehensive insights from invited speakers and organizers who are leading contributors in these respective areas. For researchers outside this area, the workshop will serve as a structured introduction to the rapidly evolving landscape of latent and implicit reasoning with neural networks. We invite original submissions on, but not limited to, the following themes:

- **Speical Thinking Tokens**: Explicit CoT compressed to special tokens (e.g., continuous thought tokens [14], VQ-VAE codes [40], gist token [25, 57]). CoT augmentation via filler [11, 30, 26] or planning tokens [44].
- **Looped Architectures**: Recurrence mechanisms (loop unrolling, dynamic halting) [4, 9, 10]. Training curricula and stability for deep iterative models [35, 34, 55, 3].
- **Stateful Reasoning**: Leveraging key/value caches for multi-step inference [24]. Comparisons between pure activation vs. cache-augmented loops.
- **Parallel Reasoning**: Diffusion models for bidirectional, iterative denoising-based reasoning [52]. Fractal generative frameworks [17] and next-block prediction [32]. Parallel CoTs generation [46, 59].

- **Training Strategies**: Training strategies to enable neural networks to reason in latent space or extract their latent thoughts. e.g. curriculum learning [5, 14], distillation [6, 37], reinforcement learning in the latent space [56] or for extracting latent CoTs [33, 7, 15], and pretraining from scratch [41].
- **Theoretical Analysis**: Theory or analysis on advantages of reasoning in the latent space. e.g. reasoning by superposition [61, 60], theoretical bounds on reasoning depth vs. layer count [34, 22], and layer-wise [38] or head-wise [8] specialization and attribution of reasoning functions [2].
- **Evaluation and Benchmarks**: Metrics for latent vs. explicit CoT capabilities. Datasets and tasks that stress ultra-deep or multi-hop latent reasoning [49, 23].
- **Limitations and Safety**: Understanding pros and cons of latent and explicit CoT, and interpretability & faithfulness of reasoning from a safety and alignment perspective.

The expected outcome of the workshop is a cohesive, cross-disciplinary understanding of implicit reasoning in neural networks. Post-workshop, we will release a summary report synthesizing discussions and open questions, encourage joint research initiatives, and maintain a public repository of slides, videos, and accepted works to maximize accessibility and long-term impact.

Previous Related ICLR Workshops

This workshop does not have previous editions. The closest related workshop at ICLR is the *Workshop on Reasoning and Planning for Large Language Models* at ICLR 2025, which addressed challenges and techniques for reasoning in LLMs. The workshop was focused on LLMs only and there was only one paper in the workshop discussing latent reasoning method [14], while the majority of the papers predominantly relied on explicit CoTs. There are also two previous workshops at NeurIPS focused on LLM reasoning: System-2 Reasoning at Scale (NeurIPS 2024) and Foundations of Reasoning in Language Models (NeurIPS 2025).

Unlike prior workshops that focused on existing reasoning paradgim of explicit CoTs in LLMs, our workshop shifts attention toward a new paradigms for reasoning that occur implicitly, within latent representations or parameter space for all deep learning models across different modalities. This new reasoning paradigm offers better reasoning efficiency and higher expressiveness potential. This timely focus aligns with emerging industrial and academic interests (e.g., reasoning-efficient architectures at Meta, Apple, NVIDIA, and Google).

We expect the workshop to break new ground rather than reiterate prior debates by unifying diverse communities—those studying implicit CoT generation, latent-state architectures, and non-autoregressive modeling. The resulting discussions will illuminate a unified conceptual framework for implicit reasoning, contributing to both scientific understanding and practical efficiency in LLMs.

Submissions

The workshop will accept papers from preliminary research results and visionary papers to full-length papers. We will host a Tiny and Short Paper Track in alignment with ICLR's initiative, explicitly welcoming late-breaking results, replication studies, and conceptual explorations that may not yet have full-paper maturity. Review criteria will emphasize clarity, originality, and potential impact rather than extensive experiments. Submissions should follow the ICLR proceedings format and choose the suitable categories as follows:

- **Tiny and Short Paper**: 2 to 4 pages + references and appendix.
- **Regular Paper**: up to 9 pages + references and appendix.

For novelty, the workshop does not accept submissions that have previously been published at ICLR or other machine learning or related venues. For openness, we encourage submissions with sufficient open-source resources (e.g., checkpoint, code, data, training details). We do not allow AI as primary authors for workshop submissions.

All accepted papers will be presented as posters. We will select around 3 papers for short oral presentations and 2 papers for outstanding paper awards with potential cash incentives. To support accessibility, we will publish accepted works on the workshop website and highlight selected tiny papers during the poster session, ensuring visibility for early-stage contributors. While the workshop is non-archival, the workshop webpage will be maintained to hold all related metarials of the workshop for future online viewers.

Timeline

To comply with ICLR's March 1, 2026 notification deadline, we propose the following timeline (all deadlines are AOE):

- **Workshop website, schedule, and call for papers release**: January 5, 2025
- **Submission deadline**: Feburary 5, 2026
- **Reviewing period**: Feburary 6 - Feburary 26, 2026
- **Notification of acceptance**: March 1, 2026
- **Camera-ready deadline**: March 15, 2026
- **Workshop date**: align with the ICLR 2026 conference date

This schedule allows sufficient time for reviewing, communication, and logistical planning.

Tentative Schedule

The workshop will include invited talks, contributed talks, and posters, as shown below.

To encourage meaningful dialogue, each keynote and contributed talk will be followed by a moderated Q&A period (5-10 min), with discussion questions collected both in-person and online. Dedicated poster sessions and a final panel discussion will provide structured venues for audience engagement. Additionally, we will organize an open "lightning round" after lunch, where participants can share 1-minute research ideas, encouraging spontaneous and inclusive participation.

While the workshop is in-person, all talks and poster abstracts will be made publicly available via our website and YouTube channel within two weeks after the event. We will also host a moderated discussion board for remote participants to post comments and questions. This ensures that researchers facing visa or financial barriers can still meaningfully engage.

Time	Arrangement
08:00–09:00	**Registration**
09:00–09:10	**Opening Remarks**: Welcome address and workshop overview
09:10–09:40	**Keynote Talk 1**: Yuandong Tian, Meta (confirmed)
09:40–10:10	**Keynote Talk 2**: Lisa Li, Stanford & UW (confirmed)
10:10–10:40	**Oral Presentations** from submissions
10:40–11:10	*Coffee Break*
11:10–11:40	**Keynote Talk 3**: Samy Bengio, Apple (confirmed)
11:40–12:30	**Poster Session 1**
12:30–13:30	*Lunch Break*
13:30–14:00	**Keynote Talk 4**: Beidi Chen, CMU (confirmed)
14:00–14:30	**Keynote Talk 5**: Tom Goldstein, UMD (confirmed)
14:30–15:00	**Keynote Talk 6**: Lingpeng Kong, HKU (confirmed)
15:00–16:00	**Panel Discussion**: What are the key challenges in enabling large language models to reason effectively in latent space?
16:00–16:30	*Coffee Break*
16:30–17:20	**Poster Session 2**
17:20–17:30	**Award Announcement** and **Closing Remarks**

Invited Speakers

The organizers will monitor the time of speakers and ensure that there is sufficient time for discussion. All invited speakers have confirmed to give their talks in person. For exceptional circumstances like visa issues, we will allow the speaker to present remotely via Zoom.

Lisa Xiang Li (confirmed) is a PhD candidate at Stanford University, advised by Percy Liang and Tatsunori Hashimoto, and an incoming Assistant Professor at the University of Washington. Her research focuses on developing methods to make language models more capable and controllable. Lisa is supported by the Two Sigma PhD Fellowship and the Stanford Graduate Fellowship, and is the recipient of an EMNLP Best Paper Award. *She is a pioneer in reasoning with soft tokens [19, 26] and diffusion language models [18].*

Beidi Chen (confirmed) is an Assistant Professor in the Department of Electrical and Computer Engineering at Carnegie Mellon University. She is a Visiting Research Scientist at FAIR, Meta. Before that, she was a postdoctoral scholar at Stanford University. She received her Ph.D. from Rice University in 2020 and B.S. from UC Berkeley in 2015. Her research focuses on efficient machine learning. Specifically, she designs and optimizes algorithms and models on modern hardware to accelerate large machine learning systems. Her work has won a best paper runner-up at ICML 2022, a best paper award at IISA 2018, and a best paper award at USENIX LISA 2014. She was selected as a Rising Star in EECS by MIT in 2019 and UIUC in 2021.

Samy Bengio (confirmed) is a Senior Director of AI and Machine Learning Research at Apple. Previously, he was a research scientist at Google, contributing extensively to foundational advances in deep learning, neural networks, representation learning, and machine learning systems. Samy has co-authored over 400 influential publications and is widely recognized as a leading figure in the machine learning community. He has served in numerous senior organizing roles, including general and program chair positions at premier conferences such as NeurIPS and ICLR.

Tom Goldstein (confirmed) is a Professor of Computer Science and Director of the Maryland Center for Machine Learning, with appointments in Applied Math and Electrical and Computer Engineering. His research focuses on responsibly building AI systems and draws on ideas from signal processing and applied mathematics. His work at the boundary between theory and systems, leveraging mathematical foundations and efficient hardware to improve the training, fine-tuning, benchmarking, and robustness of large models. He has been the recipient of several awards, including SIAM's DiPrima Prize, a DARPA Young Faculty Award, and a Sloan Fellowship. *Tom has done influential research on recurrent networks [36] and inference-time scaling for latent thinking [10] with them.*

Yuandong Tian (confirmed) is a Research Scientist Director in Meta GenAI, leading a group for Llama reasoning. His research direction covers multiple aspects of decision making, including reinforcement learning, planning and efficiency, as well as theoretical understanding of LLMs. Prior to that, he worked in Google Self-driving Car team in 2013-2014 and received a Ph.D in Robotics Institute, Carnegie Mellon University in 2013. He has been appointed as area chairs for NeurIPS, ICML, AAAI, CVPR, and AIStats. *Yuandong has made significant contributions to latent reasoning, particularly through soft-token-based methods [14, 40].*

Lingpeng Kong (confirmed) is an assistant professor at the University of Hong Kong (HKU), and a co-director of the HKU NLP Lab. His work lies at the intersection of natural language processing (NLP) and machine learning (ML), with a focus on representation learning, structured prediction, and generative models. Before joining HKU, he was a research scientist at DeepMind (London). Lingpeng obtained his Ph.D. from Carnegie Mellon University. *Lingpeng has been working on diffusion language models, and has been contributing important work [51, 53] to diffusion-based parallel latent reasoning.*

Diversity and Inclusion

Speakers: Our confirmed speaker lineup already reflects strong diversity in both identity and affiliation, representing leading institutions across academia and industry (Meta, HKU, Apple, Stanford, NVIDIA, CMU, and UMD). Two out of six of our invited speakers identify as female, and the group spans North America and Asia, ensuring a wide range of viewpoints and experiences.

Organizing Committee:The organizing committee similarly balances seniority (faculty, postdoctoral, and student organizers), institutional diversity (academia and industry from the United States, Canada, and Asia), and disciplinary backgrounds (reasoning, interpretability, machine learning theory, and efficient systems). Several organizers are first-time workshop organizers, supported and mentored by senior organizers with substantial experience leading high-impact events such as ICLR SCI-FM, NeurIPS M3L, and OPT@NeurIPS. This structure embodies our commitment to inclusive leadership development within the ML community.

Attendees: We strive to create a welcoming environment for attendees from diverse backgrounds, facilitating discussions among various perspectives after each talk and during the coffee break. We plan to offer a limited number of travel grants to support attendees from underrepresented groups, including but not limited to individuals with disabilities, those from developing countries, and early-career researchers facing financial constraints. We also plan to maintain hybrid accessibility by recording and streaming talks and hosting both syncronized and asynchronous discussions on the workshop website and OpenReview forum for those unable to attend in person.

Estimated Interest & Significance

There is a rapidly growing interest across academia and industry in exploring reasoning with deep neural networks. This trend is underscored by the success of related recent workshops, such as the ICLR 2025 Workshop on Reasoning and Planning for Large Language Models, which attracted over 130 accepted papers.

Based on these precedents and our collective organizing experience, we anticipate substantial community participation, estimating approximately 80-100 submissions and 100–150 attendees for this proposed workshop.

Audience Outreach and Advertising Plan To attract a broad and engaged audience, we will leverage multiple complementary outreach channels. First, we will announce the workshop through official ICLR mailing lists, Twitter/X, and academic platforms such as Open NLP Slack channel, Machine Learning News Google group and Hugging Face community forums. We will directly reach out to active research groups working on reasoning, efficiency, and foundation models across academia and industry (e.g., FAIR, Google DeepMind, NVIDIA, Apple, Stanford, CMU, TTIC, HKU, Princeton). The organizing committee will promote the call for papers via personal and lab websites, LinkedIn, and institutional newsletters. We will also coordinate with related workshops to cross-promote submissions and attendance. To engage early-career researchers, we plan to share our workshop on student and postdoc mailing lists, including those for the Rising Stars in EECS and Women in ML communities. Finally, we will maintain an active workshop website and social media presence with timely updates, teaser posts about speakers, and spotlight threads for accepted papers to sustain interest before and during the event.

LLM Usage Policy

In alignment with the ICLR 2026 Policies on Large Language Model Usage, our workshop will clearly distinguish between AI-assisted and AI-generated content. We will explicitly prohibit AI-generated submissions in the tiny and short paper tracks, while allowing limited, transparent AI assistance (e.g., for grammar correction or rephrasing) as long as the intellectual contribution and analysis remain primarily human-authored. During the review process, LLMs will not be allowed to use for automated reviewing or decision-making; all evaluations will be conducted by human reviewers. The organizers may use LLMs only for logistical purposes, such as drafting website descriptions or formatting templates, with full human oversight. Any substantial use of AI tools by authors, reviewers, or organizers must be disclosed in submission metadata or acknowledgments. This policy ensures responsible, transparent, and human-centered engagement with LLM technology throughout the workshop.

Organization Committee

Collectively, the organizing team brings deep experience in hosting major ML workshops including Open Science for Foundation Models Workshop at ICLR, OPT: Optimization for Machine Learning and Mathematics in Modern Machine Learning (M3L) at NeurIPS. Many team members also have experience with conference service (area chairs for NeurIPS, ICML, ICLR), and program-committee participation across top venues. Junior members can gain mentorship from senior organizers with prior organizing experience. All the organizers will be able to attend the workshop for organizing the workshop and hosting panel discussions in person, except for unexpected circumstances.

Xinyi Wang (xw2259@princeton.edu) is a postdoctoral researcher at Princeton University, and an incoming assistant professor at the University at Buffalo, SUNY. She recently defended her Ph.D. at the University of California, Santa Barbara (UCSB). She has received a J.P. Morgan AI Ph.D. Fellowship. Her research centers on developing a principled understanding of large foundation models, with the aim of enhancing their capabilities and addressing their limitations.

Nikunj Saunshi (nsaunshi@google.com) is a Senior Research Scientist at Google. His research interests lie in interweaving theory and empirics to design efficient and reliable learning algorithms and to demystify deep learning and AI. His research has spanned topics like reasoning in large language models, self-supervised representation learning, meta-learning, NLP, interpretability of deep learning models. He received a PhD in Computer Science from Princeton University and was a receipient of the 2022 Siebel Scholars award.

Rui-Jie Zhu (ridger@ucsc.edu) is a Ph.D. student at UC Santa Cruz, advised by Professor Jason K. Eshraghian, specializing in efficient deep learning with a particular focus on Linear Attention mechanisms. He is a key contributor to the development of RWKV. His work also includes notable projects like SpikeGPT. He is a recipient of the UC Santa Cruz Chancellor's 2024 Innovation Impact Award. Rui-Jie's research is dedicated to addressing the computational bottlenecks of traditional attention mechanisms and advancing scalable, efficient AI systems.

Liu Yang (liu.yang@wisc.edu) is a PhD student at the University of Wisconsin–Madison, advised by Prof. Robert Nowak, Prof. Dimitris Papailiopoulos, and Prof. Kangwook Lee. Her research focuses on improving the efficiency of large language models by understanding and leveraging their internal representations, such as latent embeddings and task vectors. She has co-authored multiple publications in top-tier machine learning conferences, including NeurIPS, ICLR, and ICML.

Yuntian Deng (yuntian@uwaterloo.ca) is an assistant professor at the University of Waterloo and a visiting professor at NVIDIA, working with Prof. Yejin Choi. Previously, he was a postdoc at AI2, also advised by Prof. Choi, and earned his PhD from Harvard University under Profs. Alexander Rush and Stuart Shieber. His research focuses on natural language processing and machine learning, particularly implicit chain-of-thought (CoT) reasoning. In 2023, he introduced the idea of implicit CoT reasoning using hidden states in *Implicit Chain-of-Thought Reasoning via Knowledge Distillation*, and further advanced this idea in *From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step*. His other projects include WildChat and OpenNMT.

Nishanth Dikkala (nishanthd@google.com) is a Senior Research Scientist at Google. He is interested in efficient architecture design, augmenting LLM reasoning capabilities, and learning theory. He holds a Ph.D. from MIT and a Bachelors from Indian Institute of Technology Bombay.

Jiaheng Liu (liujiaheng@nju.edu.cn) is an assistant professor at Nanjing University. His current research mainly focuses on Large Language Models (LLMs), with contributions in pre-training, alignment, and open science of LLMs. He has published 60+ top conference/journal papers, including the ACL 2024 Outstanding Paper Award. Additionally, he has served as the invited speaker of the first GLOW workshop at IJCAI 2023, and the *organizer of the first SCI-FM workshop at ICLR 2025*.

Zhiyuan Li (zhiyuanli@ttic.edu) is an assistant professor in the Toyota Technological Institute at Chicago (TTIC) and an affiliated faculty at Uchicago CS. He is currently a visiting faculty at Google Research. He obtained his Ph.D. in computer science at Princeton University in 2022 and spent one year for postdoc at Stanford CS. His research focuses on machine learning theory. He is a recipient of Microsoft Research PhD Fellowship. *He is the leading organizer of the first and second workshop for Mathematics in Modern Machine Learning (M3L) at NeurIPS 2023 and 2024.*

Program Committee

We aim to have every submission reviewed by at least three Program Committee members. We will use OpenReview's automated conflict-of-interest detection system based on email domains, co-authorship, and institutional affiliation. Additionally, organizers will manually verify conflicts to ensure fairness: no organizer will review or handle submissions from their own institution or recent collaborators. Submissions with potential conflicts will be reassigned to other reviewers or guest meta-reviewers. The Program Committee is as follows:

Confirmed: Vasilis Papageorgiou (UW-Madison), Ziyang Cai (UW-Madison), Sachin Goyal (CMU), Lakshya A Agrawal (UC Berkeley), Violet Xiang (Stanford), Tengxiao Liu (UCSB), Nathan Roll (Stanford), Wentao Zhang (University of Waterloo), Xin Yan (University of Waterloo), Ziwei Tang (University of Waterloo), Akira Kudo (UBC), Liliana Hotsko (University of Waterloo), Woojeong Kim (Cornell), Zhen Wu (CMU), Xiangjue Dong (TA&MU), Qi Zhang (Temple University), Guanchao Feng (Stony Brook University), Rushil Gupta (Mila), Janvijay Singh (UIUC), Shen Nie (Renmin University), Jihoon Tack (KAIST), Yunting Yin (Eastern Michigan University), Angeliki Giannou (UW-Madison), Yufan Zhuang (UCSD), Sean McLeish (UMD), Abhishek Panigrahi (Princeton), Xingyu Qu (MBZUAI), Peigeng Huang (Nanjing University), Yuan-Hong Liao (University of Toronto), Jishen Yang (Amazon), Gyuwan Kim (UCSB), Zheyang Xiong (UW-Madison)

Tentative: Chenyang Zhao (UCLA), Wenda Li (University of Michigan), Yuanxing Zhang (Kwai Tech), Zelei Chen (Northwestern University), Yue Zhang (ByteDance), Hanqing Wang (ShanghaiTech University), Chujie Zhang (Tsinghua University), Haoran Wang (Tsinghua University), Zekun Wang (Kwai Tech), Chenchen Zhang (Tencent), Jian Yang (Alibaba, Qwen), Wangchunshu Zhou (OPPO), Zili Wang (Inf Tech), Xinrun Du (01.AI), Tianyu Zheng (01.AI), Jie Liu (CUHK), Qian Liu (Sea AI Lab), Yizhi Li (University of Manchester), Ge Zhang (ByteDance), Tara Kaul (University of Manchester), Yiming Liang (UCAS), Wei Fan (University of Oxford), Yanjun Shao (Yale), Yanan Ma (University of Manchester), Xingwei Qu (University of Manchester), Hangyu Guo (Alibaba), Siwei Wu (University of Manchester), Shuyue Guo (Alibaba), Haoran Que (Shanghai AI Lab), Michael Saxon (UCSB), Alon Albalak (SynthLabs), Antonis Antoniades (UCSB), Alfonso Amayuelas (UCSB), He Zhu (Alibaba), Meng Cao (MBZUAI), Yinghui Li (Tsinghua University), Tom Palczewski (SAP), Wen-Ding Li (Cornell), Xinyuan Lu (NUS), Tan Yu (NVIDIA), Hanhua Hong (University of Manchester)

References

[1] N. Butt, A. Kwiatkowski, I. Labiad, J. Kempe, and Y. Ollivier. Soft tokens, hard truths. *arXiv preprint arXiv:2509.19170*, 2025.

[2] Y. Cai, D. Cao, X. Xu, Z. Yao, Y. Huang, Z. Tan, B. Zhang, G. Liu, and J. Fang. On predictability of reinforcement learning dynamics for large language models. *arXiv preprint arXiv:2510.00553*, 2025.

[3] A. B. De Luca and K. Fountoulakis. Simulation of graph algorithms with looped transformers. *arXiv preprint arXiv:2402.01107*, 2024.

[4] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[5] Y. Deng, Y. Choi, and S. Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

[6] Y. Deng, K. Prasad, R. Fernandez, P. Smolensky, V. Chaudhary, and S. Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

[7] Q. Dong, L. Dong, Y. Tang, T. Ye, Y. Sun, Z. Sui, and F. Wei. Reinforcement pre-training. *arXiv preprint arXiv:2506.08007*, 2025.

[8] W. Du, L. Jiang, K. Tao, X. Liu, and H. Wang. Which heads matter for reasoning? rl-guided kv cache compression. *arXiv preprint arxiv:2510.08525*, 2025.

[9] Y. Gao, C. Zheng, E. Xie, H. Shi, T. Hu, Y. Li, M. Ng, Z. Li, and Z. Liu. Algoformer: An efficient transformer framework with algorithmic structures. *Transactions on Machine Learning Research*, 2024.

[10] J. Geiping, S. McLeish, N. Jain, J. Kirchenbauer, S. Singh, B. R. Bartoldson, B. Kailkhura, A. Bhatele, and T. Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

[11] S. Goyal, Z. Ji, A. S. Rawat, A. K. Menon, S. Kumar, and V. Nagarajan. Think before you speak: Training language models with pause tokens. In *ICLR*, 2024.

[12] I. Gulrajani and T. B. Hashimoto. Likelihood-based diffusion language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16693–16715. Curran Associates, Inc., 2023.

[13] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[14] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[15] A. Hatamizadeh, S. N. Akter, S. Prabhumoye, J. Kautz, M. Patwary, M. Shoeybi, B. Catanzaro, and Y. Choi. Rlp: Reinforcement as a pretraining objective. *arXiv preprint arXiv:2510.01265*, 2025.

[16] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] T. Li, Q. Sun, L. Fan, and K. He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025.

[18] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022.

[19] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[20] Y.-H. Liao, S. Elflein, L. He, L. Leal-Taixé, Y. Choi, S. Fidler, and D. Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025.

[21] X. Lu, S. Han, D. Acuna, H. Kim, J. Jung, S. Prabhumoye, N. Muennighoff, M. Patwary, M. Shoeybi, B. Catanzaro, and Y. Choi. Retro-search: Exploring untaken paths for deeper and efficient reasoning, 2025.

[22] W. Merrill and A. Sabharwal. A little depth goes a long way: The expressive power of log-depth transformers. *arXiv preprint arXiv:2503.03961*, 2025.

[23] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] A. Mohtashami, M. Pagliardini, and M. Jaggi. Cotformer: A chain-of-thought driven architecture with budget-adaptive computation cost at inference. *arXiv preprint arXiv:2310.10845*, 2023.

[25] J. Mu, X. Li, and N. Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352, 2023.

[26] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[27] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

[28] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena. Show your work: Scratchpads for intermediate computation with language models, 2021.

[29] OpenAI. Learning to reason with llms, September 2024.

[30] J. Pfau, W. Merrill, and S. R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Conference on Language Modeling (CoLM)*, 2024.

[31] T.-H. Pham and C. Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

[32] S. Ren, S. Ma, X. Sun, and F. Wei. Next block prediction: Video generation via semi-auto-regressive modeling. *arXiv preprint arXiv:2502.07737*, 2025.

[33] Y. Ruan, N. Band, C. J. Maddison, and T. Hashimoto. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

[34] N. Saunshi, N. Dikkala, Z. Li, S. Kumar, and S. J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. *arXiv preprint arXiv:2502.17416*, 2025.

[35] N. Saunshi, S. Karp, S. Krishnan, S. Miryoosefi, S. Jakkam Reddi, and S. Kumar. On the inductive bias of stacking towards improving reasoning. *NeurIPS*, 37:71437–71464, 2024.

[36] A. Schwarzschild, E. Borgnia, A. Gupta, F. Huang, U. Vishkin, M. Goldblum, and T. Goldstein. Can you learn an algorithm? generalizing from easy to hard problems with recurrent networks. In *Advances in Neural Information Processing Systems*, 2021.

[37] Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, and Y. He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

[38] O. Skean, M. R. Arefin, Y. LeCun, and R. Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024.

[39] D. Su, S. Sukhbaatar, M. Rabbat, Y. Tian, and Q. Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. *arXiv preprint arXiv:2410.09918*, 2024.

[40] D. Su, H. Zhu, Y. Xu, J. Jiao, Y. Tian, and Q. Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 2025.

[41] J. Tack, J. Lanchantin, J. Yu, A. Cohen, I. Kulikov, J. Lan, S. Hao, Y. Tian, J. Weston, and X. Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025.

[42] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[43] G. Wang, J. Li, Y. Sun, X. Chen, C. Liu, Y. Wu, M. Lu, S. Song, and Y. A. Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

[44] X. Wang, L. Caccia, O. Ostapenko, X. Yuan, W. Y. Wang, and A. Sordoni. Guiding language model reasoning with planning tokens. In *First Conference on Language Modeling*, 2024.

[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

[46] H. Wen, Y. Su, F. Zhang, Y. Liu, Y. Liu, Y.-Q. Zhang, and Y. Li. Parathinker: Native parallel thinking as a new paradigm to scale llm test-time compute. *arXiv preprint arXiv:2509.04475*, 2025.

[47] V. Xiang, C. Snell, K. Gandhi, A. Albalak, A. Singh, C. Blagden, D. Phung, R. Rafailov, N. Lile, D. Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*, 2025.

[48] L. Yang, K. Lee, R. D. Nowak, and D. Papailiopoulos. Looped transformers are better at learning learning algorithms. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] S. Yang, E. Gribovskaya, N. Kassner, M. Geva, and S. Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[50] Z. Yang, X. Yu, D. Chen, M. Shen, and C. Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[51] J. Ye, J. Gao, S. Gong, L. Zheng, X. Jiang, Z. Li, and L. Kong. Beyond autoregression: Discrete diffusion for complex reasoning and planning. In *ICLR*, 2025.

[52] J. Ye, S. Gong, L. Chen, L. Zheng, J. Gao, H. Shi, C. Wu, X. Jiang, Z. Li, W. Bi, et al. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[53] J. Ye, Z. Wu, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong. Implicit search via discrete diffusion: A study on chess. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] P. Yu, J. Xu, J. Weston, and I. Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

[55] Q. Yu, Z. He, S. Li, X. Zhou, J. Zhang, J. Xu, and D. He. Enhancing auto-regressive chain-of-thought through loop-aligned reasoning. *arXiv preprint arXiv:2502.08482*, 2025.

[56] Z. Yue, B. Jin, H. Zeng, H. Zhuang, Z. Qin, J. Yoon, L. Shang, J. Han, and D. Wang. Hybrid latent reasoning via reinforcement learning. *arXiv preprint arXiv:2505.18454*, 2025.

[57] J. Zhang, Y. Zhu, M. Sun, Y. Luo, S. Qiao, L. Du, D. Zheng, H. Chen, and N. Zhang. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*, 2025.

[58] Z. Zhang, X. He, W. Yan, A. Shen, C. Zhao, S. Wang, Y. Shen, and X. E. Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025.

[59] T. Zheng, H. Zhang, W. Yu, X. Wang, X. Yang, R. Dai, R. Liu, H. Bao, C. Huang, H. Huang, et al. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*, 2025.

[60] H. Zhu, S. Hao, Z. Hu, J. Jiao, S. Russell, and Y. Tian. Emergence of superposition: Unveiling the training dynamics of chain of continuous thought. *arXiv preprint arXiv:2509.23365*, 2025.

[61] H. Zhu, S. Hao, Z. Hu, J. Jiao, S. Russell, and Y. Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *arXiv preprint arXiv:2505.12514*, 2025.

[62] Y. Zhuang, L. Liu, C. Singh, J. Shang, and J. Gao. Text generation beyond discrete token sampling. *arXiv preprint arXiv:2505.14827*, 2025.