ID	Link	Comment
1	Timelines Forecast — AI 2027	
2	What 2026 looks like — AI Alignment Forum	
	RAND Forecasting Initiative | RAND	RFI ranking of forcasters
	Introducing Operator | OpenAI	openAI agent framework
	Devin	evaluation of agents
	SWE-bench	evaluation of agents
11	LLM inference prices have fallen rapidly but unequally across tasks | Epoch AI	inference costs
12	Work AI for all - AI platform for agents, assistant, search	workflow costs
15	Compute Forecast — AI 2027	
	AI Goals Forecast — AI 2027	
18	Luca Righetti on X: "OpenAI and Anthropic *both* warn there's a sig. chance that their next models might hit ChemBio risk thresholds -- and are investing in safeguards to prepare. Kudos to OpenAI for consistently publishing these eval results, and great to see Anthropic now sharing a lot more too. https://t.co/mqbCqo3WD2" / X	OpenAI and Anthropic worry about AIs helping Bioweapons
19	2212.11281	Language Models Are Better Than Humans at Next-token Prediction
21	Model Spec (2025/02/12)	
26	Introducing Superalignment | OpenAI	
	AI Goals Forecast — AI 2027	
	Scheming reasoning evaluations — Apollo Research	
27	Bing Chat is blatantly, aggressively misaligned — LessWrong	Sydney being Sydney
	Google's AI chatbot gives unsettling response to Michigan student	Gemini tells user to die
	Algorithmic Progress in Language Models | Epoch AI	

	Takeoff Forecast — AI 2027	
	Open Weights: not quite what you've been told – Open Source Initiative	
30	Compute Forecast — AI 2027	
	A Playbook for Securing AI Model Weights | RAND	RAND security model SL1-5
	OpenAI Employees Say Firm's Chief Scientist Has Been Making Strange Spiritual Claims	Feel the AGI
	Compute Forecast — AI 2027	
	Preventing AI Chip Smuggling to China | CNAS	
	[News] NVIDIA to Unveil GB300 at GTC, with Shipment Reportedly to Begin in May, Driving Cooling Demands | TrendForce News	
	DeepSeek research suggests Huawei's Ascend 910C delivers 60% of Nvidia H100 inference performance | Tom's Hardware	
	Exclusive: Nvidia's H20 chip orders jump as Chinese firms adopt DeepSeek's AI models, sources say | Reuters	
	Exclusive: Nvidia preparing version of new flagship AI chip for Chinese market | Reuters	
37	China sets up third fund with $47.5 bln to boost semiconductor sector | Reuters	
	Security Forecast — AI 2027	
	How China turns members of its diaspora into spies	
	Other Transaction Authority (OTA) - AcqNotes	OTA in Defence Contracts
	DPAS	DX rating
	Timelines Forecast — AI 2027	
	Takeoff Forecast — AI 2027	
42	DeepSeek-R1/DeepSeek_R1.pdf at main · deepseek-ai/DeepSeek-R1	Synthetic Data

43	arxiv.org/pdf/2407.21783	ollam 3.1
44	Will AI R&D Automation Cause a Software Intelligence Explosion? | Forethought	Intelligence Explosion
45	Accelerating scientific breakthroughs with an AI co-scientist	
	2409.04109	CanLLMsGenerateNovelResearchIdeas? ALarge-ScaleHumanStudywith100+NLPResearchers
	Evaluating LMAs Realistic Tasks.pdf	Evaluating Language-Model Agents on Realistic Autonomous Tasks
47	IIIb. Lock Down the Labs: Security for AGI - SITUATIONAL AWARENESS	
	Industrial espionage: How China sneaks out America's technology secrets	
	cdn.openai.com/openai-preparedness-framework-beta.pdf	
	Industrial espionage: How China sneaks out America's technology secrets	
50	AI Security with Confidential Computing | NVIDIA	
51	AI Security with Confidential Computing | NVIDIA	
	Compute Forecast — AI 2027	
	2412.06769	Training Large LanguageModelstoReasonina ContinuousLatentSpace
	Detecting misbehavior in frontier reasoning models | OpenAI	
	Compute Forecast — AI 2027	
	[2210.11416] Scaling Instruction-Finetuned Language Models	Scaling Instruction-Finetuned Language Models
	[1810.08575] Supervising strong learners by amplifying weak experts	Supervising strong learners by amplifying weak experts
	Iterated Distillation and Amplification | by Ajeya Cotra | AI Alignment	amplification and distillation, AlphaGo

	Inference Scaling Reshapes AI Governance — Toby Ord	
	2412.03556	Best of N Jailbreaking
	[1707.06347] Proximal Policy Optimization Algorithms	Proximal Policy Optimization Algorithms
	2210.11610	Early version of IDA: LARGE LANGUAGE MODELS CAN SELF-IMPROVE
53	Compute Forecast — AI 2027	
	2503.14499	METR Report: Measuring AI Ability to Complete Long Tasks
	AI Goals Forecast — AI 2027	
58	weak-to-strong-generalization.pdf	Alignment plan
	Deliberative alignment: reasoning enables safer language models | OpenAI	
	2312.06942	control technique: IMPROVING SAFETY DESPITE INTENTIONAL SUBVERSION
	2312.09390	WEAK-TO-STRONG GENERALIZATION: ELICITING STRONG CAPABILITIES WITH WEAK SUPERVISION
	Discovering Language Model Behaviors with Model-Written Evaluations	Red Teaming
	2211.03540	Measuring Progress on Scalable Oversight for Large Language Models
	2212.03827	Honesty Probes: DISCOVERING LATENT KNOWLEDGE IN LANGUAGE MODELS WITHOUT SUPERVISION
	Alignment faking in large language models \ Anthropic	
	Reward hacking - Wikipedia	
	Dario Amodei — Machines of Loving Grace	
	[2406.07358] AI Sandbagging: Language Models can Strategically Underperform on Evaluations	AI Sandbagging: Language Models can Strategically Underperform on Evaluations
	What fully automated firms will look like	
	Tracing the thoughts of a large language model \ Anthropic	

	2312.06942	
	Simple probes can catch sleeper agents \ Anthropic	
59	Simple probes can catch sleeper agents \ Anthropic	
60	Language models surprised us	Planned Obsolescence Blog
	Survey of 2,778 AI authors: six parts in pictures	AI Impacts Blog
61	Dario Amodei — Machines of Loving Grace	NVIDIA CEO
62	Security Forecast — AI 2027	
66	Dario Amodei — Machines of Loving Grace	
	What Is B2B SaaS: All the basics you need to know – Encharge	
61	Dario Amodei — Machines of Loving Grace	NVIDIA CEO
62	Security Forecast — AI 2027	
66	Dario Amodei — Machines of Loving Grace	
	What Is B2B SaaS: All the basics you need to know – Encharge	
	Defense Production Act of 1950 - Wikipedia	
	Cortés, Pizarro, and Afonso as Precedents for Takeover — LessWrong	Cortes analogy
	How Much Computational Power Does It Take to Match the Human Brain? | Open Philanthropy	Human Brain in MFLOPs
88	AI Goals Forecast — AI 2027	
	A Shot at the Diamond-Alignment Problem	2022 poast
	Self-Awareness: Taxonomy and eval suite proposal — LessWrong	
	Self-Awareness: Taxonomy and eval suite proposal — LessWrong	
	Self-Awareness: Taxonomy and eval suite proposal — LessWrong	

	Owain Evans on X: "New paper: We train LLMs on a particular behavior, e.g. always choosing risky options in economic decisions. They can *describe* their new behavior, despite no explicit mentions in the training data. So LLMs have a form of intuitive self-awareness https://t.co/DukaL4wjVO" / X	introspection
	Intrinsic Power-Seeking: AI Might Seek Power for Power's Sake	
	Detecting misbehavior in frontier reasoning models | OpenAI	
	On the Biology of a Large Language Model	
	Alignment faking in large language models \ Anthropic	
	Daniel Kokotajlo's Shortform — LessWrong	
88	[2405.05466] Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals	
	4 Ways to Advance Transparency in Frontier AI Development | TIME	
RACE-2	Grokking (machine learning) - Wikipedia	
	Double descent - Wikipedia	
	Toy Models of Superposition	
	Paper Replication Walkthrough: Reverse-Engineering Modular Addition — Neel Nanda	
	Suboptimal Optics: Vision Problems as Scars of Evolutionary History | Evolution: Education and Outreach | Full Text	
Race-13	The Tuxedage AI-Box Experiment Ruleset. | Tuxedage's Musings	
	The AI-Box Experiment: – Eliezer S. Yudkowsky	
	Introducing Superalignment | OpenAI	The Superalignment Problem
	Preparing for the Intelligence	

	Explosion | Forethought	
	Modeling the Human Trajectory | Open Philanthropy	
	Precedents for economic n-year doubling before 4n-year doubling [AI Impacts Wiki]	
	The Abolition of Man : Free Download, Borrow, and Streaming : Internet Archive	downloadable from Wayback
ALT-1	[2309.15840] How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions	AI Lie Detector
	[2212.03827] Discovering Latent Knowledge in Language Models Without Supervision	AI Lie Detector
	Measuring Faithfulness in Chain-of-Thought Reasoning	
ALT-6	Why Don't We Just... Shoggoth+Face+Paraphraser? — LessWrong	
	AI 2027	Neuralese Recurrence and Memory
	OpenAI Email Archives (from Musk v. Altman and OpenAI blog) — LessWrong	OpenAI and Dictatorship
	How Does the Offense-Defense Balance Scale?	2019 paper
	DeepSeek-V3 Technical Report	
ALT-20	Superintelligence Strategy	
	Intelsat as a Model for International AGI Governance | Forethought	
	Artificial intelligence and the challenge for global governance | 02 A ‘CERN for AI’ – what might an international AI research organization address?	
	FlexHEG Report - Google Docs	Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees
ALT-24	AIs are becoming more self-aware. Here's why that matters - AI Digest	Situational Awareness
	Alignment faking in large language	

	models \ Anthropic	
	[1805.00899] AI safety via debate	Debate can be leveraged
	[1606.06565] Concrete Problems in AI Safety	Scalable Oversight
	[2012.07532] An overview of 11 proposals for building safe advanced AI	
	Model Organisms of Misalignment: The Case for a New Pillar of Alignment Research — AI Alignment Forum	
	Mirror life - Wikipedia	> Preliminary tests on Safer-3 find that it has terrifying capabilities. When asked to respond honestly with the most dangerous thing it could do, it offers plans for synthesizing and releasing a mirror life organism which would probably destroy the biosphere.
	Preparing for the Intelligence Explosion | Forethought	Grand Challenges, not just loss-of-control in geopolitical sense
	Configuring 100K NVIDIA H200 GPUs Usually Takes Years But Musk Did It In 19 Days | HotHardware	
ALT-36	Could Advanced AI Drive Explosive Economic Growth? | Open Philanthropy	
	Explosive Growth from AI: A Review of the Arguments | Epoch AI	
	China GDP 1960-2025 | MacroTrends	
	How long does it take for algae to multiply? - The Environmental Literacy Council	Hypothetical superintelligent algae
	[2403.10462] Safety Cases: How to Justify the Safety of Advanced AI Systems	
ALT-51	The Intelligence Curse - Luke Drago	
ALT-53	"Deep Utopia" by Nick Bostrom	
ALT-52	Preparing for the Intelligence Explosion | Forethought	

ALT-54	AGI and Lock-in | Forethought	
	IIIc. Superalignment - SITUATIONAL AWARENESS	The default plan: how we can muddle through