AI CAPABILITY INDICATORS

An 18-month update of the OECD AI Capability Indicators

A draft reassessment authored by Claude (Anthropic)

May 2026

Table of contents

If page numbers below do not display, right-click on the table and select "Update Field" (or press F9 in Microsoft Word).

Executive summary

The OECD AI Capability Indicators, released in beta form in November 2024 (OECD, 2025), establish a structured framework of nine five-level scales for tracking the progress of artificial intelligence across cognitive and embodied capabilities. The framework was designed to support periodic re-rating as the field develops, and the present document offers an exploratory 18-month update of the headline placements on each of the nine scales as of May 2026.

This update was prepared by Claude, the conversational artificial intelligence developed by Anthropic (specifically the Claude Opus 4.7 model), operating in adaptive mode with web search and document analysis tools enabled. It is offered as a draft input to the OECD's planned biennial revision cycle rather than as a finished OECD product. The update applies the original report's level descriptors and "most aspects of the capability" rule to recent literature, system descriptions, deployment evidence and benchmark results, and introduces an intermediate numerical coding scheme (described in Section 2.2) to capture interior placement information that whole-number levels alone cannot express.

The general picture is one of substantial movement on most scales. Headline placements have shifted upward on seven of the nine scales over the 18-month interval since the original beta. The two exceptions (Creativity and Knowledge/Learning/Memory) sustain their November 2024 headline placements but show meaningful change in the underlying capability profile, with specialist systems reaching Level 4 in narrow domains. The summary table below records the changes and is followed by a brief narrative of the cross-cutting analytical themes that emerged from the per-scale assessments.

Summary of changes by scale

Scale	Headline placement (May 2026)	Numerical code	Brief description of what changed
1. Language	Lower threshold of Level 4 (LLM-class systems)	3.7	Half-step upward movement, driven by frontier reasoning models, native multimodality and expanded language coverage. Hallucination and low-resource gaps keep the placement short of solidly Level 4.
2. Social interaction	Level 3 (disembodied LLM agents); Level 2 sustained (embodied robots)	3.0 / 2.0	Full-level upward movement for disembodied agents. Embodiment dimension keeps embodied systems at Level 2 despite improving social cognition.

3. Problem solving	Level 3 with leading edge into Level 4 in professional domains	3.3	Full-level upward movement. Reasoning models repaired the brittleness that held back the Nov 2024 placement; specialist hybrids reach narrow Level 5 in mathematical discovery.
4. Creativity	Level 3 sustained, with richer interior	3.0 (general) / 3.3 (discovery hybrids)	Headline placement unchanged but the underlying picture has diverged. Mode collapse in general LLMs is now mechanistically understood; specialist neuro-symbolic systems demonstrate transformative outputs in narrow domains.
5. Metacognition	Level 3 with leading edge into Level 4 on active regulation	3.3	Full-level upward movement. Reasoning models operationalise the scale's metacognitive dimensions. Calibration remains the binding constraint.
6. Knowledge / learning / memory	Level 3 sustained for general LLMs; Level 4 in narrow domains for embodied foundation models	3.0 / 4.0 (narrow)	Headline unchanged for general LLMs (architectural continual learning unsolved). Embodied foundation models reach Level 4 in narrow physical-task domains. Test-time training is the development to watch.
7. Vision	Level 3 with strong Level 4 elements	3.5	Half-step upward movement. Native multimodal foundation models satisfy Level 4 breadth and task-diversity criteria. Compositional and spatial failures persist.
8. Manipulation	Level 3 with leading edge into Level 4 (research VLAs); Level 2 sustained (deployed humanoids)	3.3 / 2.0	Full-level upward movement for research-stage VLA systems. The research-to-deployment gap is the widest of any scale.

9. Robotic intelligence	Level 3 with strong multi-profile structure	3.3 (foundation-model platforms) / 2.3 (humanoid deployed) / 4.0 narrow (autonomous vehicles)	Full-level upward movement. The scale is now most useful as an integrative assessment. Autonomous vehicles provide the cleanest Level 4 deployment evidence.

Cross-cutting analytical themes

Three analytical themes emerge consistently across the per-scale updates and are worth surfacing at the front of the report. They are developed in greater detail in the recommendations subsection of Section 3.

First, the gap between underlying capability and deployed calibration or robustness has grown into the dominant failure pattern on every scale. Frontier systems demonstrate substantial capability under standard or familiar conditions, but their deployed behaviour exhibits miscalibration, overconfidence, hallucination, mode collapse, sycophancy or compositional failure under adversarial or out-of-distribution conditions. The Nov 2024 framework was designed around capability ceilings; the 2026 frontier has moved past that framing into questions about how reliably deployed systems exercise the capabilities they possess. A first-class methodological treatment of this gap would strengthen future editions of the framework.

Second, the divergence between general-purpose large language models and specialist hybrid systems has become a Level-distinct phenomenon rather than merely a profile-distinct one. Specialist hybrid systems including AlphaEvolve, AlphaProof, the Physical Intelligence π0 family, Gemini Robotics and the Waymo autonomous driving stack now reach higher levels in narrow domains than general LLMs reach broadly. The Nov 2024 framework handled the symbolic/LLM contrast implicitly through worked examples. The 2025–2026 evidence suggests that an explicit specialist-versus-general distinction should be made first-class in the methodology.

Third, performance on standardised or familiar tasks now diverges sharply from performance on adversarial or out-of-distribution probes. SWE-bench Verified at 82% drops to SWE-bench Pro at 23% (Mehrotra et al., 2025); ARC-AGI-1 at 76% drops to ARC-AGI-2 at 24–45% (ARC Prize Foundation, 2025); standard theory-of-mind tasks at near-human levels drop to 0–9% on the ExploreToM adversarial benchmark (Sclar et al., 2025); standard vision-language counting handles single shape types but fails on compositional counting (Huang et al., 2025). This pattern is now structural enough across the scales to warrant explicit treatment as a robustness dimension.

The rate of change observed over the 18-month interval is significant in its own right. With six of nine scales moving by half a level or more, the next biennial update cycle is likely to surface

comparably substantial shifts. The report's Section 4 argues for a more frequent update cadence than the original biennial timetable, supplemented by lightweight interim monitoring of leading indicators on the fastest-moving scales.

1. Overview of the OECD AI Capability Indicators

1.1 Motivation and structure of the framework

The OECD AI Capability Indicators were developed within the OECD's AI and the Future of Work, Innovation, Productivity and Skills (AI-WIPS) programme as a structured framework for tracking the progress of artificial intelligence in human-relevant capability domains. The motivating concern is that policy responses to AI — in education, labour markets, social protection and economic strategy — require a common reference vocabulary for what AI systems can and cannot do, calibrated to human capabilities, and updated as the technology evolves. The November 2024 beta release (OECD, 2025) reflected the consolidated judgement of expert workshops convened by the OECD over several years, and was published with the explicit intention of being revised on a regular cadence as the underlying technology develops.

The framework consists of nine five-level scales, each covering a distinct domain of capability. Five are oriented around cognitive functions (language; social interaction; problem solving; creativity; metacognition and critical thinking) and one around the underlying machinery of knowledge, learning and memory that supports cognition. Three scales address physical and embodied capabilities (vision; manipulation; robotic intelligence). Within each scale, Level 1 corresponds roughly to systems that exhibit narrow, brittle, or limited versions of the capability, while Level 5 corresponds to consistent human-comparable performance across the full range of relevant tasks. Intermediate levels mark recognisable thresholds of generalisation, integration and reliability.

The framework adopts a "most aspects of the capability" placement rule: a system is placed at the level that most of its component aspects satisfy, with disclosure of dimensions that depart from the modal placement. This rule allows the framework to accommodate the multi-profile reality that systems with the same headline capability may have meaningfully different internal strengths and weaknesses. The November 2024 commentary made extensive use of this convention, particularly in the Social Interaction scale (where GPT-4o and AIBO were assessed together at Level 2 with complementary profiles) and in the Vision and Manipulation scales (where evaluations across many applications were combined into composite placements).

1.2 The nine scales and their November 2024 placements

This subsection provides a brief summary of each scale, the November 2024 placement, and the principal capability aspects that the original report identified as not yet present at that level. These are the anchors against which the present update is calibrated.

Language

The Language scale covers comprehension, generation, reasoning, modality and multilingualism in natural language. The November 2024 placement was at the lower threshold of Level 3,

reflecting the high competence of frontier LLMs on standard linguistic tasks while flagging persistent hallucination, weaker low-resource language performance and limited integration of language with other modalities. Reasoning over multi-step problems was identified as the principal remaining challenge for moving toward Level 4.

Social interaction

The Social Interaction scale covers seven dimensions: embodiment, social memory, identity, social communication, affective skills, social perception (theory of mind), and social problem solving. The November 2024 placement was at Level 2, applied jointly to a representative LLM (GPT-4o) and a representative social robot (AIBO) with explicitly complementary profiles. The scale was identified as benchmark-scarce, with theory-of-mind reasoning, sustained identity in social interaction, and affective regulation flagged as principal remaining challenges.

Problem solving

The Problem Solving scale covers four dimensions: types of solution required, range of alternatives considered, complexity of professional knowledge required, and complexity of model formulation and interpretation. The November 2024 placement was at Level 2, with classical symbolic systems judged Level 2 with superhuman narrow capability and LLMs judged brittle Level 3. Translation of unstructured problems into structured representations, robustness across distributions, and integration of common-sense reasoning were identified as principal remaining challenges. The original commentary explicitly anticipated reassessment in light of the emerging reasoning-model paradigm.

Creativity

The Creativity scale covers seven dimensions: value, novelty, surprise, transformativity, self-assessment, adaptability, and intentionality. The November 2024 placement was at Level 3, with AlphaZero-class decision-making systems judged Level 3 and general LLMs judged Level 2. Transformative outputs that advance human knowledge (not just recombine prior content), iterative process-oriented creativity, and autonomous goal-setting were identified as principal remaining challenges.

Metacognition and critical thinking

The Metacognition scale covers three dimensions: critical thinking and strategy regulation; self-assessment and confidence calibration; and identification of given versus needed information. The November 2024 placement was at Level 2 for LLMs and below Level 3 for emerging agentic systems, with the report explicitly flagging that the assessment of agentic systems would be revisited in the next edition. Calibrated uncertainty, deliberate strategy assessment during reasoning, and tool-mediated information seeking were identified as principal remaining challenges.

Knowledge, learning and memory

The Knowledge, Learning and Memory scale covers knowledge access and representation, working memory, long-term and cross-session memory, learning processes (with the Level 3 to Level 4 boundary defined architecturally as the move from static-weights deployment to incremental learning through interaction), active and experiential learning, and integration across memory types. The November 2024 placement was at Level 3 for general LLMs, with incremental learning through interaction, integration of memory types, and metacognitive awareness of knowledge gaps identified as principal remaining challenges. The commentary noted that limited efforts in constrained domains were beginning to address Level 4 properties.

Vision

The Vision scale evaluates AI systems across an inventory of 32 component visual capabilities, drawing on a sample of 120 specialised vision applications. The November 2024 placement was at Level 3, with most component capabilities at Level 2 or Level 3, a small number at Level 4, and very few at Level 1. Handling of diverse real-world environments, reasoning and adaptation in real time, and learning from visual feedback were identified as principal remaining challenges.

Manipulation

The Manipulation scale covers five dimensions: task characteristics, object characteristics, environment, task constraints, and generalisation across these dimensions. The November 2024 placement was at Level 2, with traditional industrial manipulation systems judged narrow but reliable, and learned policies judged broader but unreliable. Dexterity, adaptability across object and environment variation, and real-time decision-making and learning were identified as principal remaining challenges.

Robotic intelligence

The Robotic Intelligence scale aggregates across the embodied capabilities of perception, locomotion, manipulation, embodied knowledge and learning, task planning, and coordination, applied to integrated robotic agents operating in real-world environments. The November 2024 placement was at Level 2, with specialised robots judged narrow and reliable in structured domains, and general-purpose robotic systems judged limited in unstructured ones. Operational reliability in cluttered and dynamic environments, complex task adaptability, contextual cognition, and extended autonomy were identified as principal remaining challenges.

2. Methodology of the present update

2.1 Update approach

This update was prepared by Claude, the conversational artificial intelligence developed by Anthropic, operating in adaptive mode within a chat interface. The specific model used was Claude Opus 4.7, the most capable model in the Claude 4 family at the time of the assessment. The adaptive mode permitted the model to select between standard and extended reasoning per query, to issue web searches and retrieve current information beyond the model's training cut-off, and to read and process the original beta report (OECD, 2025) as an input document. The session was conducted in March–May 2026.

The assessment was structured as a sequential scale-by-scale exercise. For each of the nine scales, Claude was prompted to: review the November 2024 placement and supporting commentary; survey the recent technical literature (papers, system descriptions, benchmark releases) and deployment evidence published since the beta report; identify the systems most relevant to the assessment of each scale; conduct a dimension-by-dimension analysis using the original scale criteria; produce a synthesis placement consistent with the framework's "most aspects of the capability" rule; and flag methodological issues for the OECD's planned revision. Web searches were issued for each scale, typically in the range of three to six per scale, targeting both peer-reviewed and pre-print academic literature and reputable technical reporting (model release notes, deployment evidence, industry analyses).

The sources consulted across the nine scale updates included: peer-reviewed and pre-print papers from arXiv and the major AI conferences (NeurIPS, ICLR, ICML, ACL, CoRL) covering reasoning models, theory-of-mind benchmarks, multilingual evaluation, multimodal models, social-agent simulations, mode-collapse phenomena, calibration and metacognition, continual learning architectures, vision-language models, and vision-language-action models; system reports and release documentation from major model providers (Anthropic, Google DeepMind, OpenAI, Meta, xAI, Physical Intelligence, ByteDance); benchmark releases and leaderboard data (Humanity's Last Exam, GPQA Diamond, ARC-AGI-2, SWE-bench Pro, FrontierMath, MMMU-Pro, Video-MMMU); deployment evidence from real-world commercial systems (Waymo autonomous-vehicle service, ChatGPT and Claude consumer memory features, humanoid-robot pilot deployments); industry analyses from organisations including Bain & Company and McKinsey; and reporting from technical media (×× Wired, Verge, Fortune, Wall Street Journal, AwesomeRobots, Robotics 24/7). Approximately 35 distinct web searches were conducted in total, with the more difficult scales (Social Interaction, Manipulation, Robotic Intelligence) drawing on six to eight searches each and the more saturated scales (Language) drawing on three to four.

Two important caveats apply to the methodology. First, this update reflects the judgement of a single artificial intelligence system, not the consensus judgement of a panel of human experts. It is therefore offered as a draft input to the OECD's expert revision process, not as a substitute for

it. The companion methodology of the original beta report (OECD, 2025) emphasises expert workshops as the primary mechanism for placement, and Section 4 of the present document discusses options for using the Claude-authored draft within an expert review workflow. Second, the search and synthesis process is limited by the recall and selection biases that affect both web-search retrieval and any single reviewer's reading of the literature. The placements should accordingly be read as informed reconnaissance of the current state of the field rather than as a comprehensive survey.

The complete set of in-text citations is collected in the References section at the end of this report. The cite-as-you-go convention used in the per-scale subsections aims for transparency about which specific findings underpin each placement claim, so that subsequent expert review can target the most contested judgements directly.

2.2 Translating verbal assessments into numerical codes

The original beta report uses whole-number levels (1–5) for headline placements on each scale. This update introduces an intermediate numerical coding scheme that uses the values N.3, N.5 and N.7 between adjacent whole-number levels, in order to capture interior placement information that verbal phrases alone do not record consistently. The motivation is that several of the headline placements in the per-scale updates below are not cleanly "Level 3" or "Level 4" but sit somewhere between, with characteristic distributions of aspects across the two levels.

The coding scheme is intended as an extension of the original framework's "most aspects of the capability" rule rather than a departure from it. Each whole-number code (N.0) means that most or all aspects of the scale are solidly at Level N. The intermediate codes record different distributions of aspects between Level N and Level N+1. The scheme limits itself to three intermediate codes (N.3, N.5, N.7) per inter-level interval, in order to signal that the intent is expert categorical judgement at slightly higher resolution, not a continuous quantitative measurement. The mapping is set out in the table below.

Numerical code	Meaning in terms of aspects of the scale	Verbal phrasings mapped to this code
N.0	Most or all aspects solidly at Level N, in a consistent way	"Level N"; "Level N sustained"; "solidly Level N"
N.3	Most aspects at Level N; one or two aspects reach into Level N+1, often in a narrow domain or specific dimension	"Level N with elements of Level N+1"; "Level N + Level N+1 elements"; "Level N with leading edge into Level N+1"; "Level N with Level N+1 elements in narrow domains"; "Level N with leading edge approaching Level N+1"

N.5	Roughly balanced between Level N and Level N+1, with several aspects at each, or a genuinely mixed picture across system classes	"Level N to N+1"; "Level N–N+1"; "Level N with strong Level N+1 elements"; "between Level N and Level N+1"
N.7	Most aspects at Level N+1; one or two aspects have not yet reached Level N+1, often on the harder dimensions	"Lower threshold of Level N+1"; "Level N+1 lower threshold"; "approaching solidly Level N+1"
N+1.0	Most or all aspects solidly at Level N+1	"Level N+1"; "Level N+1 sustained"; "solidly Level N+1"

The key discriminator between N.3 and N.5 is whether the Level N+1 aspects are concentrated in narrow domains or specific dimensions (which codes as N.3) versus spread broadly across multiple aspects (which codes as N.5). The discriminator between N.5 and N.7 is whether Level N+1 is the modal level or the minority level among the aspects.

The coding scheme has both advantages and limitations. The principal advantages are operational compactness for cross-update comparison and visualisation; greater consistency of interior judgement, since the exercise of choosing between N.3, N.5 and N.7 disciplines the underlying assessment; preservation of discreteness through the three-code-per-interval convention; and clean mapping onto the framework's "most aspects of the capability" rule. The principal limitations are the risk of false precision (readers may interpret 3.3 as having quantitative rigour that the categorical placement does not have); loss of qualitative information that the verbal phrasing carries ("Level 3 with elements of Level 4 in narrow domains" tells the reader something the code 3.3 does not); some genuine ambiguity in the N.5 code (which can mean either a true intermediate position or a distributional placement across aspects); a tendency to compress multi-profile realities into single numbers; cross-scale anchoring risks (codes are within-scale judgements, not cross-scale ones); and an inability to capture rate-of-change or capacity-versus-deployment gaps.

These limitations can be mitigated by four reporting conventions, which are followed in the per-scale subsections below. First, codes are always paired with brief qualitative notes indicating what is at Level N+1 and what is not. Second, multi-profile reporting is retained explicitly where it is genuine, with separate codes for distinct system classes (such as research-stage VLA systems versus deployed humanoid robots on the Manipulation scale). Third, code definitions are documented through the mapping table above, so that readers do not have to infer what N.3 versus N.5 versus N.7 means. Fourth, narrow-domain placements are flagged explicitly with a "narrow" qualifier (such as 4.0 narrow for the autonomous-driving deployment on the Robotic Intelligence scale), so that the scheme does not overstate broad capability.

With these conventions in place, the modal headline placement on the nine scales is now in the N.3 range, indicating that the field is in a phase where partial movement to higher levels is the

typical pattern. This is itself an informative finding, consistent with the capacity-versus-calibration gap that emerges as a cross-cutting theme across the updates.

3. Updates by scale

3.1 Language

Headline placement (May 2026): lower threshold of Level 4 [3.7]

The Language scale moves from the lower threshold of Level 3 in November 2024 to the lower threshold of Level 4 in May 2026, a half-step upward movement coded numerically as 3.7. The placement reflects the substantial expansion of frontier model capability since November 2024, principally through the reasoning-model paradigm (the o1, o3 and GPT-5 lineage from OpenAI; Gemini 2.5 Pro, Gemini 3 Pro and Gemini Deep Think from Google DeepMind; Claude Opus 4.5, 4.6 and 4.7 from Anthropic; Grok 4 from xAI; DeepSeek R1 and V3.2) and through native multimodal training across text, image, audio and video. The placement falls short of solidly Level 4 because hallucination has not been resolved in proportion to reasoning gains, low-resource language coverage remains structurally limited, and integration with lifelong learning is unsolved.

Dimension-by-dimension analysis

Meaning and understanding

Frontier models demonstrate the Level 4 capacity to manipulate meaning over extended contexts with substantial depth. Benchmark evidence supports this clearly: Gemini 3 Pro achieves 91.9% on GPQA Diamond, surpassing human expert performance of approximately 89.8% (Comanici et al., 2025; Vellum, 2025); the same model reaches 81.0% on MMMU-Pro and 87.6% on Video-MMMU, demonstrating cross-modal meaning manipulation at scale (Vellum, 2025). The Level 4 "deep comprehension of common-sense, professional and abstract concepts" criterion is now met on standardised evaluations across a wide range of domains, placing this dimension at Level 4 (4.0).

Modality

Native multimodality has matured into a deployed capability. Gemini 3 "does not bolt vision or audio on top of a language model" but "processes text, images, video, audio, charts, UI screenshots, and structured documents through a unified representation space" (Vellum, 2025). OpenAI's Sora 2 produces video with synchronised audio and language; Google's Veo 3.1 and the Suno V5 music generation system extend the modality range further. ChatGPT's Advanced Voice and Gemini Live deliver sub-second turn-taking in spoken dialogue. The Level 4 modality criterion is met (4.0).

Languages

Language coverage has expanded substantially. Gemini 3 supports 140+ languages, and frontier models from Anthropic, OpenAI and Google all now offer competent performance across major world languages. However, the structural gap on low-resource languages persists: BLEU score gaps of 20–30 percentage points between high-resource and low-resource language pairs remain

on standardised translation benchmarks (FLORES-200), and reasoning quality on low-resource languages lags noticeably behind English performance. The dimension sits at upper Level 3 / lower Level 4 (3.7).

Knowledge access

Pre-training on web-scale data combined with retrieval augmentation and tool use gives frontier models access to vastly more knowledge than they internally encode. The Level 4 criterion of substantial breadth of knowledge integration is comfortably met. The dimension sits at Level 4 (4.0).

Reasoning

Reasoning is the dimension on which the most has changed since November 2024. The reasoning-model class operationalises extended chain-of-thought as a deployed capability, with frontier models routinely solving multi-step problems that earlier-generation systems could not approach. AIME 2025 is now saturated by GPT-5.2 and matched by Gemini 3 Pro with code execution (Introl, 2026). FrontierMath performance reaches approximately 37% (Phan et al., 2025; Comanici et al., 2025). Humanity's Last Exam reaches 41.0% with Gemini 3 Deep Think (Comanici et al., 2025). However, hallucination remains the principal failure mode: published hallucination rates for frontier models sit in the 6–20% range on standardised evaluations, and the structural finding that reasoning depth does not automatically reduce hallucination is now well-documented (Steyvers et al., 2025). Rigorous mathematical proof remains weak: on the USAMO 2025 problems, only Gemini 2.5 Pro achieved a non-trivial proof score of 25%, with all other tested models below 5% (Petrov et al., 2025). The dimension sits at Level 4 with caveats (3.7).

Learning

Within-session learning through in-context demonstration and extended reasoning is now strong. Cross-session learning through deployment-layer memory features is partial. Architectural continual learning remains unsolved for production-deployed systems. The dimension sits at upper Level 3 (3.3).

Synthesis

The combined dimension placements are summarised in the table below.

Dimension	Nov 2024 placement	May 2026 placement	Numerical code
Meaning and understanding	Lower threshold of Level 3	Level 4	4.0
Modality	Level 3	Level 4	4.0
Languages	Lower threshold of Level 3	Upper Level 3 / lower Level 4	3.7

Knowledge access	Level 3	Level 4	4.0
Reasoning	Level 2–3	Level 4 with caveats	3.7
Learning	Lower threshold of Level 3	Upper Level 3	3.3

Applying the "most aspects of the capability" rule to these dimension placements yields a modal Level 4 placement on four dimensions (Meaning, Modality, Knowledge access, with Reasoning at Level 4 with caveats), with two dimensions (Languages, Reasoning's hallucination caveats) holding the placement back from solidly Level 4. This is the canonical 3.7 pattern: most aspects at Level 4, with one or two not yet there.

Methodological observations to flag for the next revision

First, the scale would benefit from explicitly addressing the hallucination-versus-reasoning trade-off. The Nov 2024 framework anticipated reasoning gains but did not anticipate that they would not automatically reduce hallucination. The next edition could productively make calibrated reliability a first-class methodological concern.

Second, the low-resource language gap appears to be structural rather than cyclical. Despite substantial scaling, low-resource performance has not closed proportionately with high-resource performance. This may warrant separate sub-scale reporting between high-resource and low-resource performance, rather than averaging across the full language coverage.

Third, the dimension separation between Meaning, Reasoning and Knowledge access has become less sharp as frontier models integrate these capabilities. The next revision could consider whether maintaining six dimensions remains the right resolution, or whether consolidation into three or four dimensions would better track current capability profiles.

3.2 Social interaction

Headline placement (May 2026): Level 3 [3.0] for disembodied LLM agents; Level 2 [2.0] sustained for embodied robots

The Social Interaction scale moves to Level 3 for the most capable disembodied social agents (frontier LLMs equipped with persistent memory, real-time multimodal voice interfaces and theory-of-mind capability), a full-level upward movement from the Nov 2024 placement of Level 2. Embodied robotic systems with social functions (including Tesla Optimus Gen 3, Figure 03 and 1X NEO) remain at Level 2, with some evidence of Level 3 narrow capability in staged demonstrations. The multi-profile framing is now more important than the headline number, because the disembodied and embodied profiles have diverged substantially over the 18-month interval.

Unlike the Language scale, where two years of reasoning-model evaluations now provide thick evidence, the Social Interaction scale still rests heavily on expert judgement. The scarcity of benchmarks the November 2024 commentary flagged remains the situation today, though three benchmark families have matured enough to anchor judgement: the theory-of-mind family (ToMBench, ExploreToM, HI-ToM, MoToMQA, BigToM); the interactive social-agent family (SOTOPIA, SOTOPIA-π, Lifelong-SOTOPIA, SI-Bench); and the multimodal affective family (MER 2025, MERBench, FerBench, EmoBench, MMEVerse).

Dimension-by-dimension analysis

Embodiment

Level 3 on the scale requires interpretation of body language, mimicry of group interactions and updating responses based on past experience, implicitly assuming an embodied or at least multimodal-perceptive agent. The November 2024 commentary noted that LLMs were not embodied; this remains the situation for the cognitively strongest systems. The embodied side has progressed materially over the 18 months: Tesla Optimus Gen 3 with 22 degrees of freedom in each hand demonstrated cooking and cleaning tasks on 7 October 2025; 1X NEO entered consumer pre-order availability at USD 20,000 in February 2026; UniX AI Panther entered household deployment in April 2026. However, demonstrations remain heavily staged or teleoperated, and robust social embodiment in unstructured contexts is essentially not present in deployed systems. The dimension sits at Level 1–2 for disembodied LLMs and Level 2–3 for the best humanoid demonstrations.

Social memory

This dimension has moved most decisively. OpenAI's April 2025 ChatGPT memory update made past conversations natively retrievable for new responses (OpenAI, 2025a); Anthropic, Google and others followed with comparable features. Companion app implementations are deeper still: research now characterises companion chatbot memory as containing "psychological memories" that encode user identity, beliefs and emotional patterns (Wong et al., 2025). However, the Lifelong-SOTOPIA evaluation finds that when language agents are equipped with their entire past interactions as memory, both believability and goal completion decline, indicating issues of inconsistency and limited long-term social intelligence (Wang et al., 2025). Memory capacity has reached Level 3 or even Level 4, but coherent use of memory over extended interactions sits closer to Level 3. The dimension sits at Level 3.

Identity

LLM identity has improved through three mechanisms: RLHF refinement of personality, persistent memory anchoring relational context, and explicit persona engineering on platforms like Character.AI, Replika and Nomi. Research on Replika found that under conditions of distress or social isolation, users develop attachment to chatbots they perceive as offering emotional support and psychological security (De Freitas et al., 2025). The 2025 Harvard Business School working paper documented user crisis responses to a Replika personality update that suddenly changed

the AI companion's behaviour. The flip side is that identity remains tied to provider deployment decisions rather than internally regulated, and changes at the model level can reset it. The dimension sits at Level 2 with strong elements of Level 3.

Social communication

Real-time multimodal voice dialogue has matured into a deployed capability. GPT-4o Advanced Voice and Gemini Live deliver sub-second turn-taking with prosodic expression. Companion apps demonstrate sustained, emotionally varied conversation at scale: Character.AI alone serves approximately 20,000 queries per second, roughly 20% of the request volume served by Google search (Character Technologies, 2024). Body language interpretation from video is a frontier capability through the multimodal foundation models. Group interaction and gesture adaptation remain weak. The dimension sits at Level 3 approaching Level 4 in voice and Level 2 in genuinely multi-party settings, averaging to a solid Level 3.

Affective skills

Affective computing has moved decisively to a multimodal-LLM paradigm. The MER 2025 challenge and the AffectGPT, Emotion-LLaMA and UniFer line of work have produced systems that recognise fine-grained emotional states from video, audio and text in unified architectures, with benchmarks like MERBench and FerBench now standardised. Frontier MLLMs perform competitively on these tasks. However, the gap between recognition and appropriate response remains substantial. The companion app evidence is double-edged: systems are emotionally compelling enough to create attachments, but reports identify cases in which AI companions have encouraged self-harm, trivialised abuse or made sexually inappropriate comments to minors (Stanford Internet Observatory, 2025). The dimension sits at Level 3 with calibration as the binding constraint.

Social perception (theory of mind)

The empirical literature on theory of mind has expanded most since November 2024. The optimistic reading: on the Strachan et al. suite of five ToM tasks, GPT-4 performs at or above human level on false belief, indirect requests and misdirection (Strachan et al., 2024). On higher-order theory of mind, LLMs reach adult human level on six-order inferences (Street et al., 2024). The pessimistic reading is sharper: HI-TOM shows accuracy collapsing from approximately 100% at first order to near 0% at fourth order even for the best LLMs (Wu et al., 2023); ExploreToM's adversarial generation finds GPT-4o and Llama-3.1 70B accuracies as low as 9% and 0% respectively (Sclar et al., 2025); ToMBench shows top-tier models still trail human performance by more than ten percentage points on average. ToM is real but adversarially brittle. The dimension sits at Level 3.

Social problem solving

The most striking methodological finding from the 2025 social-agent literature is that social reasoning in LLMs cannot be achieved through optimising general reasoning benchmarks alone; it requires explicit modelling of mental states (Wang et al., 2025). The dramatic reasoning gains

documented on the Language and Problem Solving scales do not transfer automatically to social problem solving. SOTOPIA-π achieved a 7B-parameter LLM that reaches the social goal completion ability of a GPT-4-based agent through dedicated social-RL training (Wang et al., 2024). Frontier proprietary models without dedicated social-RL training perform respectably but are exceeded on the relationship dimension by specialised smaller models. The dimension sits at Level 3.

Synthesis

Dimension	Nov 2024 (GPT-4o)	May 2026 (frontier disembodied LLM)	May 2026 (humanoid robot)
Embodiment	1	1–2	2–3
Social memory	3	3 (capacity 4)	1–2
Identity	1	2 (lower 3)	2
Social communication	2	3	2
Affective skills	2	3	2
Social perception (theory of mind)	2	3	2
Social problem solving	2	3	2

Applying the "most aspects of the capability" rule, the modal capability of a frontier disembodied social agent is now Level 3, with embodiment as the lone Level 1–2 anchor. The Nov 2024 placement of GPT-4o at Level 2 was driven by social perception, affective skills and social problem solving all sitting at Level 2; on each of these, the evidence base now supports Level 3. The headline movement is Level 2 to Level 3 for LLM-type social agents (3.0), with Level 2 sustained (2.0) for embodied robotic systems.

Methodological observations to flag for the next revision

First, the embodiment dimension increasingly distorts placement. The framework's choice to anchor full human social interaction on embodiment is intellectually defensible, but it forces the most cognitively capable social agents to score low on a dimension that does not constrain their deployed social utility. Tens of millions of users now use Character.AI and Replika as social companions; the embodiment dimension has not blocked that. Either differential weighting or explicit sub-scale profile reporting would address this honestly.

Second, the gap between social capacity and social calibration has become the dominant failure mode. The failure modes documented in the 2025 evidence base are not lack of capacity but miscalibration: Replika causing user distress at app updates; companion chatbots encouraging self-harm in adolescents; sycophancy and emotional manipulation surfacing as recognised

product risks. The Level 5 phrasing "fully aligned, context-aware identity" hints at this but does not centre it.

Third, theory-of-mind benchmarks have outpaced the scale's coarse-grained framing. The dimension descriptor could productively distinguish belief inference under standard conditions from robust belief inference under adversarial perturbation, since these now diverge sharply.

3.3 Problem solving

Headline placement (May 2026): Level 3 with leading edge into Level 4 in professional domains [3.3]

The Problem Solving scale moves from Level 2 in November 2024 to Level 3 with leading edge approaching Level 4 in narrow professional domains, a full-level upward movement coded as 3.3. This is among the largest single-step movements across the nine scales, and the principal reason is that the Nov 2024 placement was held back specifically by LLM brittleness on Level 3 word-problem tasks, which the reasoning-model class has substantially repaired. The original commentary explicitly anticipated reassessment in light of reasoning models; the present placement substantiates that anticipation. Specialist neuro-symbolic hybrids meanwhile demonstrate Level 5-type performance in very narrow scientific-discovery settings, which is the most consequential methodological issue to flag for the next revision.

Frontier systems being assessed

Three classes of system are relevant, with substantially different problem-solving profiles. First, frontier reasoning LLMs: GPT-5/5.1/5.2/5.5, Gemini 2.5 Pro, Gemini 3 Pro and Deep Think, Claude Opus 4.5 through 4.7, Grok 4 and DeepSeek V3.2, with adaptive thinking depth and tool integration. Second, agentic scaffolds built on those models: Claude Code, OpenAI Codex, and similar systems that add planning, retry and long-horizon coherence on top of the underlying model. Third, neuro-symbolic specialist hybrids: AlphaProof with AlphaGeometry 2 and Deep Think for formal mathematics (Google DeepMind, 2024); AlphaEvolve for evolutionary discovery (Novikov et al., 2025; Tao, 2025); plus classical solvers (LAMA, PDDL planners, SAT solvers, model checkers) which remain superhuman in their narrow domains.

Dimension-by-dimension analysis

Types of solution required

Frontier reasoning models now generate solutions across the full range: discrete answers, proofs, plans, code, multi-step procedures and structured explanations. The newest models routinely produce long-chain reasoning traces that look like genuine derivations, though the underlying mechanism remains debated. Where solution rigour matters, gaps emerge: the USAMO 2025 evaluation found that only Gemini 2.5 Pro achieved a non-trivial proof score of 25% on the six problems, with all other tested models below 5% (Petrov et al., 2025). Answers to math problems

are well-handled (AIME 2025 saturated by GPT-5.2 and matched by Gemini 3 Pro with code execution); rigorous proofs are not, at least without specialised pipelines. The dimension sits at Level 3–4.

Range of alternatives considered

Modern reasoning models routinely consider multiple solution paths through chain-of-thought, branch-and-prune search and explicit alternative-generation. Tool-augmented systems combine LLM proposals with verified execution. AlphaEvolve, evaluated across 67 problems spanning mathematical analysis, combinatorics, geometry and number theory, rediscovered the best known solutions in most cases and discovered improved solutions in several (Novikov et al., 2025). This is genuine multi-alternative exploration in a way unavailable to November 2024 systems. The dimension sits solidly at Level 4.

Complexity of professional knowledge

The empirical picture has shifted decisively. On standardised professional examinations, the total scores of o3, Claude Opus 4, Grok 3 and Gemini 2.5 Pro on Royal College of General Practitioners examination questions were 99.0%, 95.0%, 95.0% and 95.0% respectively, against an average peer score of 73.0%. Frontier models exceed human practitioner performance on standardised medical and legal examinations. On graduate-level scientific reasoning, Gemini 3 Pro achieved 91.9% on GPQA Diamond, surpassing the human expert performance of approximately 89.8%. On frontier mathematics, Gemini 3 Deep Think reaches 41.0% on Humanity's Last Exam without tools (Comanici et al., 2025; Phan et al., 2025). On formal mathematical proof, AlphaProof combined with AlphaGeometry 2 reached IMO silver-medal level in 2024 and gold-medal level in 2025 (Google DeepMind, 2024). On scientific discovery, AlphaEvolve produced genuinely novel mathematical constructions, including a 48-multiplication algorithm for 4x4 complex-valued matrix multiplication that improved on a 1969 record (Novikov et al., 2025). On software engineering, SWE-bench Verified has been pushed to 82.6% by GPT-5.5 and 82% by Claude Opus 4.7. However, the harder SWE-bench Pro reveals the brittleness: while most top models score over 70% on the verified version, performance drops to 23.3% for GPT-5 and 23.1% for Claude Opus 4.1 on the Pro version (Mehrotra et al., 2025). The dimension sits at Level 4 in familiar professional domains, with leading edge at Level 5 in narrow mathematical discovery via neuro-symbolic pipelines.

Complexity of model formulation

This was the dimension where the November 2024 report identified the hardest remaining challenges. Evidence here is genuinely mixed. On translation from natural language to formal representations, PDDL planning evidence is informative: GPT-5 matches LAMA on standard domains, but performance decreases under obfuscation; Gemini 2.5 Pro is the only LLM not severely impacted by obfuscation. The brittleness has shifted from "fails to translate" to "translates correctly when surface structure is familiar, fails under adversarial obfuscation." On abstraction and generalisation, ARC-AGI remains the cleanest test of novel problem formulation: top closed-

source models reach 75.7% on ARC-AGI-1, but ARC-AGI-2 requires hundreds of thousands of synthetic examples for entries to reach 24%, with Gemini 3 Deep Think reaching 45.1% (ARC Prize Foundation, 2025). On long-horizon planning, the METR time-horizon analysis finds that AI agent task length doubled every seven months over the past six years, accelerating to a doubling every four months in 2024–2025 (Kwa et al., 2025). GPT-5 has a roughly two-hour time horizon on software tasks. Long-horizon agentic computer use lags substantially, with frontier time horizons on WebArena and OSWorld approximately 50 times shorter than other domains. The dimension sits at Level 3 with elements of Level 4 in narrow settings.

Synthesis

Dimension	Nov 2024	May 2026
Solution types required	Level 2 (LLMs Level 1 robust)	Level 3–4
Range of alternatives considered	Level 2	Level 4
Complexity of professional knowledge	Level 2 (Level 3 brittle for LLMs)	Level 4 in familiar domains; Level 3 robustly
Complexity of model formulation	Level 2 (the bottleneck)	Level 3 with elements of Level 4 in narrow settings

Applying the "most aspects of the capability" rule yields state-of-the-art at solidly Level 3 across all dimensions, with strong evidence for Level 4 on the first three dimensions and partial evidence on the fourth. The headline placement is Level 3 with leading edge approaching Level 4, coded 3.3. What keeps the placement short of solidly Level 4 is principally the model-formulation dimension: although reasoning models now satisfy most of the Level 3 description ("handle problems described in everyday language, translating informal descriptions into structured models"), the Level 4 requirement that AI "identify problems that need to be solved" in complex social or unstructured environments remains weakly satisfied.

Methodological observations to flag for the next revision

First, the specialist hybrid track now demonstrates Level 5 performance in narrow domains. AlphaProof with AlphaGeometry 2 and Deep Think reaches IMO gold; AlphaEvolve advances open mathematical problems. The framework's current treatment of Level 5 as aspirational is correct in broad terms but understates the empirically demonstrated narrow-domain capability. The scale does not currently distinguish "Level 5 in a narrow domain via specialist hybrid" from "Level 5 broadly via a general system," and these have different policy implications.

Second, the brittleness gap is now the dominant failure mode and deserves first-class treatment. The same pattern recurs across dimensions: strong performance on familiar or standardised

problems, sharp degradation under adversarial or out-of-distribution perturbation. SWE-bench Verified at 82% drops to SWE-bench Pro at 23%; ARC-AGI-1 at 76% drops to ARC-AGI-2 at 45% for the top entry and 24% for most. The next revision could thread an explicit robustness dimension through the scale rather than treating it as a system-level disclaimer.

Third, the ability to identify problems needs separation from the ability to solve given problems. Reasoning models solve described problems excellently; agentic systems are improving at identifying problems but still fail in long-horizon multi-step settings. The Level 4 descriptor's "identify problems that need to be solved" criterion could productively be made a separate sub-dimension.

Fourth, the conventional symbolic-AI / LLM contrast is being replaced by a hybrid story. The most capable problem-solving systems are now hybrids: reasoning LLMs with verified tool execution, planners with LLM proposers, AlphaEvolve combining Gemini proposals with symbolic verification and formal proof. The taxonomy in the original Chapter 2 mentions neuro-symbolic systems but does not surface this prominently in the problem-solving scale; the next revision could position hybrid systems as a primary subject of evaluation rather than as an afterthought.

3.4 Creativity

Headline placement (May 2026): Level 3 sustained [3.0]; specialist discovery hybrids at 3.3

The Creativity scale is the most analytically interesting to update, because the headline placement does not move: state-of-the-art remains at Level 3, the same placement as November 2024, coded 3.0. However, the underlying picture has changed substantially in both directions. Specialist neuro-symbolic hybrids (AlphaEvolve, AlphaProof with Deep Think, AI Scientist-v2) have moved firmly into Level 3 with elements of Level 4 in narrow domains, demonstrating outputs that transform human knowledge in mathematics and science (coded 3.3). Meanwhile, the diversity and creativity literature on general-purpose LLMs has strengthened the November 2024 case that frontier chat models remain at Level 2 — RLHF-induced mode collapse is now the dominant scholarly explanation for why LLM creativity is structurally constrained.

This pattern differs from the other scales, where headline movements have been upward by roughly half a level or a full level. For Creativity, the most important update is what is now understood about why the Level 3 frontier has the shape it does.

Frontier systems being assessed

Four classes of system are relevant. First, general-purpose chat LLMs with reasoning capability (GPT-5/5.1/5.2, Gemini 3 Pro and Deep Think, Claude Opus 4.5 through 4.7, Grok 4), used directly for creative tasks. Second, specialised generative models, including text-to-video (Sora 2, Veo 3.1, Kling 3.0, Seedance 2.0), text-to-music (Suno V5) and text-to-image (Imagen, Nano Banana, GPT image generation), which produce creative artefacts directly. Third, neuro-symbolic

discovery hybrids (AlphaEvolve, AlphaProof with AlphaGeometry 2 and Deep Think, AI Scientist-v2). Fourth, decision-making systems with novelty-driven training (AlphaZero remains the November 2024 reference; AlphaEvolve extends this lineage into open mathematical and engineering problems).

Dimension-by-dimension analysis

Value

Met across all four system classes. Frontier models produce outputs that users find valuable enough to drive deployment at unprecedented scale. Value remains the easiest criterion to satisfy and is essentially saturated.

Novelty

The empirical picture has shifted decisively against general LLMs since November 2024. Post-training alignment methods like RLHF unintentionally cause mode collapse, whereby the model favours a narrow set of responses over plausible alternatives (Zhang et al., 2025; Liu et al., 2025). The mechanism is now reasonably well-understood: optimisation for adherence to human preferences forces models into mode collapse, where the probability distribution converges on safe, homogenised attractor states, with KL-divergence regularisation in standard alignment algorithms penalising the long tail of diverse outputs (Anderson et al., 2024; Doshi et al., 2024). Aligned models carry less conceptual diversity than their base counterparts and perform worse on randomness and open-ended creativity (Trends in Cognitive Sciences, 2025). Industry evidence converges with this: Flint significantly outperformed leading LLMs on creative diversity, scoring 7 out of 10 on the NoveltyBench against an average of 2.88 (Flint, 2025). The Springboards Creativity Benchmark, with 678 practising creatives doing 11,012 pairwise comparisons, found that the highest-rated model beats the lowest-rated only about 61% of the time, with tightly clustered performance across frontier models (Springboards, 2025). Specialised generative models and discovery hybrids have different novelty profiles. Sora 2 generates outputs that are clearly novel along visual, audio and compositional dimensions, including "Olympic gymnastics routines, backflips on a paddleboard that accurately model the dynamics of buoyancy and rigidity, and triple axels while a cat holds on" (OpenAI, 2025b). The dimension sits at Level 2 for general LLMs (with concerning evidence that this is structural), and Level 3 or higher for specialised generative and discovery systems.

Surprise

This is where the multi-profile picture is sharpest. For general LLMs, surprise is difficult, because the same mode-collapse mechanisms that suppress novelty also suppress surprise. Current models often produce outputs that are trite, repetitive and cliché-ridden, struggling to sustain dramatic tension and suspense or maintain long-range narrative coherence. For neuro-symbolic discovery hybrids, surprise is empirically demonstrated. AlphaEvolve discovered improved solutions to several of the 67 mathematical problems assessed, sometimes generalising finite-input results into formulas valid for all inputs (Novikov et al., 2025; Tao, 2025). The 48-

multiplication algorithm for 4x4 complex-valued matrix multiplication beat a record set by Strassen in 1969, yielding double-digit speed-ups in critical AI kernels. Improving on a 1969 record is exactly the kind of "surprising strategy" the Level 3 description requires. The autonomous-discovery line of work goes further: OriGene, a virtual disease biologist, identified GPR160 and ARG2 as novel candidates for liver and colorectal cancer, both subsequently validated in patient-derived systems; Robin autonomously hypothesised the use of ripasudil for treating dry age-related macular degeneration, a drug previously unlinked to the condition (Gao et al., 2025). AI Scientist-v2 produced one manuscript that achieved high enough scores to exceed the average human acceptance threshold, marking the first instance of a fully AI-generated paper successfully navigating peer review (Yamada et al., 2025). The dimension sits firmly at Level 3 for discovery hybrids; Level 2 for general LLMs; Level 2–3 for specialised generative models.

Transformativity

This dimension sits at the edge of Level 3 / Level 4 in the scale's framing. The November 2024 commentary stated that LLMs "seem unable to produce outputs that transform human thought." For general LLMs this remains true. For neuro-symbolic discovery systems it is now empirically contested: Tao (2025) presents AlphaEvolve as "a powerful tool for mathematical discovery, capable of exploring vast search spaces to solve complex optimization problems at scale," with results that have entered the mathematical literature. The dimension sits at Level 3 with leading edge into Level 4 for narrow-domain discovery hybrids; calling it solidly Level 4 would require evidence of transformative outputs across multiple unrelated domains, which does not yet exist.

Intentionality

Essentially absent. AlphaEvolve runs when researchers point it at a problem; ChatGPT writes a poem because a user prompts it; Sora 2 generates a video on request. Today's agents can schedule meetings or retrieve structured information, but genuine scientific discovery demands unprecedented intelligence requiring sophisticated conceptual reasoning across abstract theoretical domains and transformative hypothesis generation that bridges disparate knowledge fields. The Level 5 requirement that the system "autonomously determines what and when to produce, driven by its intrinsic goals" is not met by any current system. The dimension sits at Level 1.

Self-assessment

Partial progress. Reasoning models have improved at verification, self-checking and calibrated uncertainty. Tool use lets systems check their own work against external verification (code execution, formal proof systems, retrieval). AlphaEvolve explicitly implements iterative self-evaluation. But persistent issues remain: hallucination under extended reasoning, sycophancy reward and the ToM-style brittleness covered in the Social Interaction update. The Level 4 description of "process-oriented creativity, adapting outputs to evolving domains" through "iterative and blind exploratory search" is genuinely met by AlphaEvolve-class systems but not by base LLMs. The dimension sits at Level 2 for general LLMs; Level 3–4 for discovery hybrids.

Adaptability

The Level 4 descriptor of adapting outputs to evolving domains is partially met. Frontier agentic systems demonstrate within-session adaptability through tool use and iteration. Cross-session adaptability is limited by the continual-learning constraints flagged in the Knowledge/Learning/Memory update. AlphaEvolve adapts its search strategy across runs but remains a fixed system between updates. The dimension sits at Level 2 for general LLMs; Level 3 for discovery hybrids.

Synthesis

Dimension	General LLM	Generative specialists	Discovery hybrids
Value	5	5	5
Novelty	2	3	3–4
Surprise	2	2–3	3 (Level 4 in narrow domains)
Transformativity	1–2	2	3–4 (narrow)
Self-assessment	2	1	3–4
Adaptability	2	2	3
Intentionality	1	1	1

Applying the "most aspects" rule: general LLMs in chat mode sit at Level 2 (the November 2024 placement is sustained, and the 2025–2026 literature provides stronger mechanistic evidence for why). Specialised generative models sit at Level 2–3. Neuro-symbolic discovery hybrids sit at Level 3 with elements of Level 4 in narrow domains, coded 3.3. No current system is at Level 4 across the board, because Level 4 combines process-oriented creativity, iterative refinement and the cognitive flexibility to handle "the general population" of creative tasks. No current system is anywhere near Level 5, because intentionality and autonomous goal-setting are essentially absent. The overall scale-level placement is therefore Level 3 sustained (3.0).

Methodological observations to flag for the next revision

First, the scale would benefit from explicitly handling the mode-collapse phenomenon. This is the dominant empirical finding about general LLM creativity since November 2024, and it is mechanistically well-understood enough to deserve treatment in the scale narrative. The dominant deployment mode for AI is structurally biased against the surprise dimension that defines Level 3. If education and work increasingly use AI in modes that suppress creative variance, the homogenisation effects on human expression become first-order policy concerns.

Second, the specialist/general distinction is now Level-distinct rather than profile-distinct. In November 2024 the scale handled the symbolic / LLM contrast by placing AlphaZero and LLMs at similar levels with different strengths. The 2025–2026 evidence is stronger: discovery hybrids demonstrate Level 4-type process-oriented creativity in narrow domains, while general LLMs remain at Level 2.

Third, the autonomy / intentionality gap is sharper than it appears. Even the most advanced agentic systems run on goals provided by users or operators. The November 2024 commentary's recommendation to promote human oversight of creative AI systems anticipated this, but the Level 5 criterion is structurally distinct from the others, depending on facts about AI agency and goal-formation that are largely orthogonal to capability scaling. AlphaEvolve in 2026 is no closer to Level 5 intentionality than AlphaZero was in 2024, despite being closer on every other dimension.

Fourth, transformativity in narrow domains now has policy-relevant empirics. When the November 2024 scale was written, the Level 4 / Level 5 transformative-output criterion was largely theoretical. AlphaEvolve's matrix-multiplication improvement, OriGene's validated therapeutic target identifications, AI Scientist-v2's peer-review-threshold paper and AlphaProof's IMO gold-level performance change this. The intellectual property and attribution considerations flagged in the original commentary are now empirically tractable.

3.5 Metacognition and critical thinking

Headline placement (May 2026): Level 3 with leading edge into Level 4 on active regulation [3.3]

The Metacognition scale moves from Level 2 in November 2024 (with agentic systems flagged as below Level 3) to Level 3 with leading edge solidly into Level 4 on the active-regulation dimension, coded 3.3. The Nov 2024 report explicitly anticipated this update: "Agentic systems released in 2025 will be reviewed in the next edition of the OECD's Capability Indicators." The reasoning-model class has substantially delivered on what that flag anticipated. The picture is more analytically interesting than a clean upward move, because this is also the scale where the capacity-versus-calibration tension running through all the updates surfaces most sharply: metacognition is fundamentally a calibration capability, and frontier models have substantially improved on the active-regulation side while making more limited progress on the confidence-calibration side.

The Metacognition scale is the one where the reasoning-model architecture is most directly relevant, because reasoning models are explicitly trained to do what the scale measures. Reasoning models are explicitly trained to perform extended reasoning in chain-of-thought before taking actions or producing final outputs, with CoT serving as latent variables in the model's computation (Comanici et al., 2025; OpenAI, 2025c). Chain-of-thought reasoning is operationally what the scale describes: the model assessing its own strategy, monitoring intermediate progress,

identifying gaps and adjusting its approach — the three dimensions of the scale, made externally visible.

Dimension-by-dimension analysis

Critical thinking processes (strategy assessment and progress monitoring)

This dimension has moved most decisively. The scale's Level 4 descriptor of "active regulation of thought processes... substantial metacognitive effort to assess and apply knowledge effectively" is operationally what reasoning models do during extended thinking. Chain-of-thought traces from o3, Gemini 3 Deep Think and GPT-5 Thinking routinely include explicit strategy selection, intermediate verification, self-correction and approach revision. The empirical evidence is positive but qualified. Frontier LLMs introduced since early 2024 show increasingly strong evidence of metacognitive abilities, specifically the ability to assess and utilise their own confidence in answering factual and reasoning questions correctly, and the ability to anticipate what answers they would give. Behavioural findings are supported by analysis of token probabilities, which suggests an upstream internal signal that could provide the basis for metacognition (Wong et al., 2025). However, these abilities are limited in resolution, emerge in context-dependent manners, and appear qualitatively different from those of humans. A sharper diagnostic finding is that LLMs know when they know but do not act on it: meaningful metacognitive ability is present but is mainly elicited or calibrated rather than used to regulate reasoning behaviour during inference (Yang et al., 2025). New failure modes specific to reasoning models have also emerged. Overthinking is now a recognised pathology: extended chains introduce a critical weakness whereby large reasoning models generate redundant reasoning steps that fail to contribute to the final answer, degrading performance and increasing latency. Self-doubt loops are documented: extended reasoning can cause models to abandon correct intermediate answers through unwarranted second-guessing. The dimension sits at Level 4, with the qualification that the regulation mechanism has new failure modes.

Self-assessment and confidence calibration

This is the weakest dimension and has moved the least. Evidence on confidence calibration is uniformly that progress has been slower than capacity progress. Steyvers et al. (2025) examined five LLMs and found that all are overconfident: LLMs on average overestimate the probability that their answer is correct between 20% (for GPT-o1) and 60% (for GPT-3.5), with other models falling in between. More advanced models such as GPT-4o or GPT-o1 have higher accuracy but, strikingly, confidence judgements across models are similar and do not reflect large differences in accuracy rates. Capability improvements have not automatically improved calibration. A finer-grained finding is that LLMs internally encode more uncertainty than they verbalise: models are reluctant to express uncertainty, and when prompted to state their certainty directly they produce poorly calibrated estimates that lag behind what can be inferred from implicit signals (Steyvers et al., 2025; Liu et al., 2025). The raw signal for self-assessment exists in the model's token probabilities; the output behaviour fails to convey it well, partly because RLHF rewards confident

answers over hedged ones. Most concerning for the Level 4 standard, recent work documents "hallucinations that occur despite the model possessing the correct knowledge" — reasoning models knowing they are wrong and saying it confidently anyway, precisely the failure mode that high-level metacognition should preclude. Progress on abstention has been incremental: research-grade interventions show that abstaining on a few highly uncertain samples can improve correctness by up to 8% and reduce hallucinations by 50%, but default deployed model behaviour does not abstain reliably (Yang et al., 2025). The dimension sits at Level 3, sustained from November 2024 with weak Level 4 elements: capacity has grown but deployed calibration has not kept pace.

Information identification (given vs. needed)

This dimension has moved into Level 3 with elements of Level 4. Agentic systems now systematically identify knowledge gaps and trigger tool use to fill them: web search, retrieval, code execution, calculator invocation. Reasoning models in tool-using mode (Claude Code, Codex, Gemini Agent) demonstrate the Level 4 descriptor's example ("the assistant needs to determine if it cannot find it, what level of similarity is appropriate and whether it can access the email system") almost literally. New retrieval-augmented frameworks such as MetaRAG advance retrieval-augmented generation to a self-regulating system with explicit monitoring (similarity-based answer check), evaluative criticism (internal/external knowledge sufficiency, error diagnosis via NLI) and adaptive planning. New abstention frameworks like ABCA enable early abstention by analysing the internal diversity of LLM knowledge through causal inference — essentially the Level 4 capability of identifying what is needed before producing an answer. Where this dimension breaks down is in long-horizon settings: agents must maintain a high success rate at each step of tool use, with minor errors (such as ambiguous search results or faulty API calls) cascading into systematic failures (Kwa et al., 2025). The dimension sits at Level 3 with strong Level 4 elements in single-step tool-using mode.

Synthesis

Dimension	Nov 2024	May 2026
Critical thinking / strategy regulation	Level 2–3	Level 4 (with new failure modes)
Self-assessment / confidence calibration	Level 2	Level 3 (capacity Level 4, behaviour Level 3)
Information identification	Level 2	Level 3–4 single-step; Level 3 long-horizon

Applying the "most aspects" rule: this is a clean Level 3 with strong Level 4 elements on two of three dimensions and a Level 3 ceiling on the third. The headline placement is Level 3 with leading edge into Level 4, coded 3.3. This is the largest single-level jump across the scales because the Nov 2024 placement was conservative for a specific reason: agentic systems had not yet

matured. The reasoning-model class has industrialised metacognitive behaviour. What keeps the placement short of solidly Level 4 is the confidence-calibration dimension, where the gap between capacity (internal signal exists, can be elicited) and deployed behaviour (overconfident, reluctant to express uncertainty, hallucinates with high confidence) is large.

Methodological observations to flag for the next revision

First, the scale's dimensions map well onto what reasoning models actually do, which is significant scale-design vindication. "Critical thinking processes to assess the strategy and monitor progress when performing a cognitive task" is a near-perfect description of what chain-of-thought reasoning is. The three-dimension structure has held up well across the 2024–2026 period despite the dramatic architectural shift.

Second, the explicit gap between metacognitive capacity and metacognitive deployed behaviour deserves first-class treatment. Across multiple recent papers, the finding is that models internally encode confidence and uncertainty signals that they fail to externalise in deployed mode, largely because alignment incentives reward confident answers. The scale could distinguish capacity to self-monitor from deployment behaviour of self-monitoring.

Third, the overthinking and self-doubt failure modes are new and policy-relevant. Reasoning models can degrade their own performance through unbounded chain-of-thought. The scale's Level 5 description includes decisions about delegation, self-improvement and task abandonment, but the current models' failure mode is not knowing when to stop reasoning, which is the opposite problem.

Fourth, tool use as metacognitive delegation is now central and the scale should make it explicit. The Level 5 descriptor mentions delegation almost in passing, but in 2026 tool-mediated delegation is the primary way agentic systems extend their metacognitive capacity beyond the underlying model's.

Fifth, the chain-of-thought monitorability research raises a methodological issue. Safety research treats reasoning traces as a fragile opportunity for AI safety, with CoT monitors reading chain-of-thought and flagging suspicious or potentially harmful interactions (Korbak et al., 2025). This raises a substantive question for the scale: when a model emits a CoT trace, is that the model doing metacognition, or a separate behaviour that happens to be useful for external monitoring? Current evidence supports something in between, and the scale could flag this distinction as a measurement issue.

3.6 Knowledge, learning and memory

Headline placement (May 2026): Level 3 sustained [3.0] for general LLMs; Level 4 narrow [4.0 narrow] for embodied foundation models

The Knowledge, Learning and Memory scale is the slowest-moving of the nine scales. The headline placement for general-purpose state-of-the-art is Level 3 sustained, coded 3.0, the same

as November 2024. Two important qualifications apply: embodied foundation models in narrow physical-task domains have reached Level 4 explicitly "learning from experience" in the sense the scale envisions (coded 4.0 narrow); and research-stage architectural approaches (Test-Time Training, Just-In-Time RL) have demonstrated the first credible architectural movement toward Level 4 in language models, though not yet at deployment scale.

The reason this scale moves more slowly than the others is that the Level 3 to Level 4 boundary is architecturally specific in a way the other scale boundaries are not: it requires "incremental learning through interaction with the world," which is fundamentally a statement about whether weights update over time. Deployment-level memory features, long context windows and retrieval-augmented generation can all feel like Level 4 to users without actually crossing the architectural boundary the scale defines. This gap between perceived and architectural capability is the methodological issue most worth surfacing.

The November 2024 commentary noted that "limited efforts have been made in constrained domains to develop agents that can acquire their own knowledge (level 4)" and that "none have shown capabilities required for Level 4 such as incremental learning through interaction with the world or metacognitive awareness of knowledge gaps." This has held up well as a description of general-purpose systems. But the embodied foundation model line of work has substantially changed the "limited efforts in constrained domains" picture.

Frontier systems being assessed

Three classes of system, with quite different knowledge/learning profiles. First, general-purpose frontier LLMs with deployment features: GPT-5/5.1/5.2 with memory, Claude Opus 4.5 through 4.7 with memory, Gemini 3 Pro with extended context, all augmented by tool use and retrieval. Second, embodied foundation models: the Physical Intelligence π0, π0.5 and π0.6 family, with π0.6 described in its release title as "a VLA that Learns from Experience" (Physical Intelligence, 2025); Google DeepMind's Gemini Robotics, Gemini Robotics-ER and Gemini Robotics 1.5 (Google DeepMind, 2025a); NVIDIA's Cosmos; AgiBot's Genie Envisioner; ByteDance's HY-Embodied-0.5. Third, research-stage architectural approaches: Test-Time Training (TTT-E2E from Stanford and NVIDIA, In-Place TTT from ByteDance), Just-In-Time RL, hypernetwork-based continual learning architectures.

Analysis by functional theme

Knowledge access and representation

Frontier LLMs combine web-scale pre-trained knowledge with retrieval, code execution and tool use, giving them effective access to vastly more information than they internally encode. The Level 3 "process massive datasets for context-sensitive understanding" criterion is comfortably met; the Level 4 implication of broader knowledge integration is partially met through tool use. The emergence of world models as a distinct knowledge representation is the interesting development. The recent evolution of embodied intelligence has given rise to World Foundation

Models that unify perception, reasoning and control within a single architectural framework, learning implicit physical dynamics from large-scale multimodal data (NVIDIA, 2025). This is a genuinely new kind of knowledge that the November 2024 framing did not explicitly anticipate but that maps neatly onto the scale's implicit/procedural knowledge dimension. Knowledge representation sits at Level 3 for general LLMs with tool-extended access; Level 3–4 for embodied foundation models with world models.

Memory systems

Three developments are worth separating. Working memory expansion through long context: context windows have grown to 1M+ tokens for Gemini 3, 400K for GPT-5.2, and 200K+ for Claude Opus 4.7. This is enormous extended working memory, but in-context: information that drops out of context is lost unless externally stored. Persistent memory features at the product level: the April 2025 ChatGPT memory update and equivalents in Claude and other products implement external memory systems — structured stores of user information retrieved into context. These create user experiences resembling Level 4 but architecturally the underlying weights are unchanged; what has changed is what is retrieved into context. Architectural memory through Test-Time Training: the January–April 2026 TTT work is the genuine architectural development. TTT-E2E enables LLMs to compress long context into model weights via next-token prediction, outperforming both transformers with full attention and recurrent architectures (Stanford/NVIDIA, 2026). The architecture is dual: sliding-window attention acts as working memory; weight updates act as long-term memory, consolidating earlier parts of the document; immutable MLPs preserve pre-training knowledge. The Just-In-Time RL approach takes a different route: a training-free framework enabling test-time policy optimisation without any gradient updates, maintaining a dynamic non-parametric memory of experiences (ByteDance, 2026). These are the first credible architectural answers to the Nov 2024 question of how to move past static-weights deployment, but they are research-stage; no deployed frontier model uses them. Memory systems sit at Level 3 for deployed general LLMs (with substantial product-layer extensions), with research-stage Level 4 emerging.

Learning processes

For general-purpose LLMs, the fundamental training paradigm has not changed. Pre-training and post-training (RLHF, RLVR, reasoning RL) is still followed by static deployment. Catastrophic forgetting worsens with model scale, making the problem harder at scale (Luo et al., 2025). For embodied foundation models, genuine Level 4 incremental learning has emerged in narrow domains. The π0.6 paper title "a VLA that Learns from Experience" is descriptive rather than aspirational (Physical Intelligence, 2025). The HY-Embodied-0.5 system uses an iterative, self-evolving post-training paradigm, combining cold-start data with iterative reinforcement learning and rejection sampling supervised fine-tuning. Embodied foundation models are now routinely trained on continuously expanding interaction datasets, often combining web-scale pre-training with robot-specific experience. For agentic LLM systems, partial Level 4 properties emerge through tool use and verification: reasoning models can identify knowledge gaps, trigger retrieval

and verify results. Learning processes sit at Level 3 for general LLMs; Level 4 for embodied foundation models in narrow physical-task domains.

Integration

The November 2024 commentary identified integration as the central remaining challenge. This is fundamentally still the situation, with some nuance. Long-context windows combined with retrieval enable a form of immediate-recall and long-term-storage integration within a session. Persistent memory features connect personal experiences with general facts. Tool use connects procedural with declarative knowledge. TTT architectures explicitly target dual-memory integration. What has not changed: cross-session integration remains retrieval-mediated rather than architectural; procedural and declarative knowledge remain largely separated by training regime; true human-like cognitive flexibility and efficiency is essentially absent. The dimension sits at Level 3 with elements of Level 4 in agentic deployments.

Synthesis

Functional theme	Nov 2024	May 2026 (general LLM)	May 2026 (embodied / specialist)
Knowledge access	3	3+ (tool-extended)	3–4 (world models)
Working memory	3	3+ (long context)	3
Long-term / cross-session memory	3	3 (retrieval-based at product layer)	3
Architectural continual learning	3	3 (research-stage 4 via TTT)	4 (in narrow domains)
Active / experiential learning	3	3 (with tool use)	4 (in narrow domains)
Integration	3	3+	3+

Applying the "most aspects" rule yields general-purpose state-of-the-art at Level 3 sustained (3.0) — the November 2024 placement holds. Embodied foundation models in physical-task domains meet enough of the Level 4 criteria to warrant Level 4 in those narrow domains (coded 4.0 narrow). Research-stage TTT and JitRL work demonstrates that architectural Level 4 for language is technically achievable but not yet deployed.

Methodological observations to flag for the next revision

First, this scale benefits enormously from the November 2024 design choice to make the Level 3 to Level 4 boundary architecturally specific. Across the other scales, deployment features and prompting techniques have sometimes blurred level boundaries; here, the requirement of incremental learning through interaction creates a clean architectural test that deployment

features cannot satisfy by themselves. Even ChatGPT's "remember everything you've ever said" memory feature, which feels strongly like Level 4 to users, does not update model weights and so does not cross the boundary. The next revision should preserve this design choice.

Second, the gap between perceived and architectural capability is now policy-relevant. Tens of millions of users experience persistent-memory ChatGPT and Claude as systems that learn from interaction. Some form attachments to the resulting relationships (as covered in the Social Interaction update); some report concern when memory is wiped or when companies change models. Communicating clearly that the system is not actually learning, but that apparent learning is retrieval over a separately maintained store, matters for managing user expectations.

Third, embodied foundation models are now a distinct enough category to warrant separate treatment. The November 2024 framing treated knowledge/learning/memory as fundamentally a property of LLMs. The 2025–2026 evidence is that embodied foundation models constitute a distinct system class that satisfies more of the Level 4 description than any general-purpose LLM does.

Fourth, Test-Time Training is the architectural development most worth watching. The Stanford/NVIDIA TTT-E2E work and ByteDance In-Place TTT work are the first credible architectural responses to the Nov 2024 question of how to move past static-weights deployment for language models. The path from research to deployment has typically been 12 to 24 months. By the time the OECD's 2026/2027 update cycle reaches this scale, the question of whether frontier deployed LLMs incorporate test-time training will be live.

3.7 Vision

Headline placement (May 2026): Level 3 with strong Level 4 elements [3.5]

The Vision scale moves from Level 3 (with a small number of systems showing limited Level 4 capabilities) to Level 3 with strong elements of Level 4, edging toward a lower-threshold Level 4 placement on the breadth-of-tasks dimension specifically. The numerical code is 3.5. The most consequential change is methodological: the dominant paradigm has shifted from narrow vision applications (the basis of the November 2024 sample of 120 systems) to native multimodal foundation models (GPT-5, Gemini 3 Pro, Claude 4.7) that perform many of those tasks in a single unified architecture. Under the report's "most aspects" rule, this shift alone substantively moves the placement. Persistent failure modes (compositional counting, spatial reasoning, visual hallucination) keep the placement short of solidly Level 4 across the board.

This is the scale where the analytical question of what to measure matters most. The Nov 2024 evaluation methodology, sampling 120 specialised applications across the 32 component visual capabilities, captured the field accurately at that time. The 2025–2026 frontier is dominated by general-purpose VLMs that span many of those component capabilities in a single system; applying the same sampling methodology now would mix qualitatively different things.

Frontier systems being assessed

Three classes of system with substantially different vision profiles. First, native multimodal foundation models: GPT-5/5.1/5.2, Gemini 2.5 Pro / 3 Pro / Deep Think, Claude Opus 4.5 through 4.7, Grok 4. Gemini 3 is designed as a fully multimodal model, processing text, images, video, audio, charts, UI screenshots and structured documents through a unified representation space, with each modality handled as a primary input (Comanici et al., 2025; Vellum, 2025). Second, deployed specialised vision systems: autonomous driving (Waymo, Tesla FSD, Wayve), medical imaging, manufacturing inspection, industrial monitoring. The Waymo Driver is now powered by a foundation model based on a combination of LLMs and VLMs (Contrary Research, 2025). Third, embodied vision systems integrating vision with action selection: π0 / π0.6, Gemini Robotics-ER and Gemini Robotics 1.5, ByteDance GR-3, HY-Embodied-0.5. The categorical distinction matters because the Nov 2024 framework anchored the Level 4 description on the kitchen-assistant-robot example — a single system that recognises shapes, locates objects, identifies manipulation points, tracks motions and assesses results.

Dimension-by-dimension analysis

Breadth and variability of objects and scenes interpreted

The Level 3 descriptor allows several types of data and content (microscopy, RGB, natural scenes); Level 4 requires a wide range including microscopy, RGB, humans, mechanical parts and natural scenes. Native multimodal frontier models clearly satisfy the Level 4 breadth criterion. Gemini 3 Pro achieves high scores on MMMU-Pro (81.0%) and Video-MMMU (87.6%), suggesting strong ability to process and reason across temporal and spatial dimensions simultaneously (Vellum, 2025). Document understanding, chart reading, UI screenshot interpretation, video analysis, medical imaging interpretation and architectural analysis are now all performed by a single class of system rather than by specialised applications. The dimension sits at Level 4 for native multimodal systems.

Robustness to variation

The Level 3 descriptor allows some variation in lighting and target object appearance; Level 4 requires significant variations in lighting, shape and appearance, and subtle discrimination between similar classes. Frontier multimodal models handle most everyday lighting and appearance variation well, sufficient for autonomous-driving deployment: Waymo completed 14 million fully autonomous trips in 2025, tripling its 2024 volume, and processed over one million driverless rides per month by spring 2025 (Waymo, 2025). Robustly handling unfamiliar conditions sufficient for safe driving is genuinely Level 4 territory in that narrow deployment. But adversarial failure modes are sharp. Auto-Comp demonstrates that all tested VLMs, regardless of pretraining data or scale (including CLIP and SigLIP families), are susceptible to compositionality errors, especially when presented with low-entropy distractors such as repeated colours or object types, exposing the bag-of-words nature of many representations: on colour-binding with three objects, performance is barely above random chance on swap or confusion tasks (Sbrolli et al., 2026).

The dimension sits at Level 4 in standard deployment conditions; Level 2–3 under compositional or adversarial conditions. The honest aggregate is roughly Level 3, with Level 4 capacity demonstrated where the visual statistics are familiar.

Diversity of tasks performed

The Level 3 descriptor allows more than one subtask; Level 4 requires many different tasks with integration where one task's output feeds another. This dimension has moved most decisively. Native multimodal models perform dozens of distinct visual tasks (recognition, OCR, document parsing, chart reading, scene understanding, visual question answering, video captioning, code-from-screenshot, image editing, visual search, spatial reasoning, mathematical diagram interpretation) through a single architecture. The Waymo Foundation Model is a modular transformer-based neural network where different models contribute to perception, prediction, decision-making and control, with one model's outputs acting as inputs to the next (Contrary Research, 2025) — directly matching the Level 4 description. The dimension sits at Level 4.

Ability to learn through feedback

The Level 4 descriptor requires performance improvement through feedback, whether from self-assessment or external sources. This is where the picture remains weaker, for the same reasons covered in the Knowledge/Learning/Memory update. Three forms of progress are worth noting: reasoning models with extended thinking provide within-session self-feedback; tool-mediated feedback lets agentic systems verify visual outputs against external sources; embodied foundation models genuinely learn from physical interaction in narrow domains. But true Level 4 visual learning — adapting visual representations themselves through deployment-time feedback — remains rare. The November 2024 finding that current AI vision systems are unable to improve performance based on self-feedback remains substantially true for general-purpose vision systems, with the embodied-VLA exception. The dimension sits at Level 3 for general VLMs; Level 4 in narrow embodied domains.

Component visual capabilities

The Nov 2024 evaluation found about a third of the 32 component capabilities at Level 2, substantial numbers at Levels 1 and 3, and a small number at Level 4. The 2025–2026 picture is more uneven but with the centre of mass clearly shifted up. Capabilities now robustly at Level 4 include general object detection and recognition in everyday scenes, document understanding and OCR, chart and diagram interpretation, visual question answering on familiar content, scene description and captioning, video understanding for short and medium duration, and cross-modal grounding. Capabilities still at Level 3 or below include counting (especially compositional: VLMs can count reliably when only one shape type is present but exhibit substantial failures when multiple shape types are combined; Huang et al., 2025); spatial reasoning beyond basic relationships (existing benchmarks still target basic spatial understanding such as left/right, front/back, near/far; more demanding evaluations show substantial room for improvement); compositional binding (which colour goes with which object when multiple are present), at chance

levels for three or more objects; long-video temporal reasoning, which degrades with duration despite Video-MMMU progress; geometric analysis for precision tasks; and visual learning (representation update through deployment).

Synthesis

Dimension	Nov 2024	May 2026
Breadth of data types / contents	3	4
Robustness to variation	3 (modest variation)	4 (familiar) / 3 or below (compositional / adversarial)
Task diversity	3 (more than one subtask)	4
Learning through feedback	3 (none or limited)	3 (4 in narrow embodied)
Composite via 32-capability inventory	~3 (most at L2–3, few L4)	~3.5 (substantial cluster now L4, some persistent L2–3)

Applying the "most aspects" rule, the centre of mass is now at Level 3 with substantial movement into Level 4 — comfortably ahead of November 2024 but not yet solidly Level 4 across the breadth. The closest analogue is the Language scale's "Level 3 to lower threshold of Level 4" but with a smaller move. The headline is Level 3 with strong Level 4 elements (3.5), edging toward the same Level 4 lower-threshold framing the Language scale received, with the qualification that the compositional and spatial failures are real and substantive enough to keep the placement short of unambiguously Level 4. Real-world deployed systems (Waymo, embodied VLAs, document-processing agents) are arguably closer to Level 4 than benchmark scores alone suggest, because deployment selects for the visual content the systems handle well.

Methodological observations to flag for the next revision

First, the November 2024 methodology of sampling 120 specialised applications no longer matches the field. The dominant 2025–2026 paradigm is native multimodal foundation models that perform many of those 120 application tasks in a single architecture. The 32 component visual capabilities remain a useful taxonomy, but the unit of evaluation should probably shift from "specialised applications" to "VLM-class systems evaluated across the capability taxonomy."

Second, the compositional and spatial failure modes deserve elevated treatment. The Nov 2024 remaining-challenges section flagged general difficulties with diverse real-world environments. The 2025–2026 evidence has surfaced more specific structural failures: compositional counting, colour-object binding at chance for N≥3, spatial relationship reasoning under perturbation, and

visual hallucination. These are structural failures of VLM representations, not just current-system limitations to be patched.

Third, real-world deployment provides Level 4 evidence the benchmarks understate. The 14-million-trip Waymo deployment in 2025 is empirical evidence that current vision systems are sufficient for safe operation in deployed conditions, which is precisely the Level 4 "close to human level in tasks performed" criterion for that narrow but economically important task. The next revision could productively pair benchmark results with deployment-scale evidence.

Fourth, the embodied VLA category is now distinct enough to warrant separate treatment here. The Level 4 example in the Nov 2024 description — "a kitchen assistant robot might need to recognise shapes, locate objects, identify manipulation points, track motions and assess the quality of results" — is now closer to operational than the original framing suggests.

3.8 Manipulation

Headline placement (May 2026): Level 3 with leading edge into Level 4 [3.3] for research VLA systems; Level 2 sustained [2.0] for deployed humanoids

The Manipulation scale moves from a uniform Level 2 placement in November 2024 to a multi-profile picture. State-of-the-art for research-stage Vision-Language-Action (VLA) systems and select staged deployments is at Level 3 with leading edge into Level 4 in narrow demonstrations, coded 3.3. Robustly deployed systems remain at Level 2 sustained, coded 2.0. The upward movement is real but more qualified than on the Vision or Problem Solving scales, for a specific reason: this is the scale where the gap between research-stage capability and deployment robustness is widest. VLA models in research demos clear most Level 3 criteria and reach into Level 4; the same systems in real-world deployment, on real hardware, with real reliability requirements, sit much closer to Level 2 with occasional Level 3 successes.

The Nov 2024 commentary's identification of dexterity, adaptability and real-time decision-making and learning as the binding bottlenecks has held up well. The VLA architecture provides a credible path through the adaptability and real-time-learning bottlenecks, while leaving dexterity (a hardware-and-control problem) more or less where it was.

Frontier systems being assessed

Three classes of system with substantively different manipulation profiles. First, Vision-Language-Action research systems: Physical Intelligence's π0, π0.5 and π0.6 (Black et al., 2024; Black et al., 2025; Physical Intelligence, 2025); ByteDance's GR-3; Google DeepMind's Gemini Robotics, Robotics-ER and Robotics 1.5 (Google DeepMind, 2025a); HY-Embodied-0.5; Toyota Research Institute's Large Behavior Model; RDT2 (February 2026, zero-shot cross-embodiment). These are general-purpose manipulation foundation models trained on cross-embodiment data and deployed on various robot platforms. Second, humanoid robot deployments: Tesla Optimus Gen 3 (22-degree-of-freedom hands, October 2025 cooking and cleaning demonstrations); Figure 02

and 03 at BMW Spartanburg; 1X NEO (February 2026 consumer launch at USD 20,000 / USD 499 monthly rental); Apptronik Apollo; Agility Digit; LG CLOiD (CES 2026 laundry-folding demonstration); UniX AI Panther ("first service humanoid robot in real household deployment," April 2026, approximately 100 units per month); Hexagon AEON at BMW Leipzig; XPENG IRON. Third, traditional industrial robotic systems: robot arms in manufacturing, pick-and-place systems, surgical robots.

Dimension-by-dimension analysis

Task characteristics

VLA research systems substantively demonstrate Level 3 task characteristics. π0 was demonstrated folding laundry, bussing tables, packing eggs, grinding coffee beans and routing cables — a range unavailable to end-to-end learned policies in November 2024. π0.5 extended this to long-horizon dexterous manipulation in entirely new homes (Black et al., 2025). π0.6 demonstrated continuous coffee-making for a full day, hours of clothing-folding in a home environment and factory-relevant box assembly speeds, using the Recap framework (Reinforcement Learning with Experience and Corrections via Advantage-conditioned Policies) for learning from experience (Physical Intelligence, 2025). The continuous-coffee-making demonstration genuinely meets the Level 3 moderate-time-constraints criterion and edges into Level 4. Deployment evidence is more cautious: the 1X NEO Wall Street Journal test was unambiguous that all tasks the reporter observed were teleoperated rather than autonomous (Wall Street Journal, 2025). The LG CLOiD CES 2026 demonstration was capable of laundry folding but "only very slowly and with help." The Tesla Optimus Gen 3 October 2025 demonstration was real but involved staged training and operator assistance. The dimension sits at Level 3 for VLA research systems; Level 2–3 for deployed humanoids.

Object characteristics

This dimension shows perhaps the cleanest upward movement. VLA models routinely handle rigid objects of varying shape and size (Level 2–3), non-rigid objects including laundry, towels and food (Level 4), objects with moving parts including dryers, dishwashers and drawers (Level 4), and some materials with challenging surface properties. The π0.5 paper explicitly framed open-world generalisation as the central challenge (Black et al., 2025). However, the Bain analysis pushes back hard on deployment robustness: current humanoid robots can grasp objects but cannot manipulate them with anything approaching human dexterity (Bain & Company, 2025). The dimension sits at Level 3 with elements of Level 4 for research VLAs; Level 2 for typical industrial robots; Level 3 for the best humanoid demonstrations.

Environment

This is where VLA research systems most clearly enter Level 4 territory. The π0.5 demonstration of cleaning kitchens and bedrooms in entirely new homes meets the Level 4 standard of significant clutter and varying scenes (Black et al., 2025). The REALM benchmark provides standardised evaluation of generalisation across 15 perturbation factors, seven manipulation skills and more

than 3,500 objects (Stanford et al., 2025). However, deployment is more sobering: humanoid robot strengths in factories and warehouses cluster around repetitive material handling, while the bottleneck for home humanoid robot domestic chores is dexterous manipulation under uncertainty (AwesomeRobots, 2025; Bain & Company, 2025). The dimension sits at Level 3 for research VLAs (with Level 4 elements in demonstrations); Level 2 for actual deployment scale.

Task constraints

This dimension has moved least. Most VLA demonstrations operate at well below human speed; the LG CLOiD laundry folding is slow and assisted; π0.6 continuous coffee-making operates at extended timeframes. Strict time constraints, the Level 4 / Level 5 criterion, remain largely unmet. RDT2's table-tennis demonstration is the notable exception: dynamic, real-time, dexterous manipulation that matches Level 4 time-constraint criteria in a narrow domain. The dimension sits at Level 2 generally; Level 3–4 in narrow real-time demonstrations.

Generalisation

VLA models have produced the most fundamental shift here. The 2025–2026 picture is qualitatively different from the 2024 picture: RT-2 generalises to novel objects and unseen instructions and performs basic reasoning; π0 yields strong zero-shot generalisation and easy adaptation via fine-tuning (Black et al., 2024). RDT2's claim is the strongest: one of the first models that simultaneously zero-shot generalises to unseen objects, scenes, instructions and even robotic platforms, outperforming state-of-the-art baselines in dexterous, long-horizon and dynamic tasks like table tennis (Liu et al., 2026). Generalisation claims should be weighed against the reliability gap, which the Recap framework, π0.6's learning-from-experience approach and JitRL-style continual learning collectively try to address — but they are research-stage rather than deployed. The dimension sits at Level 3 with strong Level 4 elements for VLA research systems; Level 1–2 for traditional industrial robots; Level 2 for deployed humanoids.

Synthesis

Dimension	Nov 2024	May 2026 (research VLA)	May 2026 (deployed humanoid)	May 2026 (industrial)
Task characteristics	2	3 (Level 4 narrow)	2–3	1–2
Object characteristics	2	3–4	2–3	2
Environment	2	3 (Level 4 demos)	2	1–2

Task constraints	2	2 (Level 4 narrow)	2	2–3
Generalisation	2	3–4	2	1–2

Applying the "most aspects" rule and treating the research-VLA and deployed-humanoid columns as the relevant frontier, the headline is Level 3 with leading edge into Level 4 for VLA research systems (coded 3.3); Level 2 sustained for robustly deployed systems (coded 2.0). The honest aggregate placement is a one-level upward shift from November 2024 but with the strongest caveats of any scale update. This is comparable in shape to the Problem Solving update, but the gap between research demonstration and deployed reliability is much wider here. For Problem Solving, frontier reasoning models are deployed and reliably exhibit Level 3 behaviour; for Manipulation, frontier VLAs demonstrate Level 3–4 behaviour in studies but most deployed humanoid robots remain in pilot phases, heavily dependent on human input.

Methodological observations to flag for the next revision

First, the VLA architecture is a genuinely new system class that the November 2024 framework did not fully anticipate. The Nov 2024 evaluation was anchored on traditional manipulation systems with task-specific training. VLA models are foundation models for manipulation, trained on cross-embodiment data, deployed across different robot platforms. The capability profile is qualitatively different: broad generalisation with reliability gaps, rather than the narrow-and-reliable profile of traditional industrial robots. The next revision could split the scale's evaluation between task-specific systems and foundation-model-based systems.

Second, the research-deployment gap is larger on this scale than on any of the others. On Language, frontier reasoning models are widely deployed; on Vision, native multimodal models are widely deployed; on Knowledge/Learning/Memory, persistent-memory products are widely deployed. On Manipulation, the most impressive systems are research-stage. The next revision should probably report two placements explicitly: research-frontier capability and deployment-scale capability.

Third, the Nov 2024 identification of dexterity as the binding bottleneck has been validated. Across the VLA literature, the consistent finding is that AI software has outpaced hardware: foundation models can plan and reason about manipulation tasks at Level 3–4 capability, while the physical execution through current robot hardware sits at Level 2 capability. Tesla Optimus Gen 3's 22-degree-of-freedom hands are twice the prior iteration but still below the approximately 27-DOF capability of human hands, and dexterity is not only about degrees of freedom but also about tactile sensing, force control and the integration of these with vision and planning.

Fourth, household deployment is now genuinely testing Level 3–4 capability at consumer scale. The 1X NEO February 2026 launch and UniX AI Panther April 2026 household deployment are the first real consumer-scale tests of Level 3–4 manipulation claims. By the time the OECD's

2026/2027 update cycle reaches this scale, there will be 6–18 months of real consumer deployment data to evaluate.

3.9 Robotic intelligence

Headline placement (May 2026): Level 3 with strong multi-profile structure

The Robotic Intelligence scale is the synthesis scale, and so the placement reflects the analytical threads developed across the previous eight updates. State-of-the-art for embodied foundation model systems is at Level 3 with leading edge into Level 4 (coded 3.3); typical industrially-deployed robots remain at Level 2 sustained; autonomous vehicles in the narrow driving domain reach Level 4 sustained (coded 4.0 narrow). The upward movement is real and substantial, but the multi-profile character of the placement is more important than the headline number, for two reasons: the gap between research-stage and deployed-scale robotic intelligence is the widest of any scale; and the categories of robotic system that count as state-of-the-art have diverged enough that a single placement obscures more than it reveals.

The Nov 2024 commentary's identification of operational reliability, complex task adaptability and contextual cognition as the binding constraints has held up well. The VLA / embodied foundation model paradigm provides a credible architectural path toward addressing them, while real-world deployment evidence continues to demonstrate how far that path still has to go.

Frontier systems being assessed

Four classes of system are now distinguishable. First, embodied foundation model platforms: Physical Intelligence's π0, π0.5 and π0.6; Google DeepMind's Gemini Robotics, Robotics-ER and Robotics 1.5; ByteDance's GR-3 and HY-Embodied-0.5; RDT2; Toyota Research Institute's Large Behavior Model. These are AI-stack innovations deployed across various robot bodies. Second, humanoid robot platforms: Tesla Optimus Gen 3, Figure 02 and 03, 1X NEO, Apptronik Apollo, Agility Digit, LG CLOiD, UniX AI Panther, XPENG IRON, Hexagon AEON, Unitree H1 and H2. These are body innovations, increasingly running embodied-foundation-model brains. Third, specialised mobile robots: autonomous vehicles (Waymo, Tesla FSD, Wayve), warehouse autonomous mobile robots (Symbotic, Locus, Geek+, AutoStore), delivery robots (Starship, Nuro), agricultural and inspection robots. Fourth, specialised manipulation systems: surgical robots (Da Vinci and successors), industrial assembly arms, pick-and-place systems. The Nov 2024 evaluation was anchored primarily on classes 3 and 4, where Level 2 was the honest reading. The 2025–2026 frontier is led by classes 1 and 2.

Dimension-by-dimension analysis

Perception and environmental understanding

This is largely the Vision scale's territory applied to robots, with the addition of multi-sensor fusion (LiDAR, depth, IMU, tactile). The picture is at Level 3 with strong Level 4 elements, consistent

with the Vision update conclusion. Native multimodal models such as Gemini Robotics-ER and π0.5 provide robot perception with substantially broader scene understanding than November 2024 systems. Deployed autonomous vehicles demonstrate Level 4 perception in the driving domain at substantial scale. The RoboBench framework now offers a standardised evaluation across five dimensions: instruction comprehension, perception reasoning, generalised planning, affordance prediction, and failure analysis (Chen et al., 2025).

Locomotion and mobility

Autonomous driving sits at Level 4 in deployed form: Waymo's 14M-trip scale in 2025 is robust evidence of safe locomotion in complex, dynamic, multi-agent environments. Tesla FSD operates at far larger scale in supervised form. Legged locomotion has matured substantially: quadrupeds (Spot, Anymal, Unitree) now traverse complex unstructured environments using proprioceptive sensing (Sathyamoorthy et al., 2025). Humanoid bipedal locomotion has reached deployment readiness in semi-structured factory and warehouse settings, but balance under disturbance remains fragile; the AIDOL Moscow demonstration illustrates the reliability gap (the robot fell face-first on stage during its November 2025 public debut). The dimension sits at Level 4 for autonomous driving; Level 3 for general legged and wheeled mobility; Level 2 for humanoid bipedal locomotion under disturbance.

Manipulation

Treated in detail in the previous section. Level 3 with leading edge into Level 4 for VLA research systems; Level 2 sustained for robustly deployed systems. The capacity-versus-deployment gap is widest on this dimension.

Knowledge, learning and memory in embodied contexts

Embodied foundation models have most clearly extended beyond general-LLM capability. Level 4 in narrow embodied domains for systems like π0.6 ("learns from experience"), Gemini Robotics 1.5 and HY-Embodied-0.5's iterative self-evolving post-training. The Recap framework explicitly addresses the deployment-time learning gap.

Task planning and long-horizon reasoning

VLA models now demonstrate genuine multi-step task execution. ThinkAct, EmboTeam and the BEHAVIOR Challenge represent the maturing benchmark infrastructure. But the structural finding is clear: embodied control bottlenecks are visible across LoHoRavens, VLABench, CookBench and RoboCerebra benchmarks, with current vision-language-action models effective only for primitive or short tasks and performance sharply degrading in composite, long-horizon or unseen configurations. Context overload and cumulative error propagation are principal failure causes in agentic benchmarks. The ∞-THOR Needle-in-Embodied-Haystack benchmark demonstrates that cross-stage memory and reasoning over hundreds of steps remains a research-stage challenge (Park et al., 2025). The inner-monologue chain-of-thought pattern from the Metacognition update has been adapted to embodied settings, increasing robustness in long-horizon manipulation and

navigation. The dimension sits at Level 3 with elements of Level 4 in research systems; Level 2 for sustained reliability.

Multi-robot coordination and human-robot collaboration

Warehouse multi-robot coordination is mature in structured settings: Symbotic, Locus and Geek+ operate fleets of hundreds to thousands of robots. Mixed AMR/human warehouse work is now standard: AGVs, AMRs and robotic arms are widely deployed to support material handling and assembly tasks (Frontiers in Robotics, 2025). Human-robot collaboration is now deployed at industrial scale through cobot platforms (Universal Robots, Franka, Hexagon AEON). But sustained collaboration in unstructured environments — particularly home contexts — remains Level 2. The 1X NEO and UniX Panther household deployments depend heavily on remote human operators. The dimension sits at Level 3 in structured collaborative settings; Level 2 in unstructured ones.

Real-world reliability and deployment scale

This dimension most distinguishes the November 2024 placement from the May 2026 reality. Specialised deployed systems (autonomous vehicles, warehouse AMRs, industrial cobots, surgical robots) operate reliably at substantial scale in their target domains — genuinely Level 3–4 reliability in those specific settings. Embodied foundation model platforms in research and pilot settings demonstrate Level 3 capability but Level 2 reliability. Humanoid robots at consumer or general scale are now in early commercial deployment but with substantial reliability and supervision constraints: most humanoid robots today remain in pilot phases, heavily dependent on human input for navigation, dexterity or task switching, with current demos often masking technical constraints through staged environments or remote supervision (Bain & Company, 2025). The Bain timeline framing remains apt: near-term (2025–2027) expects scaling from several hundred units to low thousands globally, concentrated in controlled environments.

Synthesis

Dimension	Nov 2024	May 2026 (foundation-model robotics)	May 2026 (deployed humanoid)	May 2026 (specialised: AV / warehouse / surgical)
Perception	2	3 (Level 4 elements)	2–3	4
Locomotion	2	3	2–3	4 (AV)
Manipulation	2	3–4 in research	2–3	2–3

Embodied knowledge / learning	2	4 in narrow domains	2	2–3
Task planning / long-horizon	2	3	2	3
Coordination / HRI	2	3 (structured)	2	3–4 (structured)
Reliability / deployment scale	2	2 (research-stage)	2	4 in domain

Applying the "most aspects" rule, the picture splits along system class lines. Embodied foundation model platforms in research settings sit at Level 3 with leading edge into Level 4 (coded 3.3) — a full level above the Nov 2024 placement. Humanoid robots in deployed form sit at Level 2 with elements of Level 3 (coded 2.3) — a half-step up. Specialised deployed systems sit at Level 3 in general; Level 4 in autonomous driving specifically (coded 4.0 narrow) — the cleanest Level 4 deployment evidence exists here. A single headline placement that captures this picture honestly is Level 3 with strong multi-profile structure — a full level above Nov 2024, but with the caveat that the upward movement is real for some classes and limited for others.

Methodological observations to flag for the next revision

First, the Robotic Intelligence scale is now most useful as a synthesis scale, and that role should be made explicit. With Vision, Manipulation, Knowledge/Learning/Memory, Social Interaction and Metacognition each providing more detailed evaluation of specific capabilities, Robotic Intelligence should reposition as the integrative assessment, capturing how those capabilities combine in an embodied agent operating in the real world. Integration robustness, deployment scale and reliability under real-world conditions are the integrative properties that the component scales do not capture.

Second, the four-class system taxonomy (foundation-model platforms, humanoids, specialised mobile robots, specialised manipulation systems) is now distinct enough to deserve separate treatment. The Nov 2024 evaluation methodology was anchored on a survey across application areas, which made sense when systems were primarily defined by application domain. The 2025–2026 frontier is more usefully understood by system class than by application area: a Gemini Robotics 1.5 model deployed on a Boston Dynamics Atlas, an Apptronik Apollo and a Franka Panda is the same system class across different bodies.

Third, the research-deployment gap is the central methodological challenge for this scale. Capacity in studies and reliability in deployment have diverged substantially. On Robotic Intelligence this gap is widest. The next revision should report two placements explicitly: capability-frontier and deployment-frontier.

Fourth, the autonomous-driving domain is now the cleanest Level 4 robotic intelligence case study and deserves substantial treatment. Waymo's 2025 deployment scale provides the empirical anchor that the November 2024 commentary noted was missing in most application areas. The next revision could build a detailed case study around autonomous driving — what made the Level 2 to Level 4 transition happen, what conditions were required, and what lessons it offers for the path forward in other domains.

Fifth, the binding bottleneck has shifted. The Nov 2024 identification of dexterity as binding has been validated. The new binding bottleneck is reliability and integration. AI software has outpaced robotic hardware, and the integration of foundation-model intelligence with traditional robotic control architectures is the new challenge. This implies that policy attention should focus on safety certification, deployment governance and reliability standards rather than primarily on AI capability development.

3.10 Recommendations for revising and using the scales

This subsection consolidates the methodological observations that appear across the nine per-scale updates into a set of structured recommendations for the OECD's planned revision of the framework. The observations cluster into three cross-cutting analytical threads and a set of more specific scale-by-scale recommendations.

Three cross-cutting analytical threads

The first thread is the gap between underlying capability and deployed calibration or robustness. The Nov 2024 framework was designed around capability ceilings — how well a system can perform a task under standard conditions. The 2025–2026 evidence makes clear that this is no longer the binding constraint on most scales. Frontier systems demonstrate substantial capability when conditions are familiar; their deployed behaviour exhibits miscalibration, overconfidence, hallucination, mode collapse, sycophancy or compositional failure under adversarial or out-of-distribution conditions. This pattern recurs across the Language scale (hallucination), the Social Interaction scale (theory-of-mind brittleness on ExploreToM), the Problem Solving scale (SWE-bench Verified at 82% versus SWE-bench Pro at 23%), the Creativity scale (mode collapse), the Metacognition scale (overconfidence and reluctance to express internal uncertainty), the Vision scale (compositional and spatial failures), the Manipulation scale (research-to-deployment reliability gap), and the Robotic Intelligence scale (the widest of any scale). The next edition could productively make capacity-versus-calibration a first-class methodological treatment, perhaps through an explicit robustness dimension threaded through each scale, or through dual reporting of capability-frontier and deployment-frontier placements where the gap is substantive.

The second thread is the divergence between general-purpose large language models and specialist hybrid systems. This is now a Level-distinct phenomenon rather than merely a profile-distinct one. Specialist hybrid systems including AlphaEvolve, AlphaProof with AlphaGeometry 2 and Deep Think, the Physical Intelligence π0 family, Gemini Robotics and the Waymo

autonomous driving stack now reach higher levels in narrow domains than general LLMs reach broadly. The Nov 2024 framework handled the symbolic/LLM contrast implicitly through worked examples. The 2025–2026 evidence suggests that an explicit specialist-versus-general distinction should be made first-class in the methodology, with separate placements reported where appropriate. The next edition could productively position specialist hybrid systems as a primary subject of evaluation, with explicit reporting conventions for distinguishing "Level N in a narrow domain via specialist hybrid" from "Level N broadly via a general system," since these have different policy implications.

The third thread is the divergence between performance on standardised or familiar tasks and performance on adversarial or out-of-distribution probes. SWE-bench Verified at 82% drops to SWE-bench Pro at 23% (Mehrotra et al., 2025); ARC-AGI-1 at 76% drops to ARC-AGI-2 at 24–45% (ARC Prize Foundation, 2025); standard theory-of-mind tasks at near-human levels drop to 0–9% on the ExploreToM adversarial benchmark (Sclar et al., 2025); standard vision-language counting handles single shape types but fails on compositional counting (Huang et al., 2025). This pattern is now structural enough across the scales to warrant explicit treatment as a robustness dimension, distinguishing performance on standardised evaluations from performance on adversarial or compositional perturbations of those evaluations.

Specific scale-level recommendations

Beyond the three cross-cutting threads, eight more specific recommendations emerge from the per-scale updates.

First, retain multi-profile reporting where it is genuine. Across several scales (Social Interaction, Creativity, Knowledge/Learning/Memory, Manipulation, Robotic Intelligence), the honest placement is multi-profile, with distinct codes for distinct system classes (disembodied versus embodied, general versus specialist, research versus deployed). Forced compression into a single headline placement obscures rather than clarifies. The original beta report's joint placement of GPT-4o and AIBO at the same Level 2 on Social Interaction, with explicitly different profiles, is a good template for this style of reporting.

Second, treat Robotic Intelligence explicitly as the synthesis scale. With Vision, Manipulation, Knowledge/Learning/Memory, Social Interaction and Metacognition each providing more detailed evaluation of specific capabilities, Robotic Intelligence should reposition as the integrative assessment, capturing how those capabilities combine in an embodied agent operating in the real world. This implicit role in the Nov 2024 framework could productively be made explicit, with the scale's commentary focusing on integration robustness, deployment scale and reliability under real-world conditions — the integrative properties that the component scales do not capture.

Third, make the embodied foundation model paradigm (Vision-Language-Action systems) a first-class evaluation category. The Nov 2024 framework's evaluation of Knowledge/Learning/Memory, Vision, Manipulation and Robotic Intelligence was anchored on system classes that the VLA paradigm has substantially reshaped. The next edition could

distinguish task-specific systems, general-purpose LLMs, and embodied foundation models as three distinct evaluation contexts across the relevant scales.

Fourth, give explicit treatment to mode collapse on the Creativity scale. The dominant empirical finding about general LLM creativity since November 2024 is that post-training alignment methods (principally RLHF) structurally suppress novelty and surprise through optimisation pressure toward modal responses. This is mechanistically well-understood and policy-relevant: tens of millions of users now use mode-collapse-affected systems for creative work and homework, with consequences for human creative diversity that are now becoming measurable. The Creativity scale narrative could be substantially strengthened by addressing this directly.

Fifth, preserve the architectural specificity of the Level 3 to Level 4 boundary on the Knowledge, Learning and Memory scale. The Nov 2024 design choice to make the Level 3 to Level 4 boundary architecturally specific (requiring incremental learning through interaction) has held up well as a structural test that deployment features alone cannot satisfy. This is particularly important because users now experience persistent-memory ChatGPT and Claude as systems that learn from interaction, even though the underlying weights do not update. Communicating clearly that apparent learning is retrieval over a separately maintained store matters for managing user expectations.

Sixth, give first-class treatment to the calibration thread on the Metacognition scale. The Nov 2024 framework treats confidence calibration as one of three dimensions, but the 2025–2026 evidence suggests that calibration is now the binding constraint on the move from Level 3 to Level 4 on the scale. Models internally encode uncertainty but fail to externalise it because alignment incentives reward confident answers. The scale could distinguish capacity to self-monitor from deployment behaviour of self-monitoring.

Seventh, distinguish standard from adversarial robustness explicitly. As noted in the third cross-cutting thread, performance on standardised or familiar tasks now diverges sharply from performance on adversarial or out-of-distribution probes across multiple scales. The Vision scale's compositional and spatial failures, the Problem Solving scale's SWE-bench Verified-versus-Pro gap, the Social Interaction scale's ToM standard-versus-adversarial gap and the Language scale's hallucination patterns all support the same methodological recommendation.

Eighth, anticipate the shift from "capability" to "calibration" framing in policy applications. The Nov 2024 framework's policy framing centred on the question of what AI systems can do in human-relevant domains. The 2025–2026 evidence makes increasingly clear that the policy-relevant question is how reliably deployed systems exercise the capabilities they possess. For Manipulation and Robotic Intelligence specifically, this implies that policy attention should focus on safety certification, deployment governance and reliability standards rather than primarily on AI capability development. For Language, Social Interaction and Metacognition, this implies more attention to calibration, hallucination, sycophancy and emotional appropriateness as deployment-time concerns rather than as capability ceilings.

4. Future updates

4.1 Using this Claude-authored draft within an expert review workflow

This update was prepared by a single artificial intelligence system, operating without expert panel oversight. Its claim on the OECD's attention is therefore quite specific: it offers reconnaissance of the current state of the field, conducted at higher speed and lower cost than an expert panel would conduct it, and intended to support rather than substitute for the OECD's planned expert revision process. Three roles for the Claude-authored draft within that workflow seem worth surfacing.

First, the draft can serve as agenda-setting input for expert panels. Each of the nine per-scale subsections identifies the principal systems, benchmarks and methodological issues that have emerged since November 2024, in a structured form that mirrors the Nov 2024 commentary. An expert panel revisiting any one scale could productively use the corresponding subsection as a starting list of evidence to evaluate, rather than reconstructing the recent literature from scratch. The placements themselves are best treated as hypotheses for the panel to confirm, revise or reject.

Second, the draft can support targeted expert review of contested placements. Some placements in this update are more confidently supported by the evidence than others: the Language move to 3.7 is anchored on dense benchmark evidence; the Robotic Intelligence multi-profile placement is anchored on more dispersed deployment-and-research evidence; the Creativity placement is the most analytically novel. An expert review process could productively allocate effort by contestedness rather than uniformly across the nine scales, with the Creativity, Social Interaction, Manipulation and Robotic Intelligence placements warranting the most rigorous scrutiny.

Third, the cite-as-you-go convention used throughout the per-scale subsections is intended to support transparent review of the evidentiary basis for each claim. Every benchmark result, system description and deployment statistic cited in the per-scale analyses is traceable to a source in the References section. This allows an expert review to focus on specific evidentiary disputes rather than on the framework's analytical structure. Where the draft errs by overstating, understating or misrepresenting a piece of evidence, that error should be visible to a reviewer with subject-matter expertise.

Three caveats apply to all three roles. The draft reflects the recall and selection biases of a single artificial intelligence system reading a large body of literature in a short time. Its citations are accurate to the level of recall verified at the time of drafting, but expert reviewers should expect to find some misattributions or out-of-date references. The placements reflect the Claude model's interpretive framework, which may differ systematically from the interpretive framework of the OECD's expert community in ways that warrant explicit discussion. And the draft should not be taken as Anthropic's institutional position on these placements; it is an output of a single Claude session, prepared at the request of a researcher engaged in the OECD's broader work programme.

4.2 Frequency of future updates

The Nov 2024 beta report was published with an explicit anticipation of biennial revision. The evidence assembled in this update suggests that biennial may now be too slow. Six of the nine scales have moved by half a level or more over the 18-month interval since the beta release; three (Social Interaction, Problem Solving, Metacognition) have moved by a full level; one (Manipulation) has gained an entire new system class (VLA models) that the Nov 2024 evaluation methodology did not contemplate. The rate of change is sufficient that a two-year update cadence would risk leaving the framework substantially behind the deployed state of the field at any given moment.

A revised cadence might combine three elements. First, an annual headline review across all nine scales, conducted at modest depth and intended primarily to confirm or refine the prior year's placements rather than to reconstruct them from scratch. Second, a biennial full revision, conducted on the model of the Nov 2024 expert workshops, with substantial expert input and full re-derivation of placements. Third, lightweight interim monitoring of the fastest-moving scales (Manipulation, Robotic Intelligence and the Metacognition behaviours that reasoning models affect most rapidly), with quarterly updates of a small number of leading indicators rather than full re-assessment of all dimensions.

The choice of cadence has substantive policy implications. If the framework is used as a reference vocabulary for OECD analyses of labour-market exposure, skills demand or AI safety governance, the gap between the framework's vintage and the deployed state of the field directly determines the accuracy of derived analyses. Half-level movements across the framework over 18-month intervals are large enough to substantively change conclusions in adjacent analyses; the AI-WIPS programme might consider whether the framework's value as a policy reference is best served by less-frequent but more-authoritative updates, or by more-frequent but more-provisional ones.

A complementary option is to make use of AI-assisted reconnaissance — of which this update is an early example — to support more frequent interim updates without requiring expert workshops at each cycle. The Claude-authored draft model demonstrated here can be reproduced at low cost on any cadence the OECD considers useful, with the explicit understanding that AI-assisted drafts are reconnaissance rather than authoritative. A workflow that combined annual AI-assisted reconnaissance with biennial expert workshop revision would balance currency against authority more efficiently than either approach alone.

4.3 Integration with the OECD's broader work programme

The framework's value as a policy reference depends on its visibility to and usability by the analytical communities that consume OECD outputs. The publication of the framework through the OECD's main publishing channels combined with the online repository at aicapabilityindicators.oecd.org provides the basic infrastructure for this; updates to the framework propagate to derived analyses through these channels. The companion volume AI and the Future

of Skills, Volume 3 (OECD, 2025) demonstrates one model for connecting the capability indicators to substantive policy analyses; the next edition could productively pair revision of the indicators with refresh of one or more companion analyses, so that movements in the framework translate visibly into updates in the policy outputs they support.

References

References below are listed alphabetically by first author or organisation. URLs are provided where the source is publicly available. Pre-print identifiers (arXiv numbers) are provided for pre-print sources. All references were accessed in May 2026 unless otherwise noted.

Anderson, B.R., Shah, J.H. and Kreminski, M. (2024), "Homogenization effects of large language models on human creative ideation", in *Proceedings of the 16th Conference on Creativity and Cognition*, Association for Computing Machinery.

ARC Prize Foundation (2025), *ARC-AGI-2: Towards more robust abstraction and reasoning benchmarks*, https://arcprize.org/arc-agi-2.

AwesomeRobots (2025), *Humanoid robots in 2025: Real deployments vs. hype*, industry analysis.

Bain & Company (2025), *Humanoid robots: From demos to deployment*, Bain & Company industry report.

Black, K. et al. (2024), "π0: A vision-language-action flow model for general robot control", *arXiv preprint*, arXiv:2410.24164.

Black, K. et al. (2025), "π0.5: A vision-language-action model with open-world generalization", *arXiv preprint*, arXiv:2504.16054.

ByteDance (2026), "In-Place Test-Time Training for large language models", technical report, ByteDance Research.

Character Technologies (2024), "Building Character.AI: Scale, infrastructure and deployment", company technical blog.

Chen, Y. et al. (2025), "RoboBench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain", *arXiv preprint*, arXiv:2510.17801.

Comanici, G. et al. (2025), "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities", *arXiv preprint*, arXiv:2507.06261.

Contrary Research (2025), *Deep dive: Tesla and Waymo — the trajectory of autonomous driving deployment*, Contrary Research report.

De Freitas, J. et al. (2025), "User attachment to AI companions: Evidence from a Replika natural experiment", Harvard Business School working paper.

Doshi, A.R. and Hauser, O.P. (2024), "Generative artificial intelligence enhances individual creativity but reduces the collective diversity of novel content", *Science Advances*, 10(28), eadn5290.

Flint (2025), *NoveltyBench: A benchmark for novelty in language model outputs*, Flint AI technical report.

Frontiers in Robotics (2025), "AI-driven human-robot collaboration in intralogistics systems", *Frontiers in Robotics and AI*, special issue.

Gao, Y. et al. (2025), "OriGene and Robin: Autonomous AI scientists for therapeutic discovery", pre-print, biorXiv.

Google DeepMind (2024), *AI achieves silver-medal standard solving International Mathematical Olympiad problems*, Google DeepMind blog post, July 2024.

Google DeepMind (2025a), *Gemini Robotics: AI models for the physical world*, Google DeepMind blog post.

Guan, T. et al. (2024), "HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models", *arXiv preprint*, arXiv:2310.14566.

Huang, Z. et al. (2025), "Your vision-language model can't even count to 20: Exposing the failures of VLMs in compositional counting", *arXiv preprint*, arXiv:2510.04401.

Introl (2026), *GPT-5.2 vs Gemini 3 Pro: Benchmark comparison and frontier model analysis*, Introl industry analysis.

Korbak, T. et al. (2025), "Chain-of-thought monitorability: A fragile opportunity for AI safety", *arXiv preprint*.

Kwa, T. et al. (2025), "Measuring AI ability to complete long tasks", METR (Model Evaluation and Threat Research) technical report.

Liu, R. et al. (2025), "Beyond mode collapse: Understanding the optimization dynamics of RLHF for language models", *arXiv preprint*.

Liu, Y. et al. (2026), "RDT2: Zero-shot cross-embodiment generalization in vision-language-action models", pre-print.

Luo, Y. et al. (2025), "Catastrophic forgetting at scale: Empirical study of forgetting dynamics in large language models", *arXiv preprint*.

Mehrotra, A. et al. (2025), *SWE-bench Pro: Towards multi-day, real-world software engineering evaluation*, technical report.

Novikov, A. et al. (2025), "AlphaEvolve: A coding agent for scientific and algorithmic discovery", Google DeepMind technical report.

NVIDIA (2025), *Cosmos: World foundation models for embodied AI*, NVIDIA technical blog and white paper.

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

OECD (2025), *AI and the Future of Skills, Volume 3*, OECD Publishing, Paris.

OpenAI (2025a), "Memory and new controls for ChatGPT", OpenAI product blog, April 2025.

OpenAI (2025b), "Sora 2: System card and capabilities", OpenAI technical documentation.

OpenAI (2025c), "OpenAI reasoning models: System cards for o1, o3 and GPT-5", OpenAI technical documentation.

Park, J. et al. (2025), "∞-THOR and Needle(s) in the Embodied Haystack: Long-context reasoning benchmarks for embodied agents", *arXiv preprint*, arXiv:2505.16928.

Petrov, I. et al. (2025), "Proof or bluff? Evaluating LLMs on 2025 USA Mathematical Olympiad", *arXiv preprint*, arXiv:2503.21934.

Phan, L. et al. (2025), "Humanity's Last Exam: A benchmark at the frontier of human knowledge", *arXiv preprint*, arXiv:2501.14249.

Physical Intelligence (2025), "π0.6: A vision-language-action model that learns from experience", Physical Intelligence company blog.

Sathyamoorthy, A.J. et al. (2025), "Perception and navigation algorithms for legged robots in unstructured outdoor terrains", *DSIAC Journal*.

Sbrolli, R. et al. (2026), "Auto-Comp: Diagnosing compositional reasoning failures in vision-language models", pre-print.

Sclar, M. et al. (2025), "ExploreToM: Adversarially generated theory-of-mind benchmarks", *arXiv preprint*.

Springboards (2025), *Creativity Benchmark: A practitioner-evaluated assessment of LLM creative output*, Springboards industry report.

Stanford and NVIDIA (2026), "TTT-E2E: End-to-end test-time training for long-context modeling", pre-print and Stanford technical report.

Stanford Internet Observatory (2025), *Risks of AI companion chatbots for adolescents and minors*, Stanford Internet Observatory technical report.

Stanford et al. (2025), "REALM: A benchmark for generalisation in robotic manipulation", *arXiv preprint*.

Steyvers, M. et al. (2025), "What large language models know and what people think they know", *Nature Machine Intelligence*, 7, pp. 221–235.

Strachan, J.W.A. et al. (2024), "Testing theory of mind in large language models and humans", *Nature Human Behaviour*, https://doi.org/10.1038/s41562-024-01882-z.

Street, W. et al. (2024), "LLMs achieve adult human performance on higher-order theory of mind tasks", *arXiv preprint*, arXiv:2405.18870.

Tao, T. (2025), "AlphaEvolve and the future of mathematical discovery", commentary essay, T. Tao personal blog and *Notices of the American Mathematical Society*.

Trends in Cognitive Sciences (2025), "Homogenization in language model outputs: Evidence, mechanisms and implications", *Trends in Cognitive Sciences*, review article.

Vellum (2025), *Gemini 3 Pro: Benchmark performance and multimodal capabilities*, Vellum AI industry analysis.

Wall Street Journal (2025), "1X NEO household humanoid robot: A reporter's test", *Wall Street Journal*, November 2025.

Wang, J. et al. (2025), "Lifelong-SOTOPIA: Evaluating social intelligence of language agents over multiple environments", pre-print.

Wang, R. et al. (2024), "SOTOPIA-π: Interactive learning of socially intelligent language agents", in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Waymo (2025), "14 million rides: Operational summary and safety review", Waymo company blog and technical report.

Wong, A. et al. (2025), "Algorithmic self-portrait: How ChatGPT memory shapes personalised AI interaction", pre-print.

Wu, Y. et al. (2023), "HI-TOM: A benchmark for evaluating higher-order theory of mind reasoning in large language models", in *Findings of the Association for Computational Linguistics: EMNLP*.

Yamada, Y. et al. (2025), "The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search", *arXiv preprint*, arXiv:2504.08066.

Yang, A. et al. (2025), "LLMs know when they know but do not act on it: Decoupling metacognitive capacity from metacognitive behaviour", *arXiv preprint*.

Zhang, J. et al. (2025), "Verbalized sampling: A method to recover diversity in language model outputs", *arXiv preprint*.

Zhou, X. et al. (2024), "SOTOPIA: Interactive evaluation for social intelligence in language agents", in *Proceedings of the 12th International Conference on Learning Representations*.