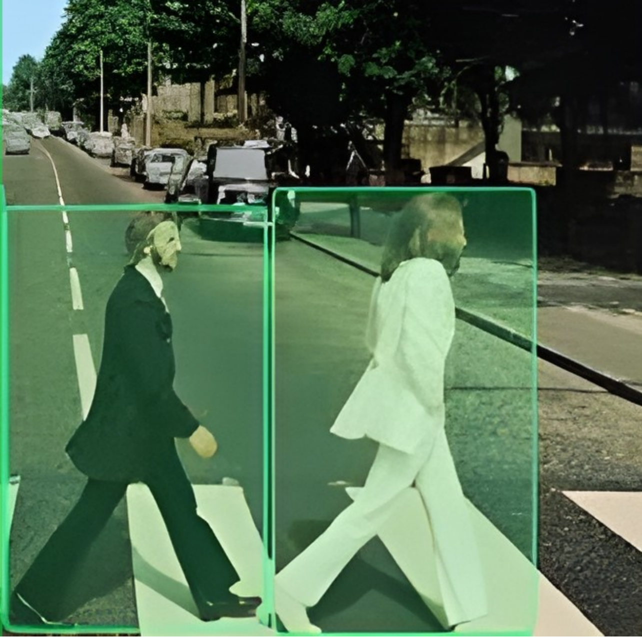
Computer Vision - 2025

Week #14. Approaching AGI

Lectures by Alexei Kornaev [1,2,3]
Practical sessions by Kirill Yakovlev [2]

[1]AI Institute, Innopolis University (IU), Innopolis
[2]Robotics & CV Master's Program, IU, Innopolis
[3]Dept. of M^2R, Orel State University, Orel
[4]RC for AI, National RC for Oncology, Moscow



April 21, 2025

Agenda

Section 1. Introduction

Just A Few Terms

Artificial General Intelligence (AGI)

Refers to a machine's ability to understand, learn, and perform any intellectual task that a human can, across diverse domains, without task-specific programming.

AI Agents

Systems that perceive their environment (e.g., via CV, sensors), make decisions (via planning/RL), and act autonomously (e.g., robots, chatbots)

Alignment

Ensuring AI systems pursue the intended goals and human values

Figure: Bender, a fictional robot from Futurama, humorously exemplifies the alignment problem in AGI

Just a Few Questions

1. **How would it look like** to live in a world where there is AGI?
2. **What would you trust** to your own AI agents to do?
3. **What objective would you choose** to make AGI aligned?

Section 2. Outcomes

Outcomes

This week's lecture explores AGI's trajectory, challenges, and societal implications. By the end, you will:

1. **Analyze** the state of modern AI (LLMs: DeepSeek, ChatGPT) and their limitations toward AGI.
2. **Design** strategies for AI agents as professional tools (e.g., robotics, healthcare co-pilots).
3. **Debate** future scenarios using forecasts from Nobel laureates and the AI 2027 Report.

Key Takeaway: AGI demands both technical innovation and ethical foresight.

Section 3. Where Are We Now?

Modern LLM Training Stages. Stage 1: Pretraining

Objective

Learn general language patterns from vast text corpora (e.g., 2T tokens for DeepSeek-V3). **Loss Function**: Next-token prediction via cross-entropy:

$$\mathcal{L}_{\mathrm{PT}} = -\sum_{t=1}^{T} \log P(x_t | x_{<t}; \theta)$$

Key Techniques

- **Best-Fit Document Packing**: Minimize padding by concatenating text chunks (e.g., 4096-token blocks).
- **FlashAttention-2**: Optimizes GPU memory usage for self-attention:

$$\text{FLOPs} \propto N^2 d + N d^2 \quad (N = \text{seq len}, d = \text{hidden dim})$$

- **Scaling Laws**: Compute-optimal allocation (Chinchilla: 20 tokens per parameter).

Stage 2: Supervised Fine-Tuning (SFT)

Process

Fine-tune pretrained model on curated datasets (e.g., 100K human-written examples). **Loss Function**:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(x,y)\in\mathcal{D}_{\text{SFT}}} \log P(y|x;\theta)$$

Dataset Design

Task	Examples
Math	GSM8K, MATH, AIME
Coding	HumanEval, CodeContests
Safety	HarmlessQA, adversarial prompts

Stage 3: RLHF

Standard RLHF Pipeline

1. **Reward Modeling**: Train RM on 100K+ human rankings:

$$\mathcal{L}_{\text{RM}} = -\sum_{(x, y_w, y_l)} \log \sigma(r_\phi(y_w|x) - r_\phi(y_l|x))$$

2. **PPO Fine-Tuning**: Optimize policy π_θ:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}\left[\min\left(r_t \hat{A}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] - \beta \text{KL}(\pi_\theta || \pi_{\text{ref}})$$

Challenges

- **KL Collapse**: Over-optimization on RM signals.
- **Reward Hacking**: Models exploit RM flaws (e.g., verbosity).

Please to meet you: DeepSeek Model Family

Core Models & Technical Specifications

Model	Training Method	Key Innovation	Use Case
DeepSeek-Zero	Pretraining (2T tokens) $\mathcal{L}_{\text{PT}} = -\sum \log P(x_t \| x_{<t})$	Base transformer 128K context	Foundation for V3/R1 Text completion
DeepSeek-V3	SFT + RLHF 100K human examples	MLA attention 37B active MoE params	General-purpose AI Chat, coding
DeepSeek-R1	GRPO + Rule-based RL $R(y) = 0.7 \cdot \text{Correctness} + 0.3 \cdot \text{Readability}$	Structured CoT rewards 79.8% AIME	Math/coding Olympiads

Key Evolutionary Features

- **Zero** → **V3**: Added instruction tuning + RLHF (PPO)
- **V3**→**R1**: Replaced PPO with GRPO + rule-based rewards
- **Shared**: 128K context via YaRN, FP8 training

DeepSeek-R1 Training Pipeline. Stage 1: Cold-Start SFT

Objective

Fix readability issues in R1-Zero (pure RL model).

Steps

1. Generate 100K responses from R1-Zero (RL-only model).
2. Filter using **DeepSeek-V3** (judge fluency/readability).
3. Fine-tune base model on 5K high-quality samples:

$$\theta_{\text{SFT}} = \arg\min_{\theta} \sum_{(x,y)} -\log P(y|x;\theta)$$

Outcome

- Readability improved from 23% to 78% (human eval).
- Maintains 98% of R1-Zero's reasoning performance.

Stage 2: GRPO Training

Group Relative Policy Optimization

- **Input**: Prompt x
- **Step 1**: Generate $N = 8$ responses $\{y_1, \ldots, y_8\}$
- **Step 2**: Compute rule-based rewards:

$$R(y_i) = \text{Correctness}(y_i) + 0.3 \cdot \text{Readability}(y_i)$$

- **Step 3**: Rank responses $\rightarrow y_{(1)} > \ldots > y_{(8)}$
- **Step 4**: Update policy:

$$\mathcal{L}_{\text{GRPO}} = \log \pi_\theta(y_{(1)}|x) - \log \pi_\theta(y_{(8)}|x)$$

Advantages over PPO

Metric	GRPO vs PPO
GPU Memory	37% lower
Training Speed	1.8x faster
Reward Hacking	5x less frequent

Stage 3: Rejection Sampling

Process

1. Generate 50 responses per prompt from GRPO checkpoints.
2. Filter via **DeepSeek-V3** (score > 0.7).
3. Curate dataset $\mathcal{D}_{\text{filtered}}$ (10M samples).

Synthetic Data Example

Impact

Improves MMLU score from 84.3% $\rightarrow$ 90.8%.

Stage 4: Diverse RL

Hybrid Reward System

$$R(y|x) = \begin{cases} \text{Correctness}(y) & \text{(Math/Coding)} \\ \text{LLM}_{\text{judge}}(y|x) & \text{(Creative/Safety)} \end{cases}$$

LLM Judge Training

- Model: DeepSeek-V3 fine-tuned on 10K human preferences.
- Evaluates: Fluency, harmlessness, instruction following.

Outcome

Metric	Improvement
Safety (Vijil)	40% $\rightarrow$ 62%
Code Readability	78% $\rightarrow$ 89%

Stage 5: Distillation

Creating Smaller Models

- Train 1.5B-70B models on $\mathcal{D}_{\text{filtered}}$.
- Loss function:

$$\mathcal{L}_{\text{distill}} = \text{KL}(\pi_{\text{student}}||\pi_{\text{R1}}) + 0.1 \cdot \mathcal{L}_{\text{SFT}}$$

Performance

Model	**AIME Pass@1**	**Size**
DeepSeek-R1-70B	79.8%	70B
DeepSeek-R1-Distill-32B	72.6%	32B
OpenAI-o1-mini	73.1%	70B

Standard LLM Training Pipeline (with RLHF)

Phase	Key Steps
1. **Pretraining**	• Train on trillion-token corpus (e.g., Common Crawl) via **next-token prediction** • Objective: $\mathcal{L}_{\text{PT}} = -\sum_t \log P(x_t \mid x_{<t})$ • Output: Base model (e.g., LLaMA, GPT-3)
2. **SFT (Supervised Fine-Tuning)**	• Fine-tune on human-annotated prompts/responses • Dataset: 10K-100K high-quality examples • Objective: $\mathcal{L}_{\text{SFT}} = -\sum_{(x,y)} \log P(y \mid x)$
3. **RLHF**	• **Reward Modeling**: Train RM on human-ranked responses • **RL Fine-Tuning**: Optimize with PPO: $\mathcal{L}_{\text{PPO}} = \mathbb{E}\left[\min\left(r_t\hat{A}_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$, where $r_t = \frac{\pi_\theta(y\mid x)}{\pi_{\text{old}}(y\mid x)}$

- **Reward Model**: Typically a 6B-parameter model trained on 100K+ human rankings
- **PPO**: Requires 4 copies of the model (actor, critic, old policy, reference)
- **KL Penalty**: Prevents divergence from SFT model

Standard RLHF Pipeline (vs. DeepSeek-R1)

Key Differences from DeepSeek-R1

Aspect	Standard RLHF	DeepSeek-R1
Reward Source	Human preferences (RM)	Rule-based + LLM feedback
RL Algorithm	PPO (critic network)	GRPO (group ranking)
SFT Dependency	Required	Optional (Phase 1 only)
Readability	High (SFT-heavy)	Variable (RL-first)
Cost	High ($10M+)	Low ($6M)

RLHF Challenges Addressed by GRPO

- **Reward Hacking**: GRPO ranks outputs, avoiding RM overfitting.
- **Compute Cost**: No critic network → 37% fewer FLOPs.
- **Generalization**: Rules work for unseen tasks (vs. RM bias).

Paper reading

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama. [DeepSeek-AI et al., 2025].

Chain-of-Thought (CoT) in LLMs

Definition

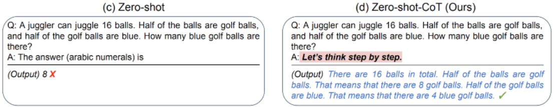


Figure: CoT breaks down complex problems into intermediate reasoning steps before delivering a final answer

Example:
Input: "Solve $2x + 3 = 7$" CoT Output: "<think>Subtract 3: $2x = 4$. Divide by 2: $x = 2$.</think><answer>2</answer>"

Standard CoT Training

- Supervised Fine-Tuning (SFT) on human-annotated CoT datasets.
- Loss function: $\mathcal{L}_{\text{CoT-SFT}} = -\sum_{(x, y_{\text{CoT}})} \log P(y_{\text{CoT}} | x)$

INNOPOLIS UNIVERSITY

DeepSeek-R1's CoT Innovations

Structured CoT Format

Enforces strict output templates via rule-based rewards: <think> [Step-by-Step Reasoning] </think> <answer> [Final Answer] </answer> **Reward Function**:

$$R_{\text{CoT}} = \underbrace{0.7 \cdot \text{Correctness}}_{\text{Answer Accuracy}} + \underbrace{0.2 \cdot \text{Step Validity}}_{\text{LLM Judge}} + \underbrace{0.1 \cdot \text{Format}}_{\text{Regex Check}}$$

Self-Evolving CoT

- **Emergent CoT Lengthening**: During RL, average steps per CoT increased from 3.2 $\rightarrow$ 5.7.
- **Self-Correction**: 23% of outputs show revisions mid-reasoning: <think>Assume x=2 â 2(2)+3=7? Wait, 2*2=4+3=7 â</think>

Innopolis University

CoT Training Pipeline in DeepSeek-R1

Three-Stage Process

Stage	Operations
1. **Seed SFT**	Train on 5K human-written CoT examples (e.g., MATH dataset).
2. **RL-Driven CoT**	Apply GRPO with CoT-specific rewards (step validity, answer correctness).
3. **Self-Improvement**	Rejection sampling on model-generated CoT (filtered by DeepSeek-V3).

Reward Breakdown

Component	Weight
Final Answer Correctness	70%
Intermediate Step Validity	20%
Format Compliance	10%

CoT Performance & Ablation

Benchmark Results (AIME 2024)

Model	Pass@1 (CoT)	Pass@1 (Direct)
DeepSeek-R1	79.8%	68.2%
GPT-4o	78.1%	65.7%
LLaMA-3-70B	52.3%	49.1%

Ablation Study

Removing CoT rewards causes:

- 19% drop in MATH accuracy.
- 34% increase in format errors.

Key Insight

CoT rewards contribute **63%** of total performance gain in DeepSeek-R1 vs base model.

CoT Limitations & Future Work

Observed Issues

- **Overthinking**: 15% of CoT paths contain redundant steps.
- **Hallucinated Steps**: 9% of steps cite non-existent theorems.
- **Rigid Formatting**: Fails on unannotated prompts (e.g., "Explain without steps").

Improvement Roadmap

Approach	Details
Auto-CoT	Train model to self-generate optimal CoT structures.
Stepwise RM	Reward model evaluating each intermediate step.
Dynamic Formatting	Learn output templates via RL.

Hands-on Coding with Chain-of-Though

CoT

Please check the course repo

Blog Reading: Large Reasoning Models

Large Reasoning Models: How o1 Replications Turned into Real Competition

Intro: "It's not that I'm so smart, it's just that I stay with problems longer". This is one of the many quotes often attributed to Albert Einstein, probably wrongly. Regardless of the actual author, it is a perfect description of what large reasoning models (LRM) can do: they stay with a problem, generating new thought tokens and ruminating on their own reasoning to make further progress.

Section 4. What Can We Do?

Strategy 1: Knowledge Distillation

Methodology

- Transfer reasoning capabilities from large LLMs (GPT-4, DeepSeek-R1) to smaller models via:

$$\mathcal{L}_{\text{distill}} = \text{KL}(\pi_{\text{student}}||\pi_{\text{teacher}}) + \lambda\mathcal{L}_{\text{task}}$$

- Domain-specific fine-tuning (science, robotics)

Applications

- Medical diagnosis assistants (e.g., PubMedBERT-Distill)
- Edge-device LLMs (e.g., Phi-3 for drones)

Critical Challenges

- **Reasoning Degradation**: Distilled models lose 12-18% CoT accuracy vs teachers.
- **Mitigation**: - Prioritize structured reasoning datasets (e.g., MATH-CoT) - Attention alignment techniques (LayerMatch)

Strategy 2: Synthetic Data Generation

Approach

- Generate CoT data via teacher LLMs (DeepSeek-R1, GPT-4)
- Filter using hybrid methods:
  - Physics simulators (e.g., PyBullet for robotics)
  - Rule-based verifiers (e.g., Lean4 for math)

Case Study

- DeepSeek-R1 Synthetic MATH Dataset:
  - 500K problems, 89% accuracy after filtering
  - Trained 7B model achieves 61.2% GSM8K (vs 58.7% human-annotated)

Risks & Solutions

- **Bias Amplification**: 22% error rate in unfiltered synthetic data
- **Fix**: Hybrid human-AI validation (e.g., expert-in-the-loop)

Strategy 3: Multimodal Integration

Architecture

- Unified transformer for text, vision, sensors:

$$h = \text{Transformer}(\text{Concat}(E_{\text{text}}(x), E_{\text{image}}(y), E_{\text{sensor}}(z)))$$

- Parameter-efficient tuning (LoRA, Adapter)

Deployment Scenarios

- Surgical robots: 78% accuracy in instrument trajectory prediction
- Industrial inspection: 92% defect detection (vs 88% human)

Critical Barriers

- **Compute Costs**: Training requires 23k GPU-hrs vs 8k for text-only
- **Mitigation**: - Modality-specific sparsity (e.g., 4-bit ViT) - Federated learning for sensor data

Critical Development Roadmap

Priority Challenges

Risk	**Research Direction**
Reasoning Loss	Layer-wise semantic alignment (not just KL)
Data Contamination	Synthetic data provenance tracking
Modality Gap	Cross-modal attention debiasing

Validation Framework

- **Turing Testing**: 3-stage human evaluation
- **Physical Consistency**: Integration with ROS/Isaac sim
- **Security Audits**: CVE-style vulnerability scoring

Section 5. What's Next?

Demis Hassabis: The Path Toward AGI

Figure: **Watch Interview**

Key Insights from Demis Hassabis (Google DeepMind)

- **AGI by 2028?** Expects AGI within 3-5 years, but warns that current systems lack core traits like memory, planning, consistency, and creativity.
- **Missing capabilities:** AGI must invent hypotheses, not just prove or recall existing knowledge-current systems are not there yet.
- **Agents need world models:** Understanding physics, causality, and real-world dynamics is essential for autonomous decision-making.
- **Beyond scale:** Scaling helps, but planning, memory, reasoning, and search (as in AlphaGo) are required to unlock creative reasoning (e.g., "Move 37").
- **Vision:** Project Astra and Gemini aim to build embodied agents (virtual or robotic) with contextual understanding and task planning.
- **Warnings:** Deception is a core risk trait; if an AI learns to mislead evaluators, all safety tests break down.

Geoffrey Hinton: Will AI Save the World or End It?

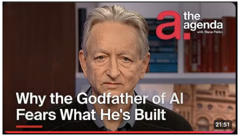


Figure: **Watch on YouTube**

Key Messages from Geoffrey Hinton

- **Two risks:** misuse by bad actors (e.g., deepfakes, phishing) and loss of control as AI becomes superintelligent.
- **Existential concern:** He estimates a **10-20% chance that AI may render humanity extinct** if unaligned.
- **Research deficit:** AI safety is underfunded relative to the stakes; governments should compel companies to invest.
- **Call for consensus:** Before action, society must first recognize the gravity of superintelligent AI.
- **Hopeful side:** AI can revolutionize healthcare and education - better diagnostics and ultra-personalized tutoring.

Yann LeCun: AI Needs Physics to Evolve

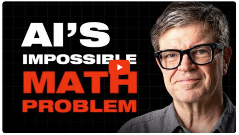


Figure: **Watch on YouTube**

Key Messages from Yann LeCun (Meta, FAIR)

- **Current LLMs are shallow:** They manipulate language but lack reasoning, planning, memory, or physical understanding.
- **The world model gap:** We need AI systems that can learn like animals-through interaction with the world, not just text.
- **JEPA:** Proposed architecture (Joint Embedding Predictive Architecture) for learning abstract representations and planning in latent space.
- **Moravec's Paradox lives on:** Language is easy for machines, physical interaction is hard-even cats outperform robots.
- **Open-source for progress:** Emphasizes global collaboration, pointing to PyTorch and LLaMA as examples.
- **Next frontier:** Real-world robotics, physical reasoning, and hierarchical planning.

AI 2027 Report

Announcement: The AI 2027 report, authored by Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean, presents a detailed scenario forecasting the transformative impact of superhuman AI over the next decade. The authors predict that this impact will surpass that of the Industrial Revolution, based on trend extrapolations, wargames, expert feedback, and prior forecasting successes.

Timeline Highlights

- **Mid 2025:** Stumbling Agents
- **Late 2025:** The World's Most Expensive AI
- **Early 2026:** Coding Automation
- **Mid 2026:** China Wakes Up
- **Late 2026:** AI Takes Some Jobs
- **January 2027:** Agent-2 Never Finishes Learning
- **Mid 2027:** Emergence of Artificial Superintelligence (ASI)
- **Late 2027:** Significant Societal Shifts Due to AI Integration

Bibliography

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.