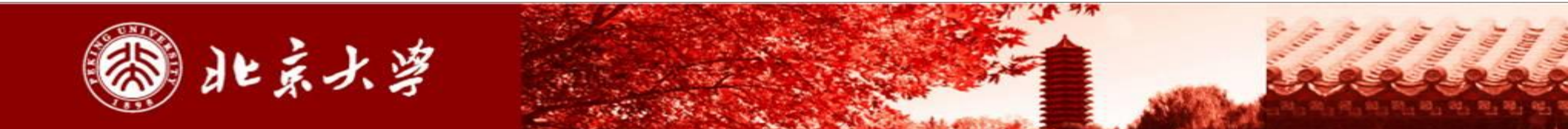
《物理与人工智能》

14. Transformer之后的发展

授课教师：马滟青

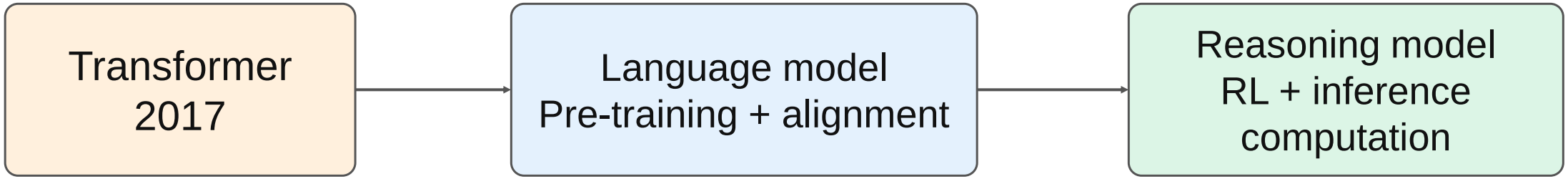
2026/10/19（第六周）



LLMs after the Transformer

Architecture, training and new capabilities

Important advances



A continuation of the attention and Transformer lecture

Sources: [1] [2]

LLM progress: nine categories

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Sources: [1] [2] [3] [4] [5]

Milestones after 2017

北京大学 PEKING UNIVERSITY

2018–2020	**GPT / BERT / GPT-3** General-purpose pre-training
2022–2023	**InstructGPT / ChatGPT / GPT-4** Instruction following
2024–2025	**GPT-4o / o1 / DeepSeek-R1** Multimodal reasoning
2026	**GPT-6 / Claude 5.5 / Gemini 3.8** Reasoning, efficiency and scientific work

Sources: [1] [2] [3] [4] [5]

01 Pre-training and scale

Section 1 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Transformer families and objectives

Scaling laws and compute allocation

In-context learning

Sources: [1] [2]

Three Transformer families

01 Pre-training and scale

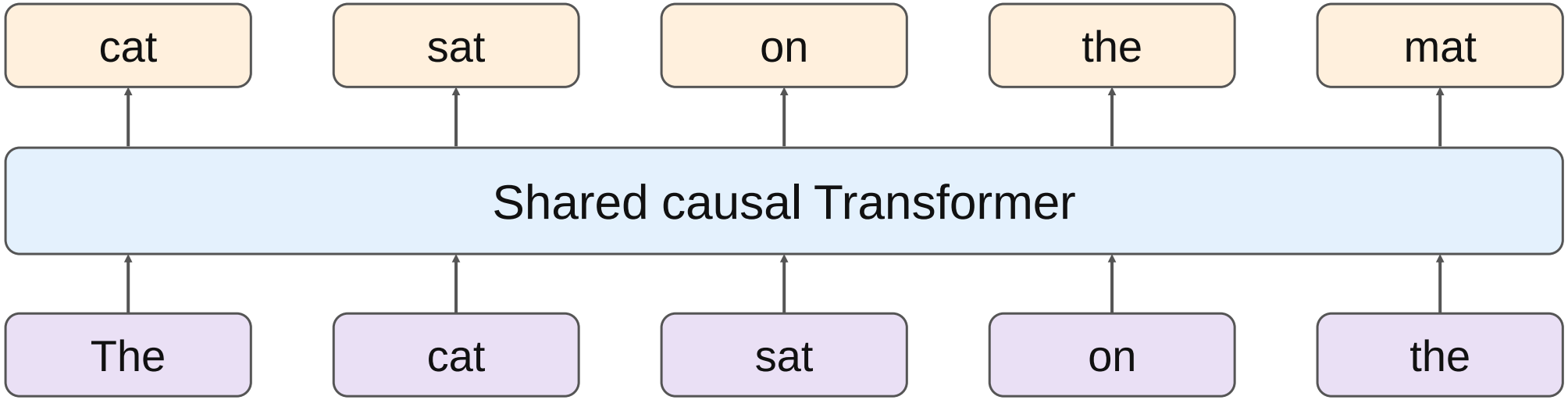
BERT	GPT	T5
Encoder only	Decoder only	Encoder + decoder
Bidirectional attention	Causal self-attention	Encoder
Feed-forward	Feed-forward	Decoder + cross-attention
Masked-token prediction Text understanding	Next-token prediction Text generation	Text-to-text learning Conditional generation

The training objective matters as much as the architecture

Sources: [1] [2] [3]

Autoregressive language modeling

Predict each token using only the tokens before it



$$L = -\Sigma_t \log p\theta(x_t \mid x_{<t})$$

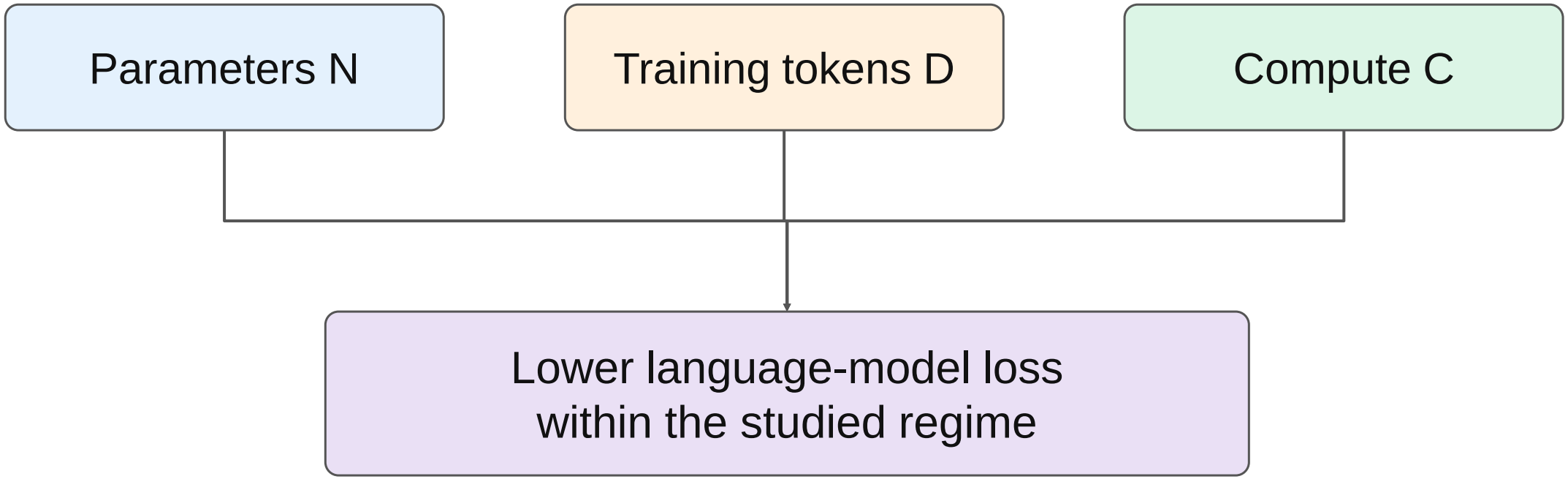
Parallel loss computation during training

Sequential decoding at inference

Sources: [1]

Scaling laws

Empirical power laws guide resource allocation



Better prediction loss does not guarantee success on every task

Sources: [1]

Compute-optimal training

Chinchilla revisited the balance between model size and data

Model	Parameters	Training data
Gopher	280B	Baseline
Chinchilla	70B	About 4× Gopher

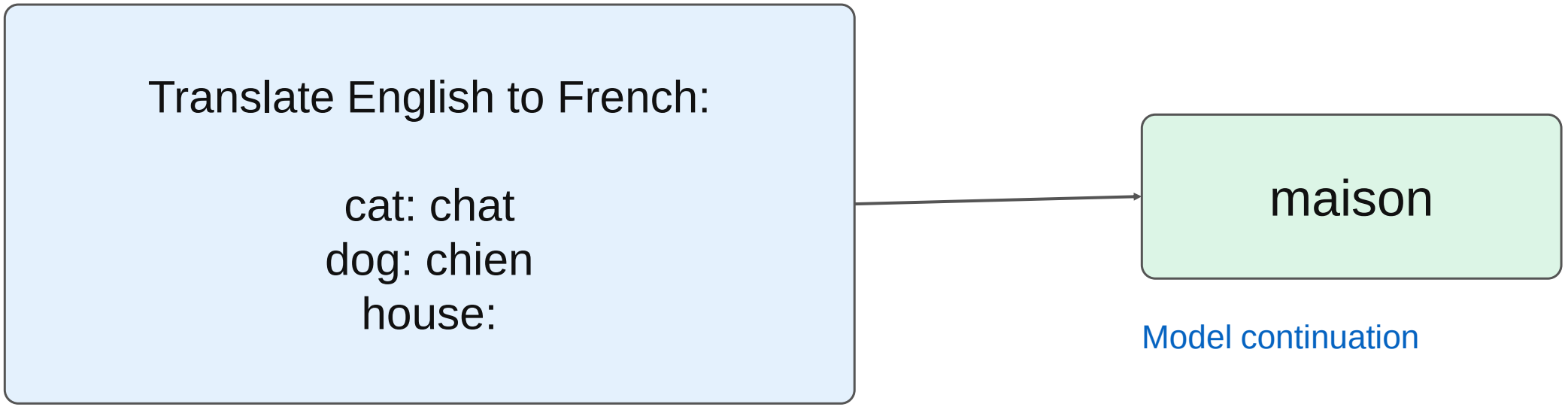
Same training compute budget

Training efficiency and deployment efficiency use different budgets

Sources: [1]

GPT-3 and in-context learning

A task can be specified by examples inside the prompt



The prompt changes. The model weights stay fixed during the request

Sources: [1]

02 Sparse and hybrid architectures

北京大学 PEKING UNIVERSITY

Section 2 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Efficient attention and state spaces

Hybrid attention layers

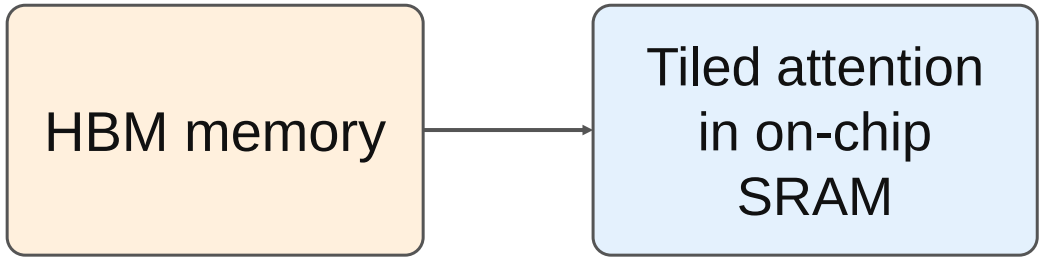
MoE routing and active parameters

Expert balance and model examples

Sources: [1] [2] [3]

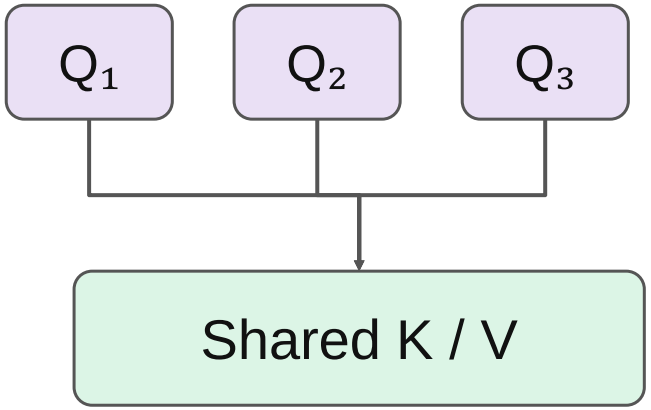
More efficient attention

FlashAttention



Less memory traffic
Exact dense attention

Grouped-query attention



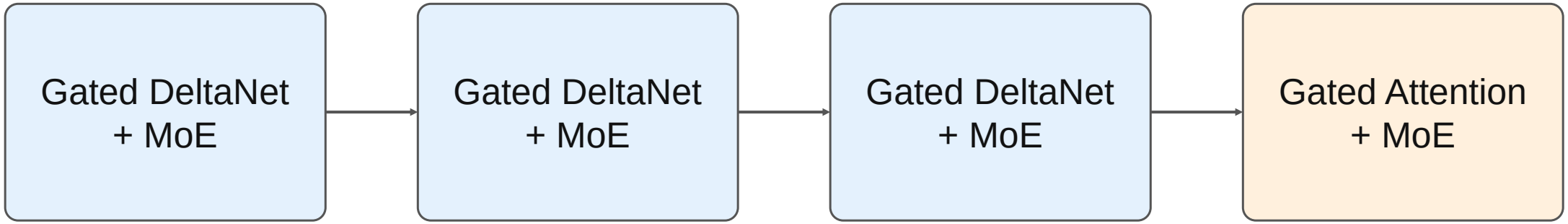
Fewer KV heads
Smaller cache during generation

Compute complexity, memory traffic and KV storage are different costs

Sources: [1] [2]

Hybrid linear attention in Qwen3-Next

48 layers = 12 repetitions of this four-layer pattern



80B total parameters

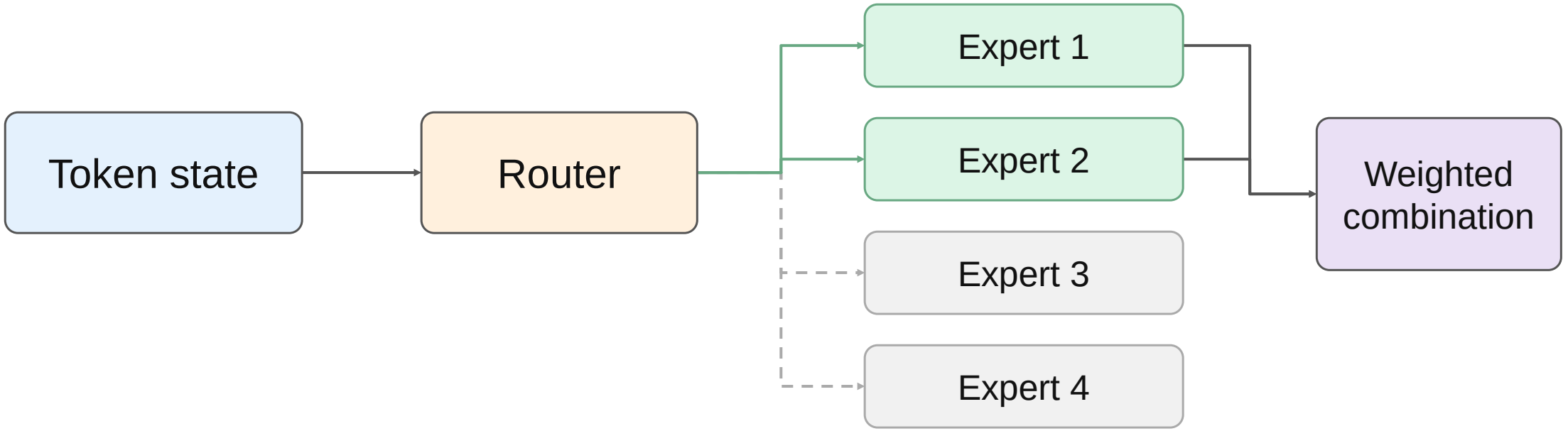
3B active per token

A specific hybrid model. Linear attention and MoE serve different roles.

Sources: [1] [2]

Mixture-of-experts models

A router activates a small subset of feed-forward experts per token

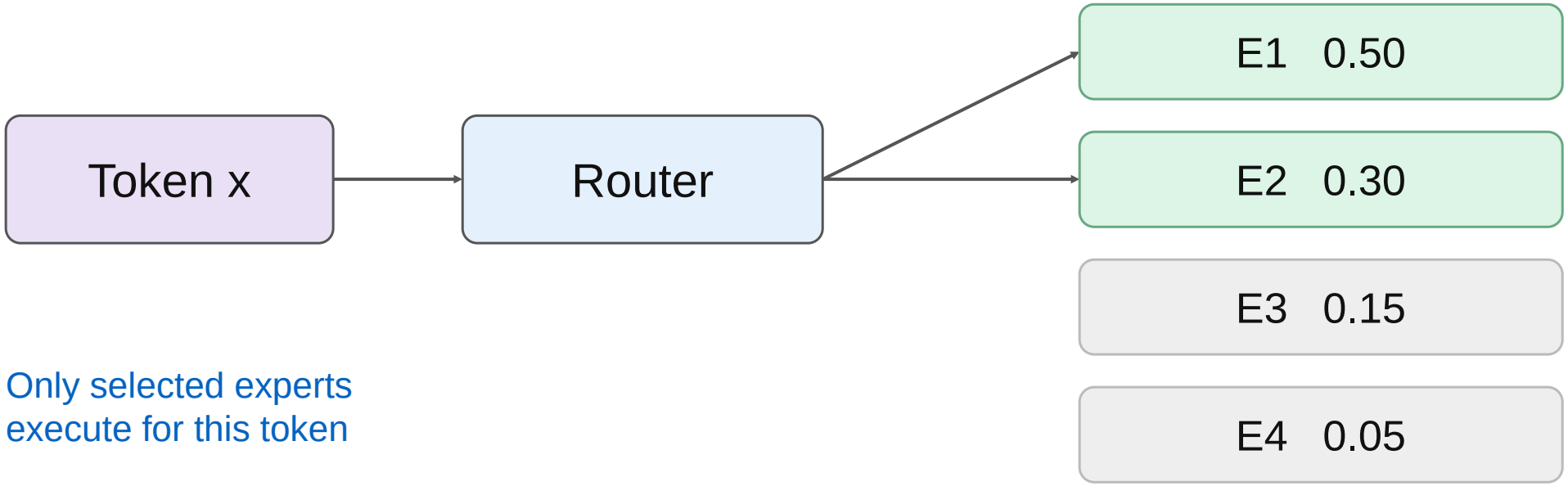


Total parameters and active parameters answer different questions

Sources: [1]

MoE routing and top-k selection

Toy example: 4 experts, top-2 routing



Only selected experts execute for this token

$$y = 0.625\ E1(x) + 0.375\ E2(x)$$

Sources: [1] [2]

MoE capacity, compute and memory

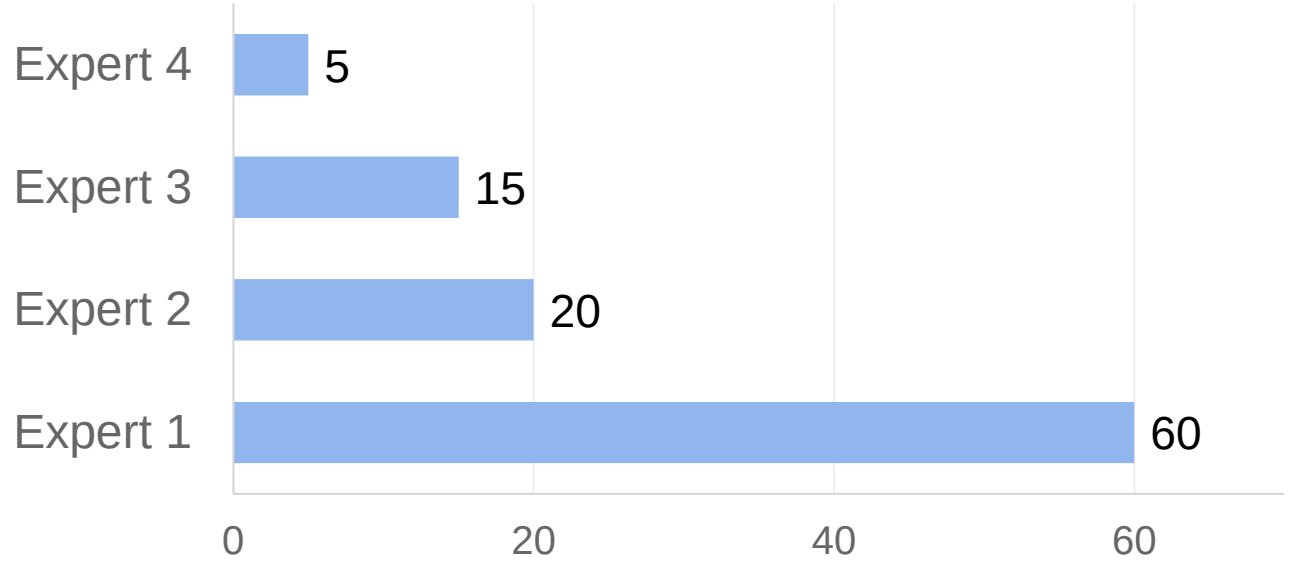
Quantity	Meaning	DeepSeek-V3
Total parameters	Weights in the main model	671B
Active parameters	Used for one token	37B
Resident weights	Storage across the system	Includes inactive experts

Latency also depends on communication, batch size and bandwidth

Sources: [1] [2]

Expert load and communication

Uneven routing can create a device bottleneck



Illustrative token counts, total = 100

Sparse compute still needs efficient distributed execution

System requirements

Load balancing
Expert capacity
All-to-all transfer
Communication overlap

Sources: [1] [2]

MoE in DeepSeek-V3 and Qwen3

Released model	Total	Active per token
DeepSeek-V3	671B	37B
Qwen3-235B-A22B	235B	22B
Qwen3-30B-A3B	30B	3B

Complementary choices in DeepSeek-V3

MLA reduces KV-cache overhead
Shared experts provide a common computation path
Multi-token prediction adds training signals

Sources: [1] [2]

03 Efficient adaptation and decoding

Section 3 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Open weights and parameter adaptation

Quantization and LoRA / QLoRA

Speculative decoding

Multi-token prediction

Sources: [1] [2] [3]

Open weights and adaptation

Open weights make controlled experiments and local adaptation possible

Approach	What changes	Typical use
Full fine-tuning	Most model weights	Deep task adaptation
LoRA	Low-rank updates	Cheaper customization
Prompt / retrieval	Input context	No weight updates

Weight access, training-data access and license terms are separate

Sources: [1] [2]

Quantization and weight memory

Illustrative 7B model: raw weights only, decimal GB

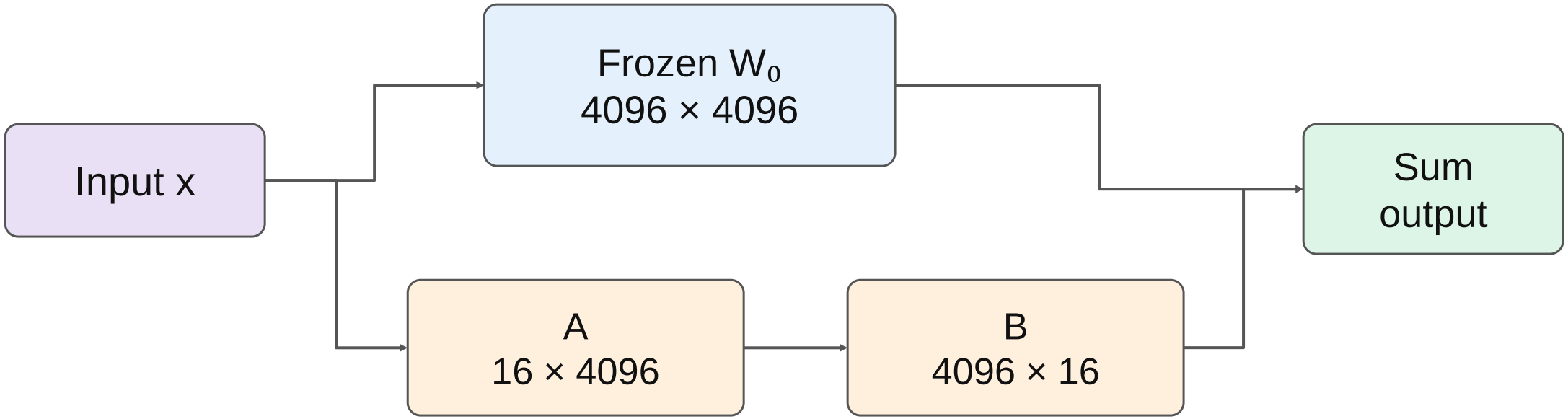
Representation	Bits per weight	Raw weight memory
FP16 / BF16	16	14 GB
INT8	8	7 GB
INT4	4	3.5 GB

Total memory also includes metadata, KV cache and activations

Sources: [1]

LoRA and QLoRA

$$W = W_0 + BA$$

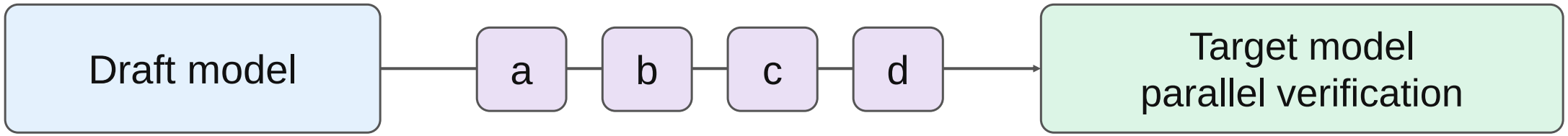


16,777,216 base weights vs. 131,072 adapter weights (0.78%)

Sources: [1] [2]

Speculative decoding

A fast draft proposes tokens. The target verifies a block.



Accepted prefix: a b c

Token d rejected. Sample from the correction distribution.

Accept with probability min(1, p(token) / q(token))

Sources: [1]

Multi-token prediction

Method	Where it acts	Main idea
Multi-token prediction	Training objective	Learn additional future positions
Speculative decoding	Inference algorithm	Draft, verify and correct
MTP as a draft source	Combined deployment	Use MTP candidates for verification

Extra prediction modules can support both learning and faster decoding

Sources: [1] [2]

04 Alignment, distillation and data

Section 4 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Instruction tuning, RLHF and DPO

Teacher–student knowledge transfer

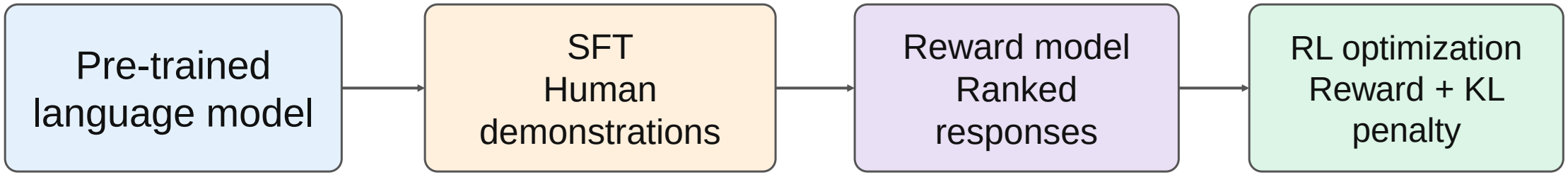
Reasoning trajectory distillation

Synthetic data quality and failure modes

Sources: [1] [2] [3] [4]

Instruction tuning and RLHF

Pre-training supplies capability. Post-training shapes behavior.



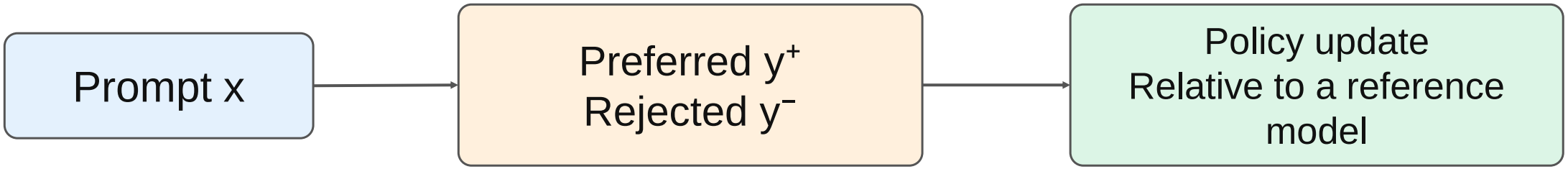
Preference data teaches which responses people favor

A preference reward is an imperfect proxy for the desired behavior

Sources: [1]

Direct preference optimization

Learn directly from preferred and rejected responses



Increase the preferred response's relative log-probability

$$\log [\pi\theta(y^+|x) / \pi\text{ref}(y^+|x)] - \log [\pi\theta(y^-|x) / \pi\text{ref}(y^-|x)]$$

Simpler optimization still depends on good preference data

Sources: [1]

ChatGPT and GPT-4

ChatGPT, November 2022

User instruction
+ dialogue history

Follow-up questions
Conversational instruction following

GPT-4, March 2023

Text and image input
Broader task coverage

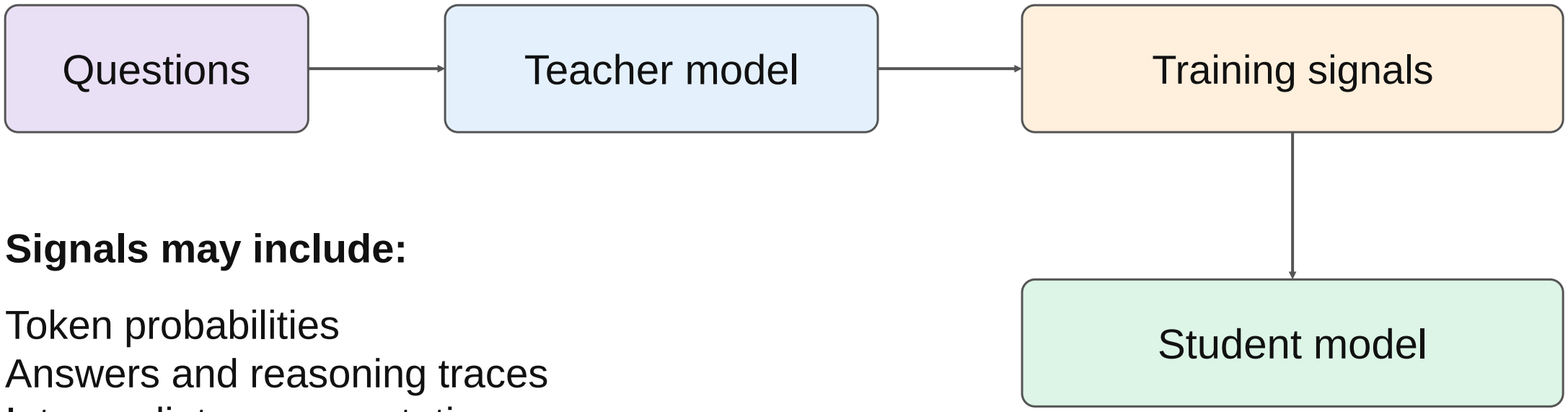
A general model for many tasks
Strong benchmark performance

Fluent answers can still contain factual errors

Sources: [1] [2]

Knowledge distillation

A smaller model learns from a stronger model's outputs



Signals may include:

Token probabilities
Answers and reasoning traces
Intermediate representations

Teacher access is needed during data creation

Sources: [1] [2]

Soft targets and temperature

$$p_i(T) = \mathrm{softmax}(z_i / T)$$

Toy logits	T = 1	T = 2
3.0	78.6%	59.0%
1.5	17.5%	27.9%
0.0	3.9%	13.2%

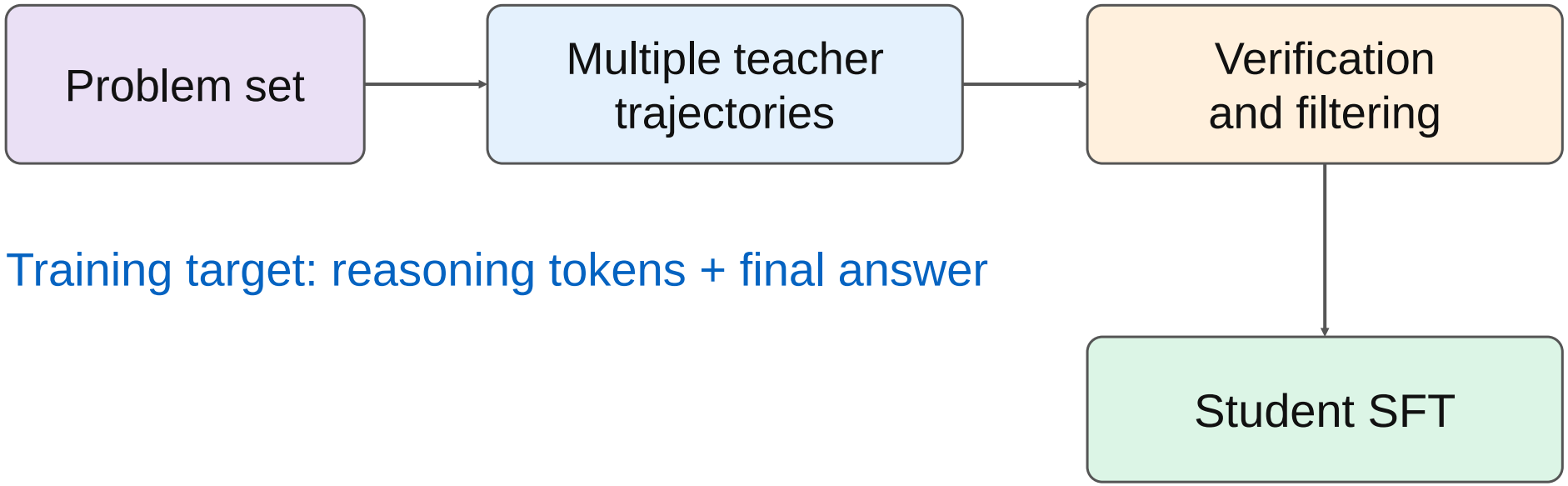
$$L = \alpha\, CE(y, pS) + (1 - \alpha)\, T^2\, KL(pT(T) \,\|\, pS(T))$$

Illustrative distribution. The logits remain fixed as T changes.

Sources: [1]

Distilling reasoning trajectories

Teacher-generated reasoning becomes supervised training data



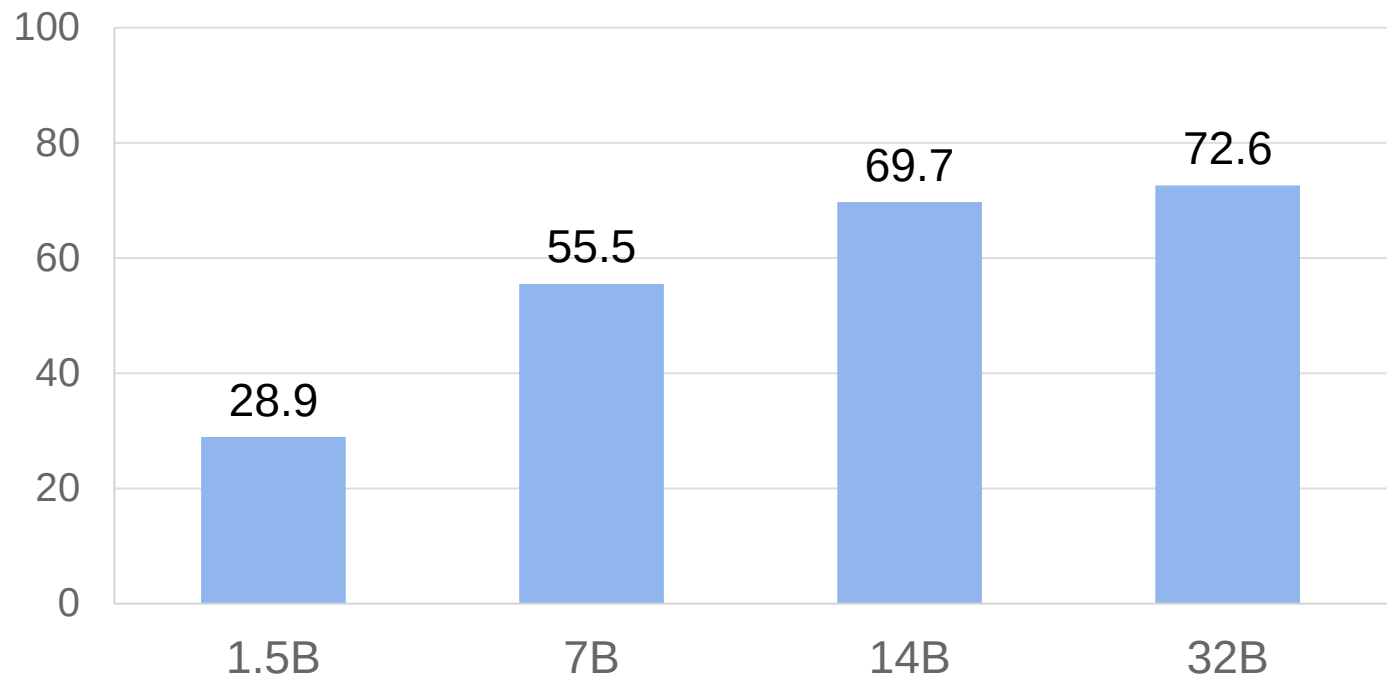
Training target: reasoning tokens + final answer

Evaluation: held-out problems and task-specific checks

Sources: [1] [2] [3]

DeepSeek-R1 distillation results

AIME 2024 pass@1 (%)



Official release
January 2025

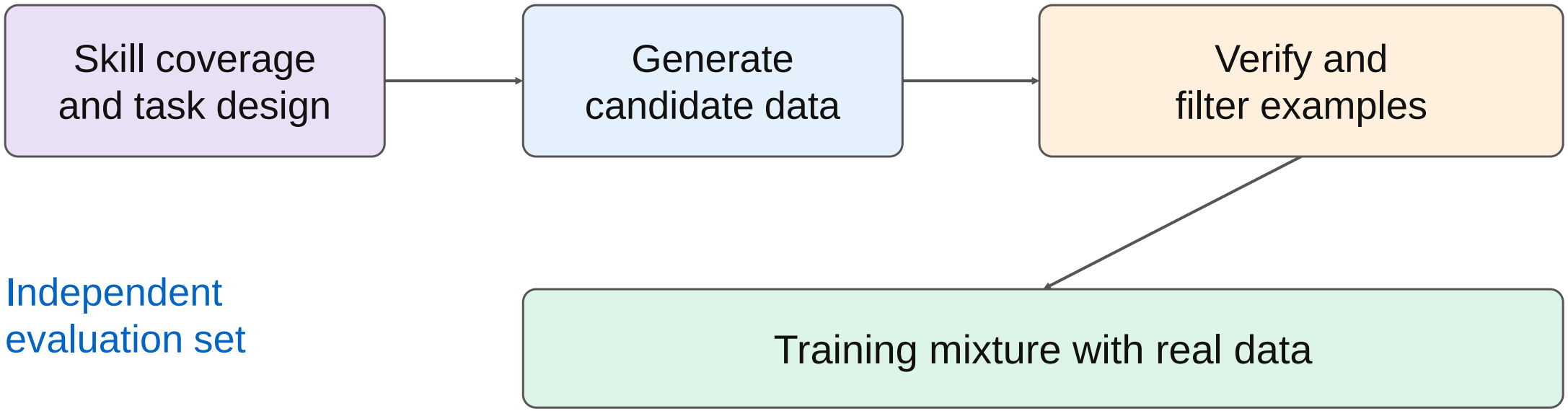
Max output: 32,768
Temperature 0.6
Top-p 0.95

64 samples per question estimate pass@1. No majority voting.

Sources: [1] [2]

Synthetic data for training

Phi-4 emphasizes data quality across training stages



Independent evaluation set

Teaching synthesis based on Phi-4 and DeepSeek-R1

Sources: [1] [2]

Synthetic data failure modes

Failure mode	Practical check
Confident but incorrect answers	Executable tests and expert review
Recursive distribution narrowing	Real data and tail coverage
Benchmark contamination	Deduplication and isolated hold-outs
Repeated templates	Diversity and difficulty audits

Risks depend on the generation and training process

Sources: [1] [2]

05 Long context and multimodality

Section 5 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Position scaling and longer sequences

Reliable use of distant information

Joint text, vision and audio training

Sources: [1] [2] [3]

Long-context language models

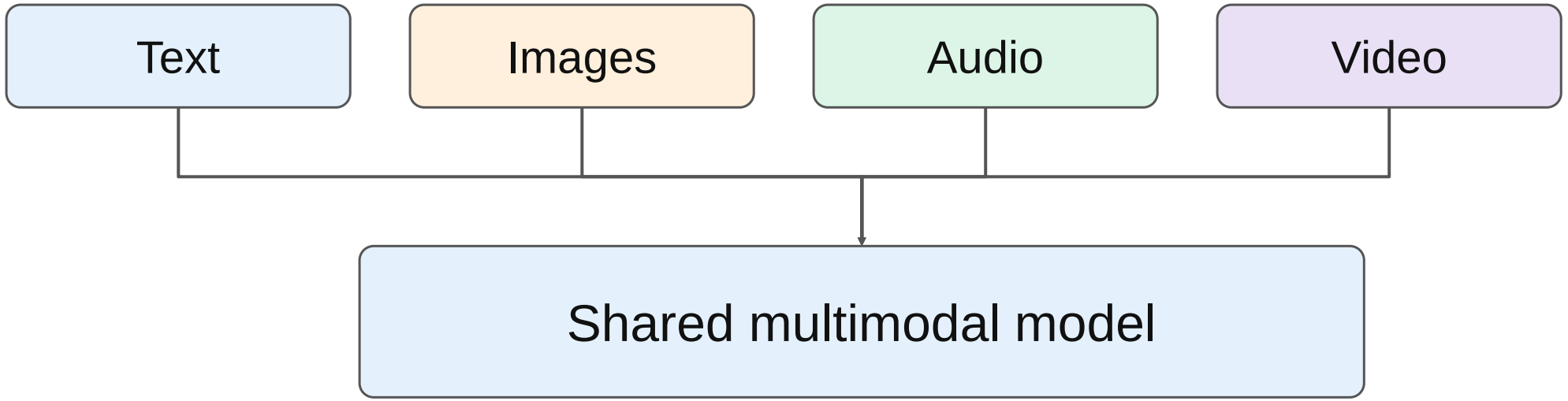
Model-level change	Purpose
Position interpolation / scaling	Extend the range of position representations
Long-sequence fine-tuning	Adapt attention to longer inputs
Information-use reliability	Use evidence across positions and lengths

Position Interpolation extended studied LLaMA models to 32,768 tokens

Sources: [1] [2]

Native multimodal models

Joint understanding supports richer interaction



GPT-4o: end-to-end training across text, vision and audio

Joint training connects language, vision and audio

Sources: [1]

06 Reasoning capability

Section 6 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Reasoning-oriented reinforcement learning

Computation allocated during inference

Verifiable rewards in DeepSeek-R1

Mathematics and formal proof results

Sources: [1] [2] [3]

Reasoning models

2022: chain-of-thought prompting

Examples with intermediate steps

Prompt-level technique

2024: RL-trained reasoning

Train a policy to spend more effort on hard problems

Training + inference technique

A written explanation still needs an independent correctness check

Sources: [1] [2]

Test-time computation

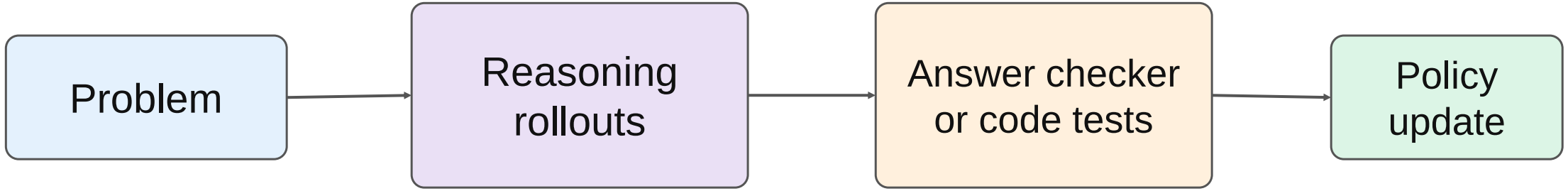
Serial reasoning	More internal steps before the answer
Parallel candidates	Generate alternatives, then select
Candidate verification	Score alternatives with a verifier

A fixed model can allocate different amounts of inference computation

Sources: [1] [2]

DeepSeek-R1 and verifiable rewards

Reinforcement learning can use correctness signals from verifiable tasks



R1-Zero: RL without initial SFT
R1: cold-start data + multi-stage training

Verifiable rewards work best where success can be checked clearly

Sources: [1]

Adaptive reasoning in general models

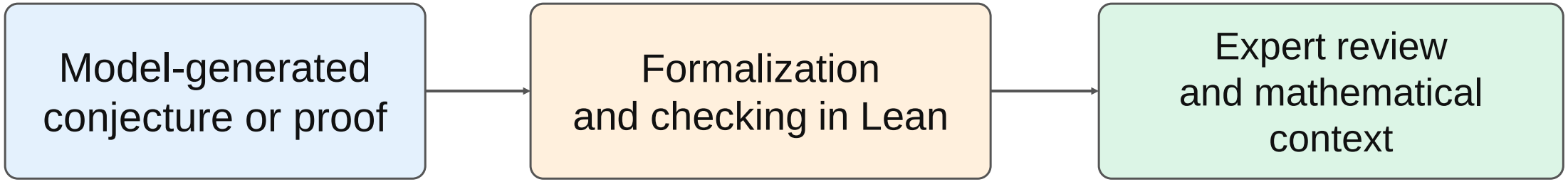
2025 milestone	What changed
DeepSeek-R1, January	Open reasoning weights and distilled models
Gemini 2.5, March	Thinking models for complex problems
GPT-5, August	Fast and reasoning models with a router

Reasoning depth becomes a controllable resource

Sources: [1] [2] [3]

Mathematics with external verification

6 October 2026: OpenAI released new mathematical results



Published artifacts include proof formalizations and research details

Machine-checked logic and expert assessment answer different questions

Sources: [1]

07 Diffusion language generation

北京大学 PEKING UNIVERSITY

Section 7 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Masked-token denoising

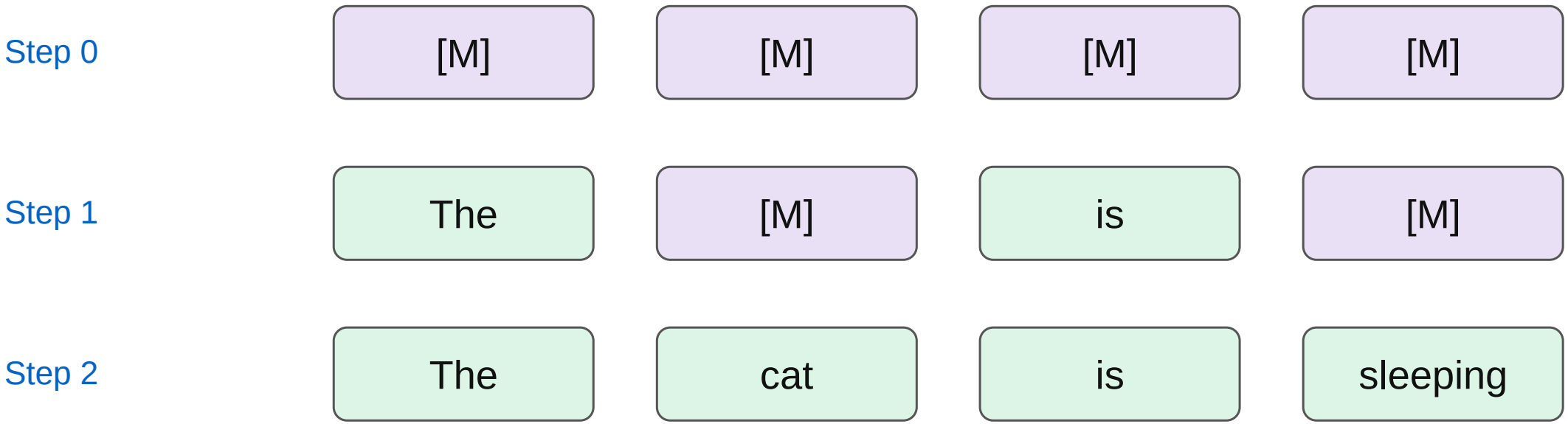
Iterative sequence revisions

LLaDA and the 2026 iLLaDA follow-up

Sources: [1] [2]

Diffusion language models

Masked-token denoising generates a sequence through revisions



Illustrative word-level example. Actual tokens and update schedules vary.

Sources: [1]

Autoregressive and diffusion routes

Dimension	Autoregressive	Masked diffusion
Training signal	Next-token prediction	Recover masked tokens
Visible context	Causal prefix	Both sides when available
Generation	Append tokens in order	Revise multiple positions
Cost trade-off	Steps and KV cache	Denoising rounds and updates

2026 example: iLLaDA extends bidirectional masked modeling

Sources: [1] [2] [3]

08 Model interpretability

Section 8 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Sparse autoencoder features

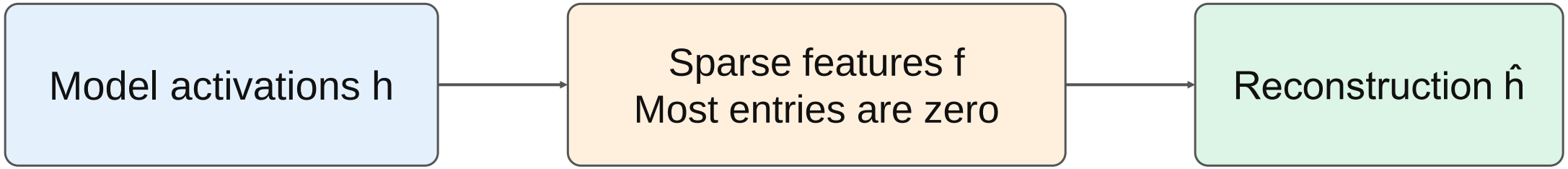
Attribution paths and circuit tracing

Causal interventions and remaining limits

Sources: [1] [2]

Sparse autoencoder features

Sparse features can separate patterns mixed in activations



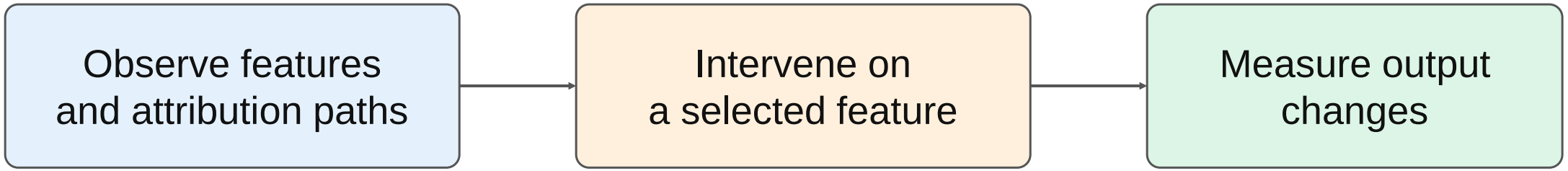
Reconstruction loss + sparsity penalty

Feature labels summarize observed examples and remain incomplete

Sources: [1]

Circuit tracing and causal tests

Written reasoning may differ from the internal computation



Limits: incomplete feature coverage, approximation errors and prompt scope

Attribution hypotheses need causal tests and replication

Sources: [1] [2] [3]

09 Model progress in 2026

Section 9 of 9

01 Pre-training and scale

02 Sparse and hybrid architectures

03 Efficient adaptation and decoding

04 Alignment, distillation and data

05 Long context and multimodality

06 Reasoning capability

07 Diffusion language generation

08 Model interpretability

09 Model progress in 2026

Topics in this section

Selected September–October releases

Reasoning, factuality and token efficiency

Architectural changes in open models

Sources: [1] [2] [3] [4] [5] [6]

Recent releases in 2026

09 Model progress in 2026

Date	Release	Main emphasis
02 Sep	Gemini 3.8 Flash / Cyber	Agents and efficient reasoning
03 Sep	GPT-6 Astra	Computer use and scientific work
10 Sep	DeepSeek V4.1-Flash	Asymmetric MoE and vision
22 Sep	Claude Opus 5.5	Long tasks with lower cost
28–29 Sep	Sonnet 5.5 / GPT-6.1 Sol	Capability at lower task cost
07 Oct	Haiku 5.5 / GPT-6 in ChatGPT	Broader, cheaper access

Selected official releases available by 13 October 2026

Sources: [1] [2] [3] [4] [5] [6] [7] [8]

GPT-6: reported model progress

09 Model progress in 2026

Release	Reported progress
Astra, 3 Sep 2026	Stronger mathematical and scientific reasoning
6.1 Sol, 29 Sep 2026	Improved factuality and complex-document understanding
Sol / Luna, 7 Oct 2026	Further updates to the GPT-6 model family

Different releases illustrate capability gains across a model family

Sources: [1] [2] [3]

Claude 5.5: capability and efficiency

Released model	Reported progress
Opus 5.5, 22 Sep	Stronger complex reasoning and model alignment
Sonnet 5.5, 28 Sep	Improved image understanding and faster generation
Haiku 5.5, 7 Oct	Stronger capabilities in a smaller, faster model

The model family advances both capability and generation efficiency

Sources: [1] [2] [3]

Gemini 3.8: reasoning and coding

Released 2 September 2026

Shared model progress	Published direction
Multi-step reasoning	More reasoning steps on difficult problems
Software understanding	Stronger performance on complex coding tasks
Security specialization	Flash Cyber focuses on vulnerability detection

The two variants share a foundational model capability

Sources: [1]

Open models: architectural progress

Model	Architecture	Parameter accounting
DeepSeek V4.1-Flash 10 Sep 2026	Causal encoder–decoder MoE Native vision	552B total 8B input / 16B output active
Qwen3.8-Flash-Next Aug 2026	Hybrid linear + sparse attention N-gram embeddings	125B LM, 6B active + 51B embedding + 4B MTP

Low active compute still requires storing or offloading the full weights

Sources: [1] [2]

Remaining model research challenges

Reasoning generalization	Transfer reasoning beyond familiar training patterns
Long-context use	Retain accuracy across positions and sequence lengths
Faithful explanations	Connect readable reasoning to internal computation
Training data quality	Control synthetic errors and distribution coverage

Sources: [1] [2] [3] [4]

Key takeaways

Pre-training	Scaling, data quality and compute allocation
Architecture	Sparse experts, hybrid attention and efficient decoding
Post-training	Preference learning and knowledge distillation
New capabilities	Reasoning, multimodality and interpretable features

Sources: [1] [2] [3] [4]

References: foundations

Radford et al., Improving Language Understanding by Generative Pre-Training, 2018

Paper / official source

Devlin et al., BERT, 2018 / NAACL 2019

Paper / official source

Raffel et al., Exploring the Limits of Transfer Learning, 2019 / JMLR 2020

Paper / official source

Brown et al., Language Models are Few-Shot Learners, NeurIPS 2020

Paper / official source

Kaplan et al., Scaling Laws for Neural Language Models, 2020

Paper / official source

Hoffmann et al., Training Compute-Optimal Large Language Models, 2022

Paper / official source

Hu et al., LoRA, 2021 / ICLR 2022

Paper / official source

Touvron et al., LLaMA, 2023

Paper / official source

Dao et al., FlashAttention, NeurIPS 2022

Paper / official source

Ainslie et al., GQA, EMNLP 2023

Paper / official source

Sources: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

References: training and generation

Ouyang et al., Training language models to follow instructions with human feedback, 2022

Paper / official source

Rafailov et al., Direct Preference Optimization, NeurIPS 2023

Paper / official source

Fedus et al., Switch Transformers, 2021 / JMLR 2022

Paper / official source

Hinton et al., Distilling the Knowledge in a Neural Network, 2015

Paper / official source

DeepSeek-AI, DeepSeek-R1, Jan 2025

Paper / official source

Microsoft Research, Phi-4 Technical Report, Dec 2024

Paper / official source

Gu and Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023

Paper / official source

Leviathan et al., Fast Inference from Transformers via Speculative Decoding, 2022 / ICML 2023

Paper / official source

Nie et al., Large Language Diffusion Models, 2025

Paper / official source

Anthropic, Tracing the thoughts of a large language model, 27 Mar 2025

Paper / official source

Sources: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

References: recent progress

OpenAI, GPT-6 Astra, 3 Sep 2026

Paper / official source

OpenAI, Introducing GPT-6.1 Sol, 29 Sep 2026

Paper / official source

OpenAI, GPT-6 Sol and Luna: October 2026 update, 7 Oct 2026

Paper / official source

Anthropic, Claude Opus 5.5, 22 Sep 2026

Paper / official source

Anthropic, Claude Sonnet 5.5, 28 Sep 2026

Paper / official source

Anthropic, Claude Haiku 5.5, 7 Oct 2026

Paper / official source

Google DeepMind, Gemini 3.8 Flash and Flash Cyber, 2 Sep 2026

Paper / official source

DeepSeek, V4.1-Flash, 10 Sep 2026

Paper / official source

Qwen, Qwen3.8-Flash-Next model card, Aug 2026

Paper / official source

OpenAI, Sharing AI progress in mathematics, 6 Oct 2026

Paper / official source

Sources: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

谢谢

北京大学
PEKING UNIVERSITY

