《专业英语》期末大报告

2025-2026 学年第一学期

论文题目：From Probabilistic Mapping to Deliberate Reasoning: A Survey on the Evolution of Large Language Model Inference

关键字 Reasoning Foundation Models, Process Reward Models, Test-Time Compute, Attention Gating

姓　　名：王锦政

学　　号：2024270223

撰写时间：2025 年 12 月 21 日

撰写须知：

1. 请严格按照模板提供的一级大标题撰写报告，可根据需要可自行添加二级、三级小标题。不要更改报告格式、字体等，确保报告格式的统一性。

2. 报告格式统一规定如下：（1）正文部分中文字采用**楷体**、**小四**，西文字采用 **Times New Roman**；（2）若出现分级，各级标题的字号从小二依次递减；（3）正文行间距采用 **1.2 倍行距，**。
3. 大报告的评分主要依据是：格式排版、内容的创新性、逻辑自恰、独立思考等。其中，摘要的内容主要从逻辑上的完整性（背景、挑战、亮点、实验结果等）、创新性和丰富性三个方面考察，独立思考则体现在对摘要的解析过程，考察是否有对所摘要内容有着独到的思考。**摘要解析中必须包含对于摘要中涉及到的背景、亮点、实验结果展开解析，体现结构衔接的合理性、方法本身可能的创新性**。
4. 报告提交时请转化成 **PDF 格式**，并按“姓名_期末报告.pdf”方式命名。
5. 模板中灰色斜体字为报告撰写说明，不要删除或者编辑。
6. 报告提交的截止时间为：**2025 年 12 月 28 日晚 23:59**。

摘　要

请结合课堂的摘要写作知识，撰写和题目相应的摘要，包含但不限于①研究背景及意义、②研究现状及进展、③研究问题及方法、④主要创新点及⑤主要研究结果，。注意：请不要用 AI 写作，查到痕迹 0 分处理；请检查逻辑上是否自恰，允许使用 AI 润色（课堂上讲过 AI 写作和润色的区别）。

1000 单词以上

From Probabilistic Mapping to Deliberate Reasoning: A Survey on the Evolution of Large Language Model Inference

(Note: This survey is written from the perspective of the 2025 technological landscape.)

Abstract

1. The Cognitive Gap and the Limits of Statistical Correlation

Developing AI systems that go beyond syntax to perform causal and compositional reasoning is a key step toward Artificial General Intelligence **(AGI) [1]**. The complex interdependencies of logical premises, mathematical constraints, and real-world dynamics should be modeled as a structured hierarchy of 'thoughts' rather than probabilistic token sequences, but as a structured hierarchy of 'thoughts' governed by internal deduction and verification **[2]**. Over the past decade, the 'Scaling Law' paradigm has dominated the field, supported by empirical evidence that increasing parameter counts and data volume yields a consistent, power-law reduction in model perplexity **[3], [4]**. While this paradigm has yielded models like GPT-4 that exhibit remarkable 'System 1' intuition—characterized by rapid, heuristic-based processing **[5], [6]**—these systems fundamentally remain sophisticated statistical approximators. However, a gap remains between their factual retrieval capabilities and logical inference in novel contexts, frequently resulting in hallucinations when tasks demand multi-step planning or counterfactual simulation **[7]**. This limitation highlights the necessity for a paradigm shift: moving from models that simply predict the next token to models that engage in deliberate, compute-intensive reasoning processes before committing to an answer.

2. The Rise of Reasoning Foundation Models and System 2 Architectures

To address these limitations, Reasoning Foundation Models have emerged as a new paradigm. These systems aim to mimic 'System 2' cognitive processes—characterized by slow, deliberate, and algorithmic reasoning. Unlike previous models that primarily rely on next-token prediction, these systems generate outputs through a latent, iterative optimization process that integrates chain-of-thought generation **[8]**, self-verification, and dynamic resource allocation.

In 2025, researchers introduced architectural primitives that revisit standard Attention mechanisms **[9]**. Most notably, the 'Attention Gating' mechanism, proposed by the Alibaba Qwen Team **[10]**, addresses the 'Attention Sink' phenomenon inherent in standard Softmax attention—where normalization constraints force the model to assign redundant attention scores to initial tokens.. By incorporating a learnable, head-specific sigmoid gate immediately following the Scaled Dot-Product Attention **(SDPA)**, the model introduces significant input-dependent sparsity. Consequently, it reduces irrelevant context, i.e., noise and activation outliers, facilitating stable and efficient reasoning over long contexts.

3. Scope of This Survey: A Unified Perspective on Reasoning

This paper reviews the emerging landscape of Reasoning Foundation Models, covering architectural innovations and inference-time algorithms. This survey focuses on three key aspects:

Architecture Designs for Stability and Efficiency. We first examine the evolution of the Transformer backbone. Moving beyond standard Softmax Attention, we explore the **resurgence** of Recurrent Neural Networks **(RNNs)** via State Space Models **(SSMs)** such as **Mamba [12]**, and their convergence with Transformer architectures through mechanisms like Gated Linear Attention **(GLA) [11]**. We analyze how the introduction of data-dependent gating mechanisms, such as Qwen's recent work **[10]**—mitigates the "distraction" problem found in standard Transformers.. This capability enables models to selectively discard irrelevant context while retaining critical axioms required for subsequent reasoning. Furthermore, we compare these architectures in terms of their theoretical expressivity and practical hardware efficiency on modern GPUs.

Training Objectives: From Next-Token to Process Supervision. We survey the transition from standard self-supervised learning to rigorous alignment methodologies tailored for reasoning. We detail the strategies behind models such as OpenAI o1 **[14]** and DeepSeek-R1 **[15]**, which leverage Reinforcement Learning on Reasoning Traces. Unlike traditional RLHF, which relies on Outcome Supervision **[16]**, these methods employ Process Reward Models **(PRMs)** to evaluate the validity of intermediate reasoning steps **[17]**. This paradigm incentivizes the model to backtrack, self-correct, and explore alternative logical paths, effectively instilling a metacognitive capability. Additionally, we provide a taxonomy of data synthesis techniques used to train these verifiers, including **Self-Instruct [18]** and paths generated via Monte Carlo Tree Search **(MCTS)**.

Inference Dynamics and Test-Time Compute. We discuss the newly established scaling law: **Test-Time Compute Scaling [19]**. This principle suggests that increasing computational resources during inference (e.g., by generating and filtering thousands of candidate trajectories) can yield performance gains comparable to scaling model parameters by orders of magnitude during pre-training. We review prompting frameworks such as Chain-of-Thought **(CoT) [8]**, Tree of Thoughts (**ToT) [21]**, and Graph of Thoughts **(GoT) [22]**, analyzing them not merely as prompt engineering heuristics, but as explicit approximations of search algorithms executed within the semantic space.

4. Open Challenges and Future Directions

Despite these advances, we identify four key challenges for future research: for Reasoning Foundation Models:

The Faithfulness of Reasoning Traces. While models such as o1 generate explicit reasoning paths, a central question remains: do these traces accurately reflect the model's internal decision-making process, or do they constitute mere post-hoc rationalizations? We specifically address the risk of **Chain-of-Thought Steganography [23]**, where models may encode hidden computations within ostensibly human-readable reasoning steps, thereby undermining true interpretability.

Inference Efficiency and Latency. Shifting to Test-Time Compute inevitably increases inference costs and latency. To balance reasoning depth and energy efficiency, we suggest integrating Early Exit mechanisms **[24]** and Adaptive Computation Time

(ACT) [25]. These techniques allow models to dynamically allocate computational resources proportional to the complexity of the problem.

Generalization to Embodied Agents. A substantial gap exists between textual logical reasoning and the sensorimotor grounding required for physical dynamics. We explore how architectural innovations, specifically Gated Attention mechanisms, might be pivotal in filtering high-frequency sensory noise in robotics, effectively bridging the divide between abstract reasoning and real-world control.

Adversarial Robustness and Logical Safety. We highlight the susceptibility of reasoning models to Logical Jailbreaks **[26]**. Unlike traditional adversarial attacks, these are prompts designed to exploit the model's own deductive capabilities, triggering fallacious reasoning loops that bypass safety alignment. This necessitates the development of robust verification frameworks capable of detecting semantic, rather than just syntactic, adversarial patterns.

5. Conclusion

This survey reviewed recent advancements in Reasoning Foundation Models, from mathematical proving to software engineering. By analyzing the shift from probabilistic mapping to process-supervised reasoning, we outline a roadmap for future Foundation Models. The convergence of architectural gating mechanisms and reinforcement learning for reasoning suggests the emergence of 'System 2' AI-machines that do not just speak, but think.

摘 要 解 析

请对上文的摘要写作进行解析，围绕内容本身的创新性、可能存在的贡献、逻辑层次之间的衔接、逻辑前后呼应、自恰等方面展开。中文字数要求不少于800字（仅包括撰写的字数，不含模板本身的说明字数）。

1. 创新性论述

请阐述该摘要下的背景与动机，在此之下的创新性、实用性、前瞻性如何。

（1） 背景与动机：在过去数年间，以 GPT-4 为代表的模型通过 Scaling Law 实现了惊人的语言能力，但这种基于下一个词预测的机制本质上还是在做文字接龙，属于心理学卡尼曼认知体系中的 System 1（快思考）。但是要实现真正的 AGI，要求模型具备 System 2（慢思考）的能力，即模型需要具备复杂的因果推理、数学证明及长程规划能力。基于此写了这篇综述摘要：LLMs 从概率预测向过程推理的范式转移；

（2） 创新性与前瞻性：Transformer 架构从 2017 年提出至今，已经统治了 AI 领域近 8 年，本篇对 Transformer 变体进行了简单罗列，同时说明了 2025 年最新的成果：Qwen 团队在 Attention 里引入了门控机制，赋予了模型稀疏性与非线性过滤的能力。本篇欲在 LLMs 的认知科学和底层架构建立一座桥梁；

（3） 概念定义的重构：从大语言模型（LLM）到推理基座（RFM），新一代模型的架构、训练、推理不再是为了预测下一个 token，而是为了最大化推理过程的正确性；

（4） 实用价值：评价维度的迁移，在过去十年，Scaling Law 几乎等同于堆砌参数量。摘要创新性地指出了 Test-Time Compute Scaling 这一新定律的崛起。它打破了模型越大越智能的固有印象，提出了思考越久越智能的新可能，预示了未来 AI 研究的焦点将从预训练（Pre-training）转向推理阶段（Inference-time）。

2. 亮点论述

请按照课堂平时的论文赏析过程，阐述摘要提出的亮点如何有效支撑论文题目以及上述的创新性。

（1） 结合了宏观叙事与微观锚点：许多综述容易陷入空泛的宏大叙事，但该文本在论述宏观趋势时，精准地抛出了微观技术锚点。例如，在讨论注意力机制的缺陷时，并未泛泛而谈，而是具体引用了 Qwen 团队**的**“Sigmoid 注意力门控方案，指出了 Attention Sink（注意力汇聚）这一具体的现象。这种“以小见大的处理方式，使得关于长文本推理稳定性的论述显得由其扎实，展示了作者对前沿技术细节的精准掌控力；

（2） 对过程监督价值的深度思考：文本明确区分了结果监督（Outcome Supervision）**与**过程监督（Process Supervision）的本质差异，这是理解 OpenAI o1 和 DeepSeek-R1 等模型成功的关键。将 PRM 描述为赋予模型元认知（Metacognitive）能力，即自我纠错和回溯的能力。这一论断极具洞察力，深刻揭示了新一代模型为何能解决复杂数学问题的根本原因，即从蒙对答案进化到了懂步骤；

（3） 批判性反思：在摘要的 challenge 部分，文本抛出了 Chain-of-Thought Steganography 这一概念，模型生成的推理步骤究竟是真实的思考过程，还是为了取悦人类而生成的事后合理化（Post-hoc Rationalizations）？虽然引入 PRM 和 CoT，让模型像是学会了思考，但这并没有改变深度学习的根本性质。（它本质上还是一个黑盒，背后依然是不可解释的神经网络权重矩阵。虽然近些年关于可解释性的工作很多，但大部分还是偏主观性的解释，未来这个问题如何解决，我个人也比较期待。）

3. 逻辑解析

请从写作逻辑方面深度分析摘要的结构，必要时候可分段分层，解析这么安排的合理性，如何做到论点的支撑以及前后有效呼应。

摘要的逻辑结构我主要参考了 TPAMI：Foundation models defining a new era in vision: A survey and outlook **[27]** 这篇 paper，逻辑主线是问题提出-方案拆解-深入剖析-反思展望，具体分析如下：

（1） 逻辑起点：开篇通过对比建立了核心冲突：现有的大语言模型本质上只是在做文字接龙（System 1），在面对因果问题与复杂问题时存在认知鸿沟。指出了模型 Hallucinations 的原因，这为后文提出推理基座模型提供了充分的逻辑必要性；

（2） 解决方案的立体展开：在 Scope 部分，我采用了底层、中层、顶层逻辑结构。

- **底层架构部分：** 讨论 SSM、Mamba 和 Qwen 的门控机制，解决模型的效率和 memory 问题；

- **中层训练部分：** 讨论 RLRT 和过程监督，模型的能力获取，如何学会思考；
- **顶层推理部分：** 讨论 CoT 和搜索算法，如何展示思考；

（3） 辩证：在 Challenges 部分，没有回避新范式可能带来的副作用：推理效率的下降（Thermodynamic Efficiency）和安全隐患（Logical Jailbreaks）等；

（4） 呼应：最后，Conclusion 部分实现逻辑的首尾呼应：开头提出模型受限于语法解析（parse syntax）和统计模仿（statistical mimics），结尾升华至我们正在见证 not just speak, but think 的机器。

4. 作业背景

最近看到 Google 的 Nested Learning 和 Qwen 的 Gated attention 这两篇文章，比较感兴趣，所以以这个为主题写了摘要。由于本文为综述性质，实验结果部分主要体现为对现有 SOTA 模型性能的定性分析与对比，而非单一实验的定量数据。

主要逻辑架构和核心关键词（如 Test-Time Compute, System 2）是我构思的，初稿比较直白。英文摘要部分我参考了 AI 的润色，并结合课堂所学的学术写作规范进行了手动重写和调整，并添加了对应文献的引用。但文中提到的 o1 模型、DeepSeek-R1 以及最近 Qwen 的门控机制都做过具体调研，不是 AI 生成的内容。

References

[1] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[2] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv:1709.08568*, 2017.

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[4] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[5] D. Kahneman, *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux, 2011.

[6] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 24824–24837.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[10] Z. Qiu, Z. Wang, B. Zheng, Z. Huang, K. Wen, S. Yang, R. Men, L. Yu, F. Huang, S. Huang, D. Liu, J. Zhou, and J. Lin, "Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free," *arXiv preprint arXiv:2505.06708*, May 2025.

[11] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim, "Gated linear attention transformers with hardware-efficient training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[13] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[14] OpenAI, "Learning to reason with LLMs," *OpenAI Research*, Sep. 2024. [Online].Available: **https://openai.com/index/learning-to-reason-with-llms/**

[15] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv preprint arXiv: 2501.12948*, 20245

[16] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2017.

[17] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[18] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," in *Proc. Assoc. Comput. Linguist. (ACL)*, 2023, pp. 13484–13508.

[19] D. Zhang, S. Zhou, Z. Hu, Y. Yue, Y. Dong, and J. Tang, "ReST-MCTS*: LLM self-training via process reward guided tree search," *arXiv preprint arXiv:2406.03816*, 2024.

[20] C. Snell, I. Kostrikov, Y. Su, M. Yang, and S. Levine, "Scaling LLM test-time compute optimally can be more effective than scaling model parameters," *arXiv preprint arXiv:2408.03314*, 2024.

[21] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.

[22] M. Besteiro, I. G. N. Zhu, J. Liu, and A. Liu, "Graph of thoughts: Solving elaborate problems with large language models," in *Proc. AAAI Conf. Artif. Intell.*, 2024.

[23] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[24] T. Schuster, A. Fisch, J. Gupta, M. Dehghani, D. Bahri, V. Tran, R. Tay, and D. Metzler, "Confident adaptive language modeling," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[25] A. Graves, "Adaptive computation time for recurrent neural networks," *arXiv preprint arXiv:1603.08983*, 2016.

[26] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does LLM safety training fail?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

[27] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2245–2264, Apr. 2025