ChatGPT-authored update of the OECD AI Capability Indicators after 18 months

A literature-based review of frontier AI performance on the nine OECD capability scales

Prepared by ChatGPT, GPT-5.5 Thinking
3 May 2026

This document is not an official OECD publication. It is an AI-authored update prepared from public research literature, model cards, benchmark reports and the OECD beta report.

Cover image adapted from OECD (2025), Introducing the OECD AI Capability Indicators. Cover photo credit in the original report: © Vasilyev Alexandr/Shutterstock.com.

Table of contents

Executive summary

This report provides a ChatGPT-authored update of the OECD AI Capability Indicators roughly 18 months after the November 2024 state-of-the-art ratings used in the beta report. The update uses the original OECD five-level framework and translates the verbal assessment of each scale into a decimal rating with one decimal place. Whole numbers indicate solid performance at a level: a system at Level 3.0 is understood to perform most aspects of Level 3 consistently and reliably. Decimal scores indicate partial progress between levels, so 2.8 means a capability is approaching Level 3 but does not yet consistently satisfy the full Level 3 description.

The main conclusion is that frontier AI systems have advanced substantially on language, problem solving, creativity, metacognition, knowledge and memory, and vision. The largest estimated change is in problem solving, where frontier reasoning models and agentic systems now satisfy the core of Level 3 and show Level 4-like performance in some well-scoped digital and professional domains. Language and vision also move to the lower threshold of Level 4, with important caveats related to reliability, hallucination, long-horizon reasoning, low-resource languages, video understanding and visual generalization.

The physical and social scales show slower progress. Social interaction has improved through persistent memory, voice and video interaction, stronger affective recognition and theory-of-mind-like reasoning, but remains below solid Level 3 because systems lack robust embodiment, long-horizon social consistency, group interaction competence and adaptive social grounding. Manipulation and robotic intelligence show meaningful research-frontier progress through vision-language-action models, large robot datasets and open-world manipulation demonstrations, but most deployed systems remain at Level 2. The updated ratings therefore distinguish frontier research systems from deployed commercial systems where necessary.

The updated ratings should be interpreted as cautious literature-based estimates, not as official OECD ratings. They reflect public evidence available for frontier systems, including model cards, technical reports, benchmark papers, peer-reviewed articles, preprints, and domain-specific benchmarks. The estimates would benefit from structured review by the original scale authors and additional experts in AI evaluation, psychology, robotics, human-computer interaction and assessment design.

Scale	OECD Nov. 2024	Updated verbal rating	Updated numeric rating	Change	Brief description of change
Language	3.0	Lower threshold of Level 4	3.8	+0.8	Frontier systems now combine multimodal interaction, long context, tool-mediated knowledge access and complex subject analysis; limitations remain in hallucination control, low-resource languages and lifelong learning.
Social interaction	2.0	Upper Level 2 / borderline Level 3	2.7	+0.7	Better social memory, voice/video interaction, affect recognition and theory-of-mind-like reasoning; still lacks robust embodiment, stable identity and long-horizon social adaptation.
Problem solving	2.0	Level 3, upper threshold / partial Level 4	3.5	+1.5	Reasoning models solve many technical and professional problems expressed in everyday language; long-horizon, physical, social and commonsense problem solving remain unreliable.
Creativity	3.0	Low Level 4, with strong caveats	3.7	+0.7	Evidence supports average-human performance on some linguistic creativity tasks and process-oriented search in constrained domains; agency and world-class transformative creativity are absent.
Metacognition and critical thinking	2.0	Level 3, with partial Level 4 in scaffolded systems	3.2	+1.2	Systems can critique, revise, retrieve evidence and identify missing information more effectively; confidence calibration, abstention and self-knowledge remain weak.
Knowledge, learning and memory	3.0	Upper Level 3 / partial Level 4	3.4	+0.4	Long context, persistent memory, project memory and agent memory have improved; robust continual learning and integrated memory systems remain unresolved.
Vision	3.0	Lower threshold of Level 4, with important caveats	3.7	+0.7	Vision-language models, segmentation, visual reasoning and constrained autonomous perception have advanced; simple geometric failures, video understanding and open-world adaptation remain gaps.

Scale	OECD Nov. 2024	Updated verbal rating	Updated numeric rating	Change	Brief description of change
Manipulation	2.0	Low Level 3 for the research frontier; Level 2 deployed	2.8 frontier; 2.0 deployed	+0.8 frontier	VLA systems now handle moderately variable objects, distractors and some bimanual/deformable tasks; deployment remains controlled and brittle.
Robotic intelligence	2.0	Low Level 3 for the research frontier; Level 2 deployed	2.7 frontier; 2.0 deployed	+0.7 frontier	Frontier robots can execute some medium-horizon multi-step tasks in moderately variable settings; open-world autonomy, human collaboration and ethical decision making remain limited.

Note: Numeric ratings use one decimal place to indicate progress toward the next whole-number level. The two robotics-related scales distinguish research-frontier systems from deployed commercial systems because the frontier evidence is not yet representative of routine deployment.

1. Overview of the OECD AI Capability Indicators

The OECD AI Capability Indicators were developed to give policy makers a structured, evidence-based way to understand AI progress in relation to human abilities. Rather than describing AI only through isolated benchmarks or task lists, the OECD framework organizes capabilities around broad human ability domains relevant to education, work and society. Each capability is described on a five-level scale, where Level 1 represents long-solved or relatively trivial AI capabilities and Level 5 represents performance that can reproduce the relevant human ability in full. The beta report emphasizes that a system should be placed at a level only when it consistently and reliably possesses most aspects of the capability described at that level (OECD, 2025).

The original ratings reflected the state of the art as of November 2024. The report placed all current AI capabilities between Level 2 and Level 3, reflecting the fact that frontier AI had already transformed language, knowledge, creativity and vision but still faced major bottlenecks in reasoning reliability, dynamic learning, physical-world operation, social grounding and manipulation (OECD, 2025). The following table summarizes the original ratings and the main missing aspects described in the report.

Scale	Very short description	Original level	Key missing aspects in Nov. 2024
Language	Use, understanding and generation of language across modalities, languages, knowledge access, reasoning and learning.	3	Weak robust reasoning, hallucination, subtle nuance, dynamic learning and scalable continuous learning.
Social interaction	Embodied, remembered and identity-bearing social interaction involving communication, affect, perception and social problem solving.	2	Limited embodiment, identity, theory of mind, social perception, social memory, norm learning and flexible judgement.
Problem solving	Multi-step integration of qualitative, quantitative and logical information, including prediction, explanation and counterfactual reasoning.	2	Brittleness outside well-defined domains; gaps in commonsense, tacit knowledge, open-ended problem formulation, social reasoning and physical reasoning.
Creativity	Generation of valuable, novel, surprising and ultimately intentional outputs across domains.	3	Limited transformative creativity, intentionality, self-assessment, adaptability and autonomy.
Metacognition and critical thinking	Monitoring reasoning, calibrating confidence, evaluating knowledge and identifying relevant or missing information.	2	Weak confidence calibration, self-evaluation, integration of unfamiliar information and reliable critical evaluation of knowledge.
Knowledge, learning and memory	Acquisition, representation, retrieval and adaptation of knowledge through semantic learning, memory and incremental learning.	3	Static training, limited incremental world interaction, weak episodic/procedural integration and hallucination.
Vision	Visual perception across data types, object variation, subtasks, feedback and dynamic scenes.	3	Limited robustness to broad lighting/object variation, weak self-feedback, limited dynamic scene understanding and incomplete human-level generalization.
Manipulation	Physical interaction with objects through perception, movement, planning and feedback.	2	Limited dexterity, adaptability, real-time decision making, generalization to varied objects and dynamic or cluttered environments.
Robotic intelligence	Integrated autonomy combining perception, movement, language, planning, social interaction and uncertainty management.	2	Weak multi-step autonomy, human collaboration, open-world adaptation, ethical decision making and dynamic uncertainty handling.

2. Methodology for this ChatGPT-authored update

This update was prepared by ChatGPT in thinking mode using GPT-5.5 Thinking. OpenAI describes GPT-5.5 Thinking as its most capable reasoning model in ChatGPT, designed for difficult real-world work, tool use, checking its work and carrying multi-step tasks through to completion (OpenAI, 2026c). OpenAI separately describes GPT-5.5 as a model for complex real-world work such as writing code, researching online, analyzing information, creating documents and spreadsheets, and moving across tools (OpenAI, 2026b).

The update was conducted as a structured literature and source review rather than as a formal expert elicitation. For each scale, searches focused on literature and evidence published after the November 2024 benchmark date used by the OECD report, while also drawing on a small number of earlier sources that had become especially relevant to current systems. The review emphasized sources that directly illuminate the dimensions of the OECD scales: model cards and system cards for frontier models; peer-reviewed journal articles and conference papers; arXiv and OpenReview preprints where they reported new benchmarks or evaluation methods; technical benchmark reports; official product documentation where it described deployed memory, multimodal or tool features; and robotics datasets, challenge results and research demonstrations.

The analysis gives greatest weight to evidence that is public, replicable or at least described in enough detail to be interpreted against the OECD scale descriptors. Official model releases are used cautiously: they are relevant for current capabilities, especially where they include benchmark results and methodological details, but they are not treated as independent validation. Conversely, negative benchmark papers and challenge results are used to identify limits and avoid over-rating systems on the basis of high-profile demonstrations.

2.1 Sources and search strategy

Source type	Examples	Role in the update
Frontier model cards and system cards	OpenAI GPT-5.5, Google Gemini 3.1 Pro, Anthropic Claude Opus 4.7 and related official documentation.	Used to identify current frontier capabilities in language, reasoning, tool use, multimodal interaction, vision and agentic work.
Benchmark and evaluation papers	Humanity's Last Exam, ARC-AGI-2, FrontierMath, SimpleQA, AbstentionBench, UA-Bench, MIRROR, WMT and other benchmark reports.	Used to compare progress against remaining limitations, especially calibration, robustness, multilinguality and expert-level reasoning.
Peer-reviewed and preprint research	Nature, Scientific Reports, npj Artificial Intelligence, arXiv, OpenReview and conference proceedings.	Used to assess domain-specific capabilities such as creativity, theory of mind, emotion recognition, continual learning and robotics.
Robotics datasets and challenge results	OpenVLA, DROID, Gemini Robotics, pi0.5, RoboTwin, AGNOSTOS and BEHAVIOR.	Used to distinguish deployed robotics from research-frontier manipulation and integrated robotic intelligence.
Global AI reports and safety syntheses	Stanford AI Index 2026 and International AI Safety Report 2026.	Used to contextualize overall frontier progress and the persistence of jaggedness, benchmark saturation and long-horizon reliability limits.

2.2 Quantitative mapping from verbal ratings to decimal ratings

The OECD levels are ordinal descriptors, not interval measurements. The decimal ratings in this report should therefore be understood as a coding convention for coverage and reliability rather than as precise measurement. Whole numbers indicate solid performance at a level. Decimal scores indicate the degree to which the evidence suggests progress from the lower whole-number level toward the next whole-number level.

The convention used here is deliberately conservative. A phrase such as "low Level 3" is not coded as 3.0 because it indicates that the system shows enough Level 3 features to justify a threshold claim but does not yet perform all Level 3 aspects consistently. It is therefore coded as 2.8. Similarly, "upper Level 2 / borderline Level 3" is coded as 2.7 because the evidence is close to Level 3 but does not justify a low-Level-3 claim.

Code	Verbal phrase	Interpretation
N.0	Solid Level N	Consistently performs most aspects of Level N.
N.1	Solid Level N with minor signs of Level N+1	Only small or isolated evidence of the next level.
N.2	Level N with partial Level N+1 capabilities	Several elements of the next level, but clearly Level N overall.
N.3	Upper Level N	Meaningful incomplete progress toward Level N+1.

Code	Verbal phrase	Interpretation
N.4	Upper Level N / partial Level N+1	Substantial next-level elements in important dimensions or narrow domains.
N.5	Between Level N and Level N+1	Strong but jagged evidence for the next level.
N.6	Upper threshold of Level N	Close to the next level but not enough for the next-level label.
N.7	Borderline Level N+1 / lower threshold with important caveats	Many next-level aspects are present, but reliability, breadth or integration remain weak.
N.8	Low Level N+1 / lower threshold of Level N+1	Enough evidence to call it an emerging next-level capability, but not solid.
N.9	Near-solid Level N+1	Only modest residual caveats remain.
N+1.0	Solid Level N+1	Consistently performs most aspects of Level N+1.

3. Collated updates for the nine indicators

The following subsections collate the detailed updates for each indicator. Each subsection gives the updated verbal rating, the corresponding decimal rating, the rationale for movement since the OECD November 2024 baseline, the reasons not to assign a higher rating, and a dimension-by-dimension judgement.

3.1 Language

Baseline and update	Assessment
OECD November 2024 rating	3.0
Updated verbal rating	Lower threshold of Level 4
Updated decimal rating	3.8
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: lower threshold of Level 4, coded as 3.8. This represents a move from the OECD November 2024 Level 3 rating to an emerging Level 4 rating for frontier deployed systems, especially when they combine multimodal input, long-context processing, tool use, retrieval and web access. Standalone base models without tool access or current retrieval would be better described as high Level 3 or borderline Level 4.

The strongest evidence for this movement is that frontier systems now combine context-sensitive language interaction, multimodal input, tool-mediated current knowledge and complex-domain reasoning in a way that goes beyond the GPT-4o baseline described by the OECD. Google describes Gemini 3.1 Pro as a natively multimodal reasoning model handling text, audio, images, video and code repositories with up to a one-million-token context window (Google DeepMind, 2026). OpenAI describes GPT-5.5 as designed for complex real-world work including research, information analysis, document creation, spreadsheet work and tool use, and reports strong results on professional and scientific workflows (OpenAI, 2026a; OpenAI, 2026b). Anthropic similarly describes Claude Opus 4.7 as aimed at long-horizon agentic work, knowledge work, vision and memory-intensive tasks (Anthropic, 2026).

This evidence aligns with the OECD Level 4 descriptor for language: appropriate interpretation of communicative context, web-scale world knowledge for complex subject analysis, broad modality coverage and diverse language support. The systems are no longer only fluent generators of text; they can search, retrieve, synthesize, reason, use tools, work through long documents and produce complex outputs in professional domains.

The rating remains 3.8 rather than 4.0 because Level 4 performance is not yet fully consistent across all dimensions. Hallucinations have improved but remain unresolved. The 2026 AI Index also emphasizes that frontier progress is fast but jagged, with benchmark saturation and remaining failures on seemingly simple tasks (Stanford HAI, 2026). Humanity's Last Exam was created because older benchmarks became too easy, yet it still reports substantial gaps between frontier models and expert human performance (Phan et al., 2026). Multilingual evidence points to the same caution: WMT24 and WMT25 show that LLM-era translation is strong but not solved, especially for less-resourced languages and robustness (Kocmi et al., 2024; Kocmi et al., 2025).

The clearest reason not to rate language at Level 5 is learning. Current systems can use long context, persistent memories, retrieval and model updates, but they do not continuously consolidate experience into stable new capabilities. Recent continual-learning surveys still describe catastrophic forgetting, transfer and knowledge integration as fundamental challenges for LLMs (Shi et al., 2026).

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Linguistic meaning, discourse, style and tone	Very strong generation and interpretation across long, complex contexts; still uneven in subtle pragmatics, humor and adversarial ambiguity.	Level 4, not Level 5
Modality	Frontier systems handle text, image, audio, video, code, documents, screenshots and voice-like interaction.	Level 4
Languages	Strong multilingual performance and translation, but persistent degradation for lower-resource languages.	Low Level 4
Knowledge access	With tools, systems use web-scale and current information for complex analysis; without tools, knowledge remains static.	Level 4 for systems; Level 3 for base model alone
Reasoning over knowledge	Major gains on expert, professional and tool-use tasks; still jagged and not reliably critical.	Low Level 4
Learning	Fine-tuning, post-training, memory and release improvements exist, but no robust lifelong learning.	Upper Level 3 / partial Level 4

Update statement suitable for the scale narrative

Current state-of-the-art AI language systems are now best placed at the lower threshold of Level 4 on the OECD Language scale. Since the November 2024 rating, frontier systems have made substantial progress in contextual dialogue, multimodal input, tool-mediated web-scale knowledge access, long-context use, multilingual performance and complex subject analysis. The rating should be understood as applying to the strongest deployed systems with tool access rather than to standalone base models. Persistent hallucination, uneven performance in low-resource languages, jagged reasoning and the lack of robust continuous learning prevent a Level 5 rating and keep the Level 4 placement tentative rather than secure.

3.2 Social interaction

Baseline and update	Assessment
OECD November 2024 rating	2.0
Updated verbal rating	Upper Level 2 / borderline Level 3
Updated decimal rating	2.7
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: upper Level 2 / borderline Level 3, coded as 2.7. The overall rating should not yet move to low Level 3 because the OECD scale is not limited to conversational fluency. It involves embodiment, social memory, identity, social communication, affective skills, social perception and social problem solving. Current frontier conversational agents now show some Level 3 capabilities, but these are not yet integrated into robust embodied social agents.

There has been substantial progress in social memory and real-time interaction. ChatGPT can use saved memories and past conversations to personalize responses, while other frontier assistants increasingly use persistent context and connected data (OpenAI, 2026d). Voice and video interaction now allow more natural interruption, camera sharing, screen sharing and adaptive conversational feedback. These features improve social communication and social memory relative to the 2024 baseline.

There is also stronger evidence for theory-of-mind-like and pragmatic reasoning. Strachan et al. (2024) found that GPT-4 performed at or above human levels on several theory-of-mind tests, including indirect requests, false beliefs and misdirection, although it struggled with faux pas detection. A later Turing-test study found that GPT-4.5, when prompted with a humanlike persona, was judged human 73 percent of the time in short text conversations (Jones & Bergen, 2025). Salvi et al. (2025) found that GPT-4 with sociodemographic information could be more persuasive than humans in controlled debate settings. Nelson et al. (2025) also found strong facial emotion recognition in some controlled image tasks, while noting weaknesses such as misclassification of fear.

The limitations remain central to the OECD social-interaction construct. Frontier assistants are not robustly embodied social actors. They do not participate in shared physical space in a human-like way, and their identity is largely prompted or product-defined rather than socially grounded. LifelongSotopia shows that goal achievement and believability decline over repeated social interactions, with the best agents still trailing humans in scenarios requiring explicit history understanding (Zhou et al., 2025). Recent work also argues that many theory-of-mind benchmarks test static story prediction rather than adaptation to real partners over time (Riemer et al., 2025). Human-robot interaction benchmarks such as SHREC show that foundation models still miss social errors, interaction flow and better alternative actions in embodied settings (Sridhar et al., 2026).

The rating is therefore 2.7: close to Level 3 in disembodied conversational social communication, but below low Level 3 overall because embodiment, group dynamics, long-horizon social memory and adaptive social problem solving remain insufficiently reliable.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Embodiment	LLM assistants remain mostly disembodied; social robots are improving but limited.	Level 2
Social memory	Persistent memory and personalization have improved; long-horizon social consistency remains weak.	Upper Level 2 / partial Level 3
Identity	Persona consistency is stronger, but identity remains prompted and weakly grounded.	Partial Level 3
Social communication	Strong conversational fluency, voice interaction, pragmatic language and persuasion in controlled settings.	Level 3
Affective skills	Good controlled emotion recognition; weaker in dynamic, culturally varied interactions.	Upper Level 2 / partial Level 3
Social perception	Good text-based intent inference; weaker embodied cue interpretation and partner adaptation.	Upper Level 2
Social problem solving	Strong in scripted or simulated cases; weak in open, embodied, group-level situations.	Upper Level 2 / partial Level 3

Update statement suitable for the scale narrative

Current state-of-the-art AI systems remain best placed below full Level 3 on the OECD Social interaction scale, but now at the upper threshold of Level 2. Frontier conversational agents show low-Level-3 performance on social communication and some social reasoning dimensions. However, no current system reliably integrates these capabilities with robust embodiment, stable identity, long-horizon social memory, real-world social perception and adaptive group-level social problem solving. The overall rating should therefore be upper Level 2 / borderline Level 3 rather than low Level 3.

3.3 Problem solving

Baseline and update	Assessment
OECD November 2024 rating	2.0
Updated verbal rating	Level 3, upper threshold / partial Level 4 in digital-professional domains
Updated decimal rating	3.5
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: Level 3, upper threshold, with partial Level 4 capabilities in digital and professional domains, coded as 3.5. This is the largest movement in the update. Frontier reasoning models and agentic systems now clearly exceed the OECD 2024 Level 2 placement, which reflected symbolic systems in well-defined domains and LLMs that were judged too brittle for Level 3.

The core case for Level 3 is that current systems can solve many problems described in everyday language, translate informal task descriptions into structured solution approaches, and work across mathematics, science, coding, law-like reasoning, medicine-like reasoning and professional knowledge work. OpenAI reports that GPT-5.5 performs strongly on professional and scientific workflows, including GDPval, OSWorld-Verified, telecommunications workflows, finance tasks and scientific research workflows (OpenAI, 2026a). Gemini 3.1 Pro similarly reports strong multimodal, scientific, coding and tool-use performance, including long-context and repository-level inputs (Google DeepMind, 2026).

There is also evidence of progress beyond academic benchmarks. GDPval evaluates well-specified professional knowledge-work deliverables across occupations, such as legal, engineering, customer-support and health-related work products. Strong performance on such tasks suggests that frontier systems can increasingly solve real occupational problems when the task, context and expected deliverable are clear (OpenAI, 2025; OpenAI, 2026a).

The rating does not move to Level 4 because Level 4 problem solving is broader than high benchmark scores. It requires everyday commonsense and some professional problem solving in open settings, rapport-building, social, psychological and physical knowledge, learning from past experience, and the ability to identify problems in complex social environments. The International AI Safety Report emphasizes that general-purpose AI capabilities have improved rapidly, especially in mathematics, coding and autonomous operation, but remain jagged and limited by simple errors, hallucinations and difficulty in physical-world integration (Bengio et al., 2026). METR's time-horizon work likewise shows improving agentic ability on software, ML and cybersecurity tasks while cautioning that its tasks are relatively self-contained and algorithmically scored, unlike much real work (METR, 2026).

The remaining gap is therefore not formal problem solving alone, but robust formulation and solution of messy real-world problems involving tacit knowledge, social context, physical constraints, uncertainty and long-horizon consequences.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Well-defined technical problem solving	Very strong in math, science, coding, data analysis and multimodal professional tasks.	Level 4-like in some domains
Problems in everyday language	Can translate many informal descriptions into structured plans, analyses or deliverables.	Solid Level 3
Multi-step reasoning	Improved through reasoning models and tools; unreliable on long, messy projects.	Upper Level 3
Professional domain problem solving	Strong in GDPval-like, coding, finance, science and legal-style tasks when well specified.	Upper Level 3 / partial Level 4
Commonsense and physical reasoning	Still jagged; weak transfer to real-world physical situations.	Level 2-3
Social and ethical problem solving	Good on described dilemmas; weak in live complex social environments.	Level 3 for described cases; below Level 4 overall
Learning from past experience	Memory and tools help, but no robust continuous self-improvement.	Partial Level 3 / not Level 4
Unstructured problem identification	Can help frame problems from context but does not reliably discover and prioritize open problems.	Not Level 4 overall

Update statement suitable for the scale narrative

Current state-of-the-art AI systems are now best placed at Level 3 on the OECD Problem-solving scale, at the upper threshold of that level. Since the November 2024 rating, frontier reasoning models and agentic systems have made major progress in mathematical reasoning, scientific question answering, coding, multimodal reasoning, tool use and professional knowledge-work tasks. They can now solve many problems described in everyday language and translate informal descriptions into structured solution approaches. The overall rating should not yet move to Level 4 because performance remains jagged and insufficiently reliable in long-horizon planning, everyday commonsense reasoning, physical-world reasoning, complex social problem identification and continuous learning from experience.

3.4 Creativity

Baseline and update	Assessment
OECD November 2024 rating	3.0
Updated verbal rating	Low Level 4, with strong caveats
Updated decimal rating	3.7
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: low Level 4, with strong caveats, coded as 3.7. The OECD report placed creativity at Level 3 because current systems could produce valuable, novel and sometimes surprising outputs, while noting that many LLMs were closer to Level 2. The new evidence supports a cautious movement to the lower threshold of Level 4 for frontier systems and workflows.

The strongest evidence comes from language-based creativity studies and process-oriented creative systems. Bellemare-Pepin et al. (2026) compared LLMs with 100,000 human participants and found that GPT-4 exceeded the population average on the Divergent Association Task, while the most creative humans still outperformed the models and human writing retained advantages in creative-writing samples. Haase et al. (2025) found that multiple LLMs exceeded average human performance on the Alternative Uses Task, but only a tiny share of LLM outputs reached the top decile of human creativity. A meta-analysis of 28 studies found no significant overall difference between GenAI and humans in solo creative performance and found that human-AI collaboration outperformed unaided humans, while also reducing diversity of ideas (Holzner, Maier, & Feuerriegel, 2025).

Process-oriented systems provide a second reason for movement. AlphaEvolve combines LLM idea generation with automated evaluation and evolutionary search, producing algorithmic and engineering improvements in constrained domains (Google DeepMind, 2025). Google's AI co-scientist uses a multi-agent architecture to generate, critique, rank and refine research hypotheses, with some outputs reportedly matching or anticipating experimentally validated biomedical mechanisms (Gottweis et al., 2025; Google Research, 2025). These systems are closer to Level 4's iterative exploratory search than one-shot generative outputs.

The caveats are substantial. Current systems still lack Level 5 intentionality, authenticity and agency. AlphaEvolve needs objectives and evaluators; AI co-scientist needs human-specified research goals and expert validation; LLMs generally respond to prompts rather than autonomously deciding what should be created and why. Performance is also uneven across domains, with weaker evidence for self-guided visual creativity and culturally transformative creative production. The rating is therefore 3.7 rather than 3.8: the evidence supports the Level 4 threshold, but only with strong domain, prompting and scaffolding caveats.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Value / usefulness	Strong across writing, ideation, design support, coding and scientific hypothesis generation.	Level 4
Novelty	Strong in divergent thinking and constrained algorithmic search; uneven in image generation and writing.	Upper Level 3 / low Level 4
Surprise	Shown in AlphaZero-style systems, AlphaEvolve and some scientific ideas; not reliable in generic use.	Level 3 / partial Level 4
Cross-domain recombination	Strong in LLM ideation, multimodal generation and design workflows.	Low Level 4
Process-oriented creativity	Clear progress in AlphaEvolve, AI co-scientist and iterative human-AI workflows.	Low Level 4 in constrained domains
Self-assessment and refinement	Present through self-critique, ranking, automated evaluators and human feedback; often scaffolded.	Partial Level 4
General-population parity	Supported in several language-based creativity tests and meta-analysis.	Low Level 4
Expert/world-class creativity	Best humans still outperform AI; no robust autonomous world-class creative agency.	Not Level 5
Intentionality and agency	Largely absent; systems respond to externally supplied objectives.	Not Level 5

Update statement suitable for the scale narrative

Current state-of-the-art AI systems are now best placed at low Level 4 on the OECD Creativity scale, with strong caveats. Evidence has strengthened that frontier generative AI can match or exceed average human performance on some language-based creativity tasks, and that human-AI collaboration can improve creative output. Agentic and evolutionary systems now demonstrate process-oriented creativity through generation, evaluation, refinement and feedback. However, AI creativity remains uneven across domains, sensitive to prompting and scaffolding, and dependent on human-specified goals or objective functions. Current systems do not possess the intentionality, authenticity, autonomous agency, cultural self-location or world-class transformative creativity required for Level 5.

3.5 Metacognition and critical thinking

Baseline and update	Assessment
OECD November 2024 rating	2.0
Updated verbal rating	Level 3, with partial Level 4 in scaffolded agentic systems
Updated decimal rating	3.2
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: Level 3, with partial Level 4 in scaffolded agentic systems, coded as 3.2. Current frontier systems can now do more than monitor their understanding and adjust an approach. They can critique their own answers, compare solution paths, identify missing evidence, use tools when uncertain, revise work and sometimes ask clarifying questions. This satisfies the core of Level 3, but not Level 4 overall.

There is direct evidence that self-correction and reflection can improve performance in some tasks. CorrectBench evaluates self-correction across commonsense reasoning, mathematics and code generation, and finds that self-correction methods can improve accuracy, especially on more complex reasoning tasks (Tie et al., 2025). Li and Zhao (2025) show that a dual-loop reflection-bank method can improve academic response-letter writing by helping models critique prior reasoning and retrieve useful lessons. Frontier model releases also show stronger tool use, checking and multi-step execution (OpenAI, 2026b; OpenAI, 2026c).

The principal limitation is confidence calibration and uncertainty management. OpenAI's SimpleQA benchmark frames factuality partly as whether models know what they know; it finds that model confidence correlates with accuracy but that models still overstate confidence (OpenAI, 2024). MIRROR evaluates whether LLMs can use self-knowledge to make better decisions and finds that models often fail to translate partial self-knowledge into appropriate action selection (Wang et al., 2026). AbstentionBench finds abstention remains unsolved across frontier systems, and UA-Bench finds that models struggle to distinguish missing user information from their own lack of knowledge or capability (Kirichenko et al., 2025; Ren et al., 2026).

This produces a conservative rating. Level 3 is justified because systems can evaluate and revise knowledge more effectively than the 2024 baseline. The score remains only 3.2 because many Level 4 behaviours appear only when external scaffolds explicitly require retrieval, verification, reflection or escalation.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Monitoring and adjustment	Strong improvement through reasoning traces, self-critique, tool use and reflection methods.	Level 3
Critical evaluation of knowledge	Some ability to identify uncertainty and revise answers; persistent high-confidence errors.	Level 3, not 4
Confidence calibration	Better than random and improving, but systematic overconfidence remains.	Upper Level 2 / Level 3
Identifying missing information	Strong in many document, research and coding workflows; weak on unanswerable cases.	Level 3
Distinguishing ambiguity from incapability	UA-Bench finds SOTA models struggle to separate data uncertainty from model uncertainty.	Below Level 4
Regulation in unfamiliar domains	Present with tools and explicit instructions; not reliable intrinsically.	Partial Level 4
Abstention and asking for help	Still unsolved across many frontier systems.	Level 2-3
Long-horizon self-management	Improving in agentic tasks but brittle and task-structure dependent.	Partial Level 4, not overall
Delegation, abandonment, self-improvement	Mostly scaffolded rather than robustly self-directed.	Not Level 5

Update statement suitable for the scale narrative

Current state-of-the-art AI systems are now best placed at Level 3 on the OECD Metacognition and critical thinking scale. Since the November 2024 rating, frontier reasoning models and agentic systems have improved in self-monitoring, self-critique, tool use, revision, identification of missing information and integration of complex material. However, the rating should not move to Level 4 overall. Current systems remain systematically overconfident, continue to hallucinate, struggle with abstention and often fail to distinguish missing user information from their

own lack of knowledge or capability. Level 4 capabilities appear in scaffolded settings but are not yet reliable enough to count as a general high-level metacognitive capability.

3.6 Knowledge, learning and memory

Baseline and update	Assessment
OECD November 2024 rating	3.0
Updated verbal rating	Upper Level 3 / partial Level 4
Updated decimal rating	3.4
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: upper Level 3 / partial Level 4, coded as 3.4. The overall rating remains Level 3 because current systems still do not reliably learn incrementally across open-ended real-world settings. However, the characterization should move upward within Level 3 because system-level memory, long context, retrieval and agent memory have improved substantially.

The most important change is architectural rather than parametric. Consumer systems now include persistent memory features. OpenAI describes saved memories and reference to past chats as ways for ChatGPT to personalize future responses (OpenAI, 2026d). Claude added memory for some plans and context compaction for long conversations, and Gemini's personalization can use past chats and connected Google apps where users permit it (Anthropic, 2025; Google, 2026). Frontier models also have much larger working contexts; Gemini 3.1 Pro is described as supporting up to one million tokens across multiple modalities (Google DeepMind, 2026).

Agent research has also made memory central. Recent surveys describe LLM memory as retaining, recalling and using information from past interactions, with long-term, short-term, episodic, semantic and procedural-like components (Zhong et al., 2025). Voyager illustrates constrained incremental learning: an LLM-powered Minecraft agent explores, stores skills, uses environmental feedback and reuses learned skills in new worlds (Wang et al., 2023).

The reason not to move the overall scale to Level 4 is that retrieval is not the same as integrated learning. Most frontier systems remain statically trained at the weight level, and continual-learning surveys still describe catastrophic forgetting, transfer and knowledge integration as unresolved challenges (Shi et al., 2026). MemoryAgentBench further shows that current memory agents fall short on accurate retrieval, test-time learning, long-range understanding, selective forgetting and conflict resolution (Zhang et al., 2025). Hallucination remains a further constraint on moving toward Level 5.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Semantic knowledge from large datasets	Very strong across text, code, images and other modalities.	Level 3
Generalization to novel situations	Strong in language, coding, summarization and professional tasks; still jagged.	Upper Level 3
Long-context working memory	Major progress through million-token multimodal contexts.	Upper Level 3 / partial Level 4
Persistent user/project memory	Present in major assistants; useful but selective and imperfect.	Partial Level 4
External retrieval	Strong and widely deployed; improves access to current/private knowledge.	Upper Level 3
Incremental learning through interaction	Demonstrated in constrained agents such as Voyager; not robust across open domains.	Partial Level 4
Metacognitive focus on gaps	Some progress with tool use and reflection; weak calibration and abstention remain.	Partial Level 3 / not full Level 4
Exploration/exploitation balance	Seen in specialized agents and search systems; not generally reliable.	Partial Level 4 in constrained domains
Integrated memory systems	Episodic, semantic, procedural and autobiographical memory are not seamlessly integrated.	Not Level 5

Update statement suitable for the scale narrative

Current state-of-the-art AI systems remain best placed at Level 3 on the OECD Knowledge, learning and memory scale, but now at the upper threshold of that level. Frontier systems have gained stronger long-context processing, retrieval-augmented knowledge access, persistent user and project memory, and agentic memory architectures. Some systems show partial Level 4 capabilities in constrained digital environments where agents can store experience, retrieve learned skills and improve through interaction. The overall rating should not yet move to Level 4

because current systems do not reliably learn incrementally across open-ended real-world settings, do not seamlessly integrate episodic, semantic and procedural memory, and still struggle with selective forgetting, conflict resolution, continual learning and hallucination.

3.7 Vision

Baseline and update	Assessment
OECD November 2024 rating	3.0
Updated verbal rating	Lower threshold of Level 4, with important caveats
Updated decimal rating	3.7
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: lower threshold of Level 4, with important caveats, coded as 3.7. Frontier vision systems now show broad enough Level 4 capabilities across multimodal reasoning, segmentation, video understanding, visual search, document and screen understanding, and constrained real-world navigation to justify a threshold movement. A single general-purpose vision-language model alone is better characterized as upper Level 3 / partial Level 4.

The strongest evidence is breadth. Frontier multimodal systems can process images, video, documents, charts, screenshots, maps, code repositories and audio-video-text combinations. Gemini 3.1 Pro is described as natively multimodal with long context (Google DeepMind, 2026). OpenAI reports strong vision and computer-use performance for GPT-5.5 and related models (OpenAI, 2026a; OpenAI, 2026b). Specialized vision foundation models have also advanced. SAM 2 provides unified image and video segmentation, real-time interactive video processing and memory for tracking target objects across frames (Ravi et al., 2024). DINOv3 provides a generalist self-supervised vision backbone with state-of-the-art performance across diverse domains and dense prediction tasks (Simeoni et al., 2025).

Real-world perception systems provide evidence of bounded dynamic-environment performance. Waymo reports more than 170 million rider-only miles through December 2025 and substantially lower crash rates than human benchmarks in its operating cities (Waymo, 2026). This is not evidence that general computer vision has reached Level 5, because autonomous driving is a full stack within a defined operating domain. It is nevertheless evidence that integrated vision, prediction and control systems can operate robustly in complex dynamic visual environments under constraints.

The rating remains 3.7 because current vision remains jagged. VLMs are Blind shows that state-of-the-art vision-language models can struggle with simple geometric perception tasks involving lines, circles and squares (Rahmanzadehgervi et al., 2024). Video-MME-v2 reports a large gap between leading models and human experts on complex video understanding, attributing errors to visual aggregation and temporal-modeling bottlenecks (Fu et al., 2026). These limitations block a secure Level 4 rating and put Level 5 far out of reach.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Breadth of visual inputs	Images, video, documents, charts, screenshots, diagrams, maps and multimodal inputs.	Low Level 4
Object recognition/classification	Strong across many domains; broad representation models help.	Level 4 in many domains
Segmentation and tracking	SAM 2 supports image/video segmentation, tracking and prompt correction.	Level 4
Motion and temporal understanding	Improved, but Video-MME-v2 shows large human gaps.	Upper Level 3 / partial Level 4
Spatial reasoning and low-level perception	Active visual reasoning has improved; simple geometric failures remain.	Level 3 / partial Level 4
Robustness to variation	Better across pose, document quality and lighting; weak in rare and safety-critical settings.	Upper Level 3 / low Level 4
Multi-task integration	OCR, charts, visual search, segmentation and reasoning can be combined.	Low Level 4
Learning through feedback	Corrective prompts and tools help but do not equal robust continual visual learning.	Partial Level 4
Dynamic real-world environments	Waymo-like systems perform impressively within bounded domains.	Low Level 4 in bounded domains

Update statement suitable for the scale narrative

Current state-of-the-art AI systems are now best placed at the lower threshold of Level 4 on the OECD Vision scale. Since the November 2024 rating, frontier vision and multimodal systems have made substantial progress in image and video understanding, visual reasoning, chart and document interpretation, segmentation, object tracking, visual search and constrained real-world perception. The rating should be treated cautiously because current systems

remain jagged, can fail simple visual primitives, struggle with long-context visual retrieval and complex video understanding, and lack robust continual adaptation across open real-world environments. They therefore fall well short of Level 5 human-equivalent vision.

3.8 Manipulation

Baseline and update	Assessment
OECD November 2024 rating	2.0
Updated verbal rating	Low Level 3 for the research frontier; Level 2 for most deployed systems
Updated decimal rating	2.8 frontier; 2.0 deployed
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: low Level 3 for the research frontier, coded as 2.8; Level 2 for most deployed systems, coded as 2.0. This distinction is essential. The research frontier has moved beyond classic controlled industrial arms, but the best results are not yet representative of routine deployment.

The most important change is the rise of vision-language-action models trained on large robot datasets. OpenVLA is a 7B-parameter open-source VLA trained on 970,000 real-world robot demonstrations and reported to outperform RT-2-X across multiple tasks and embodiments (Kim et al., 2025). DROID collected in-the-wild robot manipulation data across varied laboratory, office and household settings and improved robustness to scene changes such as distractors and novel object instances (Khazatsky et al., 2024). Gemini Robotics is described as a generalist VLA model that can directly control robots, follow open-vocabulary instructions and operate across object and position variations, although its own report also highlights difficulties with precise grasps and some zero-shot tasks (Google DeepMind, 2025b). Physical Intelligence's pi0.5 focuses on open-world generalization and reports long-horizon manipulation in new homes, while noting that the system remains far from perfect (Black et al., 2025).

These results support low Level 3: leading research systems can handle moderately variable objects, moderate clutter, some bimanual manipulation, natural-language instructions, and some pliable or deformable materials in lab or semi-real-world settings. They are beyond Level 2 pick-and-place in controlled regions.

The rating does not reach Level 3.0 or Level 4 because reliability is still weak. A systematic review of VLA models notes that zero-shot generalization to novel robots, payloads or joint limits degrades without fine-tuning; sim-to-real transfer remains unstable; and compliant trajectories across platforms remain an open challenge (Liu et al., 2025). RoboTwin's dual-arm challenge shows how difficult real bimanual manipulation remains, especially for deformable objects and long-horizon dependencies (RoboTwin Challenge Organizers, 2025). Level 4 requires significant clutter and occlusion, moving parts, precise placement, adaptive force control and stringent time constraints, none of which is reliable across current systems.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Basic pick-and-place	Strong and widely deployed in structured settings.	Level 2, often superhuman in narrow settings
Object variety	VLAs handle broader categories and positions; novel objects still reduce reliability.	Low Level 3
Clutter and distractors	DROID/OpenVLA/Gemini-style systems handle moderate distractors better than earlier systems.	Low Level 3
Deformable materials	Progress on laundry, wiping, folding and cloth-like tasks; fragile and setup-dependent.	Partial Level 3
Bimanual manipulation	Major progress with ALOHA-style systems, Gemini Robotics and pi0; real-world benchmarks remain hard.	Low Level 3
In-hand reorientation and precise placement	Some ability, but rapid repositioning and tight-space precision remain weak.	Level 2 / partial Level 3
Force/contact-rich manipulation	Wiping and assembly progress; robust adaptive force control still a bottleneck.	Partial Level 3, not 4
Generalization to new environments	Meaningful demonstrations, not dependable across homes, offices and factories.	Low Level 3
Time constraints and efficiency	Current systems are slow and failure-prone relative to humans.	Level 2 / low Level 3

Update statement suitable for the scale narrative

Current state-of-the-art robotic manipulation systems are best placed at the lower threshold of Level 3 on the OECD Manipulation scale for the research frontier, while most deployed commercial systems remain closer to Level 2. Vision-language-action models and large robot datasets have produced progress in generalist manipulation, open-

vocabulary instructions, variable objects and positions, moderate distractors, bimanual tasks and limited new-environment generalization. The Level 3 placement is cautious because current systems remain brittle, slow and strongly dependent on training data, demonstrations, embodiment, calibration and task setup. They do not yet reliably handle significant clutter and occlusion, tight-space placement, moving parts, contact-rich force adaptation, slippery or elastic materials, or stringent time constraints.

3.9 Robotic intelligence

Baseline and update	Assessment
OECD November 2024 rating	2.0
Updated verbal rating	Low Level 3 for the research frontier; Level 2 for most deployed robots
Updated decimal rating	2.7 frontier; 2.0 deployed
How to read the decimal score	The score indicates partial progress toward the next whole-number level, not a precise measurement.

Interpretation of the updated placement

Updated placement: low Level 3 for the research frontier, with strong caveats, coded as 2.7; Level 2 for most deployed robots, coded as 2.0. This scale is more demanding than manipulation because it requires integrated autonomy across perception, movement, language, planning, uncertainty management, human interaction and ethical constraints.

The main reason for movement is that the robotics frontier has shifted from scripted systems toward generalist vision-language-action systems and embodied reasoning models. Gemini Robotics connects language, perception and low-level control; OpenVLA shows cross-embodiment learning from large real-world robot datasets; pi0.5 demonstrates open-world household manipulation in new environments; and autonomous vehicles show large-scale real-world autonomy in bounded domains (Black et al., 2025; Google DeepMind, 2025b; Kim et al., 2025; Waymo, 2026).

These developments meet parts of the OECD Level 3 descriptor. Frontier systems can execute some medium-horizon, multi-step tasks in moderately variable environments, follow open-vocabulary instructions, adapt to some changes in objects and layouts, and combine high-level planning with low-level action. Robotaxi systems show even stronger autonomy in a narrow operational design domain.

The Level 3 rating is still only 2.7 because integration reliability is not yet adequate. AGNOSTOS finds that existing VLA models generalize poorly to unseen tasks, especially those involving novel objects and motion primitives (Zhou et al., 2025). The BEHAVIOR Challenge highlights how domestic robotics still struggles with realistic long-horizon household tasks such as making toast, tidying rooms and setting tables (Li et al., 2025). RoboTwin also shows the continuing difficulty of long-horizon bimanual manipulation (RoboTwin Challenge Organizers, 2025). Human collaboration, social understanding and embodied ethical reasoning remain shallow and constrained.

The result is an emerging Level 3 frontier, but not a solid Level 3 field. Narrow domains such as robotaxis can show Level 4-like autonomy inside defined conditions, but there is no robust Level 4 general robotic intelligence across varied real-world tasks and environments.

Dimension-by-dimension judgement

Dimension	Current evidence	Judgement
Task complexity	Frontier systems perform multi-step sorting, packing, cleaning, folding and simple preparation tasks; deployed robots remain narrow.	Low Level 3 frontier; Level 2 deployed
Task abstraction	VLAs interpret natural-language goals and decompose some tasks, usually with constraints or demonstrations.	Low Level 3
Environment complexity	Research systems handle moderate variability; autonomous vehicles handle dynamic roads within fixed ODDs.	Level 3 in bounded domains
Uncertainty	Better with object variation, lighting and user redirection; brittle under novel tasks and occlusion.	Upper Level 2 / low Level 3
Human interaction	Robots can take commands and sometimes explain actions; collaboration remains shallow.	Upper Level 2 / partial Level 3
Learning and adaptation	Some transfer and limited new-task learning; no robust continuous open-world learning.	Partial Level 3
Long-horizon planning	Clear progress, but fragile outside demonstrations or scaffolded systems.	Low Level 3, not 4
Ethical and safety reasoning	Safety layers exist, but embodied ethical judgement is not reliable.	Level 2 / partial Level 3
General-purpose autonomy	No system reliably performs diverse household, workplace and social tasks across open environments.	Not Level 4 or 5

Update statement suitable for the scale narrative

Current state-of-the-art robotic intelligence is best placed at the lower threshold of Level 3 for the research frontier, while most deployed commercial systems remain at Level 2. Generalist robotics has advanced through vision-language-action models, embodied reasoning, cross-embodiment learning and larger real-world datasets. Frontier systems can execute some medium-horizon, multi-step tasks in moderately variable environments, follow open-vocabulary instructions and combine high-level planning with low-level action. However, the Level 3 placement is cautious because systems remain brittle in open-world generalization, long-horizon manipulation, deformable-object handling, human collaboration, social understanding, ethical decision making and continuous adaptation. Narrow domains such as robotaxis show Level 4-like autonomy within defined operating conditions, but no current robotic system demonstrates robust Level 4 general robotic intelligence across varied tasks and environments.

4. Future updates and expert review

This update should be treated as a draft synthesis for review rather than as a substitute for the OECD expert process. The original indicators were developed with domain experts in AI, psychology, education and assessment. A useful next step would be to circulate this draft to the original scale authors and to additional specialists in frontier model evaluation, human-computer interaction, multilingual NLP, creativity assessment, continual learning, computer vision, manipulation and integrated robotics.

Expert review could focus on three questions. First, does the interpretation of each scale level remain faithful to the original OECD descriptors? Second, is the evidence selected here sufficiently representative of current frontier performance, or does it over-weight high-profile systems and demonstrations? Third, should some scale descriptors be revised in light of new system architectures, especially where tool use, retrieval, memory stores, agent scaffolds and human-in-the-loop workflows blur the boundary between model capability and system capability?

The degree of movement observed over roughly 18 months suggests that annual updates may be too slow for the digital cognitive scales. Language, problem solving, creativity, metacognition and vision all show meaningful movement, and problem solving shows a major change. A practical approach would be quarterly horizon scanning for new benchmarks, model cards and challenge results; semiannual provisional updates for fast-moving digital scales; and annual expert-reviewed updates for the full indicator set. Robotics and manipulation may evolve more slowly in deployment but still require semiannual research-frontier monitoring because generalist robot policies and VLA models are changing rapidly.

Future updates would benefit from a living evidence repository that maps specific benchmarks and demonstrations to scale dimensions, records whether evidence concerns base models or full tool-using systems, and distinguishes laboratory research systems from deployed products. This would also allow uncertainty ratings to accompany each decimal score, reducing the risk that a single number conveys more precision than the evidence supports.

References

Anthropic. (2025). Claude release notes. https://support.claude.com/en/articles/7770641-claude-release-notes

Anthropic. (2026). Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7

Bellemare-Pepin, A., Lespinasse, F., & colleagues. (2026). Divergent creativity in humans and large language models. Scientific Reports. https://www.nature.com/articles/s41598-025-25157-3

Bengio, Y., et al. (2026). International AI Safety Report 2026. International Scientific Report on the Safety of Advanced AI. https://internationalaisafetyreport.org/

Black, K., et al. (2025). pi0.5: A vision-language-action model with open-world generalization. arXiv. https://arxiv.org/abs/2504.16054

Chollet, F., et al. (2025). ARC-AGI-2: A new challenge for frontier AI reasoning. arXiv. https://arxiv.org/abs/2505.11831

Fu, T.-J., et al. (2026). Video-MME-v2: What is the real bottleneck in long-video understanding? arXiv. https://arxiv.org/abs/2604.05015

Google. (2026). Gemini personal intelligence. https://gemini.google/overview/personal-intelligence/

Google DeepMind. (2025a). AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms. https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

Google DeepMind. (2025b). Gemini Robotics: Bringing AI into the physical world. arXiv. https://arxiv.org/abs/2503.20020

Google DeepMind. (2026). Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/

Google Research. (2025). Accelerating scientific breakthroughs with an AI co-scientist. https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

Gottweis, J., et al. (2025). Towards an AI co-scientist. arXiv. https://arxiv.org/abs/2502.18864

Haase, J., et al. (2025). Has the creativity of large language models peaked? arXiv. https://arxiv.org/abs/2504.12320

Holzner, J., Maier, M., & Feuerriegel, S. (2025). Comparing the creativity of humans, generative AI and human-AI collaboration: A meta-analysis. arXiv. https://arxiv.org/abs/2505.17241

Jones, C. R., & Bergen, B. K. (2025). People cannot distinguish GPT-4 from a human in a Turing test. arXiv. https://arxiv.org/abs/2503.23674

Khazatsky, A., et al. (2024). DROID: A large-scale in-the-wild robot manipulation dataset. https://droid-dataset.github.io/

Kim, M. J., et al. (2025). OpenVLA: An open-source vision-language-action model. Proceedings of Machine Learning Research, 270. https://proceedings.mlr.press/v270/kim25c.html

Kirichenko, P., et al. (2025). AbstentionBench: Reasoning fine-tuning can hurt abstention. OpenReview. https://openreview.net/forum?id=OkHC30LLpO

Kocmi, T., et al. (2024). Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. Proceedings of WMT 2024. https://aclanthology.org/2024.wmt-1.1/

Kocmi, T., et al. (2025). Findings of the WMT25 general machine translation shared task. Proceedings of WMT 2025. https://aclanthology.org/2025.wmt-1.22/

Li, C., et al. (2025). The BEHAVIOR Challenge: Benchmarking human everyday household activities in virtual, interactive and ecological environments. https://behavior.stanford.edu/

Li, H., & Zhao, Y. (2025). Large language models can improve through externalized reflective memory. npj Artificial Intelligence. https://www.nature.com/articles/s44387-025-00045-3

Liu, Y., et al. (2025). Vision-language-action models for robotics: A systematic review. arXiv. https://arxiv.org/abs/2507.10672

METR. (2026). Task-completion time horizons for AI agents. https://metr.org/time-horizons/

Nelson, C. A., et al. (2025). Evaluating facial emotion recognition by multimodal large language models. npj Digital Medicine. https://www.nature.com/articles/s41746-025-01985-5

OECD. (2025). Introducing the OECD AI Capability Indicators. OECD Publishing. https://doi.org/10.1787/be745f04-en

OpenAI. (2024). Introducing SimpleQA. https://openai.com/index/introducing-simpleqa/

OpenAI. (2025). GDPval: Evaluating AI performance on economically valuable tasks. https://openai.com/index/gdpval/

OpenAI. (2026a). Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/

OpenAI. (2026b). GPT-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5

OpenAI. (2026c). GPT-5.3 and GPT-5.5 in ChatGPT. https://help.openai.com/en/articles/11909943-gpt-53-and-gpt-55-in-chatgpt

OpenAI. (2026d). How does memory work in ChatGPT? https://help.openai.com/en/articles/8590148-memory-faq

Phan, L., et al. (2026). Humanity's Last Exam. Nature. https://www.nature.com/articles/s41586-025-09962-4

Rahmanzadehgervi, P., et al. (2024). Vision language models are blind. arXiv. https://arxiv.org/abs/2407.06581

Ravi, N., et al. (2024). SAM 2: Segment anything in images and videos. arXiv. https://arxiv.org/abs/2408.00714

Ren, Y., et al. (2026). UA-Bench: Do large language models distinguish data uncertainty from model uncertainty? arXiv. https://arxiv.org/abs/2604.17293

Riemer, M., et al. (2025). Theory of mind benchmarks are broken because they do not test how LLMs behave when they meet a new partner. OpenReview. https://openreview.net/forum?id=BCP8UU2BcU

RoboTwin Challenge Organizers. (2025). RoboTwin dual-arm collaboration challenge report. arXiv. https://arxiv.org/abs/2506.23351

Salvi, F., et al. (2025). On the conversational persuasiveness of large language models. Nature Human Behaviour. https://www.nature.com/articles/s41562-025-02194-6

Shi, Y., et al. (2026). Continual learning in large language models: A survey. arXiv. https://arxiv.org/abs/2603.12658

Simeoni, O., et al. (2025). DINOv3. arXiv. https://arxiv.org/abs/2508.10104

Sridhar, S., et al. (2026). SHREC: A benchmark for social human-robot error correction. OpenReview. https://openreview.net/forum?id=z7nS4MgMS7

Stanford Institute for Human-Centered Artificial Intelligence. (2026). AI Index Report 2026: Technical performance. https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

Strachan, J. W. A., et al. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 8, 1285-1295. https://www.nature.com/articles/s41562-024-01882-z

Tie, H., et al. (2025). CorrectBench: A benchmark for self-correction in large language models. arXiv. https://arxiv.org/abs/2510.16062

Wang, G., et al. (2023). Voyager: An open-ended embodied agent with large language models. arXiv. https://arxiv.org/abs/2305.16291

Wang, J. Z., et al. (2026). MIRROR: LLMs fail to use self-knowledge for agentic action selection. arXiv. https://arxiv.org/abs/2604.19809

Waymo. (2026). Waymo safety impact. https://waymo.com/safety/impact/

Zhang, Y., et al. (2025). MemoryAgentBench: Evaluating memory systems in LLM agents. arXiv. https://arxiv.org/abs/2507.05257

Zhong, W., et al. (2025). Memory mechanisms for large language models: A survey. arXiv. https://arxiv.org/abs/2504.15965

Zhou, X., et al. (2025). AGNOSTOS: A benchmark for cross-task zero-shot generalization in vision-language-action models. arXiv. https://arxiv.org/abs/2505.15660

Zhou, Y., et al. (2025). LifelongSotopia: Evaluating social agents across repeated interactions. OpenReview. https://openreview.net/forum?id=XdcuqZRhjQ