人类智能启发的通用人工智能发展阶段测试：从通用到应用 *

彭玉佳 **[1,2,3] 何欣怡 [2,4] 谢宏钊 [2] 肖茜之 [5] 王愉茜 [1,2] 朱松纯 [2,3,4] 张振亮 **[2]

（[1]北京大学心理与认知科学学院，北京市行为与心理健康重点实验室，生物与机器智能教育部重点实验室，北京，100871）

（[2]北京通用人工智能研究院，跨媒体通用人工智能全国重点实验室，北京，100080）（[3]北京大学人工智能研究院，北京，100871）

（[4]北京大学智能学院，北京，100871）（[5]北京大学元培学院，北京，100871）

摘 要 定义与测量通用人工智能（AGI）是人工智能领域的核心挑战。为此，本研究提出一种受人类智能启发的 AGI 发展阶段测试框架，涵盖从通用到应用的演进过程。该框架的核心是通用测试阶段，强调 AGI 需具备跨领域的基础认知能力，并通过对标人类儿童年龄的通用测试系列评估智能体的认知水平。针对大模型的测试结果表明，其当前能力与人类儿童存在显著差距。进一步，框架扩展至“通用 – 专项 – 应用”三阶段结构：专项阶段聚焦领域内自主学习和能力提升，应用阶段则关注真实环境下的泛化能力与行业适配性。此框架旨在为 AGI 评估提供一套系统的、受人类认知发展规律启发的标准与方法，推动其从通用能力向高阶应用智能转化，逐步融入人类社会场景。

关键词 人工智能 通用人工智能 认知发展 人工智能测试 具身智能

1 问题提出

1.1 人工智能发展和通用人工智能定义

人工智能的快速发展正在深刻重塑社会，既带来了前所未有的机遇，也引发了广泛的关注与担忧。随着生成式预训练模型（generative pre-trained transformer，GPT）模型系列的发布（Brown et al., 2020; OpenAI, 2024; Radford et al., 2019），通用人工智能（artificial general intelligence, AGI）再次成为人工智能（AI）领域的关注焦点。最近的基础模型显示出在特定领域内进行泛化的能力，如自然语言处理（NLP）（OpenAI, 2024）、图像分割（Kirillov et al., 2023）以及机器人技术（Driess et al., 2023）。近几年，人工智能逐步实现实用化转型，包括小型化（丁贵广等 , 2024）、专业化（Magesh et al., 2025; McDuff et al., 2025）等。

然而，大模型仍然不能被广泛认可为 AGI。AGI 重在“通用”（general），要求智能体必须能够适应一个动态的环境，就像地球上人类所生活的真实物理和社会环境一样，能够应对无限的随机任务（Peng et al., 2024）。这意味着 AGI 必须能够适应多变的具身动态环境，应对无数意想不到的场景并表现良好。为了保证能够可靠、自主地适应和生存于动态具身物理社会交互（dynamic embodied physical and social interactive，DEPSI）环境（Shu et al., 2021），通智测试（Tong test）提出，一个典型的 AGI 系统应当具有以下 3 个基本特征：应对无限任务、自主生成任务、价值驱动任务生成（Peng et al., 2024）。前期工作虽然已明确定义了 AGI 的基本标准，但从抽象标准迁移到具象测试和应用仍存在距离。

1.2 人工智能测试现状和挑战

对于 AGI 测试与评级，传统的人工智能测评方法提供了值得借鉴的思路，但存在一定的局限性。以最经典的图灵测试为例（Oppy & Dowe, 2003），测试的核心是判断一台机器是否具备人类的智能水平。然而，图灵测试在很大程度上依赖人类判断者自身的知识和认知水平，很难达到测试的客观化和标准化。图灵测试也已经不止一次地被聊天机器人（如 Google Duplex Voice AI）打败，而这些聊天机

* 本研究得到跨媒体通用人工智能全国重点实验室、科技部科技创新 2030–“新一代人工智能（2030）”重点研发计划（2022ZD0114900）项目和国家自然科学基金项目（32471151，32200854）的资助。

** 通讯作者：彭玉佳，E-mail: yujia_peng@pku.edu.cn；张振亮，E-mail: zlzhang@bigai.ai

DOI:10.16719/j.cnki.1671–6981.20260202

器人大多基于专门设计的应对策略算法才通过图灵测试，远远谈不上具有真正的智能。

基于数据集的任务导向基准测试，是指让人工智能算法在特定的数据集上执行特定的任务，并对其在该数据集上的表现进行评估和测量。在过去的二十年中，人工智能的各个子领域涌现出了大量不同种类的数据集，如 ImageNet、CoCo、Visual Question Answering （VQA） 等（Antol et al., 2015; Deng et al., 2009; Lin et al., 2014）。这些数据集为人工智能的研究和开发提供了丰富的资源和基础。然而，这种基于数据集的测评方式最显著的问题之一是这些任务导向的数据集基准过于强调解决高度专项的问题，不适合对 AGI 进行测试。此外，这种测评方式还容易导致“刷榜”现象，即研究人员和开发者为了在基准测试中获得更高的排名，会针对特定数据集进行优化。这种优化后的模型无法应对复杂、开放的现实场景。

近年来，也有多个针对多任务的测试平台诞生，如 MMBench 和 FlagEval （He et al., 2024; Liu et al., 2024），涵盖了跨模态的多维度测试，易于部署与实现，且覆盖了一定广度，但仍然无法克服 AGI 测试中对于无限任务、自主生成任务以及价值的测试。同时，上述平台并未涉及具身领域（如可交互的虚拟平台或现实机器人环境），也缺乏对 AGI 的完整评估体系，难以指导人工智能前瞻性研究。

在理论层面，关于人工智能的评级系统，也有多家研究团队尝试给出答案。如 OpenAI、Google DeepMind，也认为当前的大模型并非 AGI 的最终形态（Morris et al., 2024）。国内外各团队尝试对 AGI 进行定义与分级，但并未给出具体、完备的定义。这些框架虽然为 AGI 的理论研究提供了重要参考，但在将这些原则转化为可量化的评估标准方面仍存在实践挑战。

1.3 通智测试和 GSA 框架

鉴于传统人工智能测评方法的局限性，通智测试（Tong test; Peng et al., 2024; 表 1）明确了 AGI 的定义，并提出了一套基于动态物理社会环境交互的评级基准与测试系统。系统旨在全面评估通用智能体（即 AGI Agent）在多样化任务场景中的表现，确保其具备广泛的知识和技能泛化能力，并能够与人类价值观对齐。然而，通智测试虽然提出了相对完备统一的 AGI 测试理论框架，更需要一套可操作的实践体系，将抽象的 AGI 评估标准转化为具体的测试范式。上海人工智能实验室的测试榜单中也罗列了“以人为本”评测、土木工程评测，但并未归纳出所涉及众多评测间的关系，人工智能的评测框架处于缺失状态。基于此，我们受人类认知发展阶段的启发（Feldman, 2006），本文进一步在理论层面，提出潜在路线图“通用 - 专项 - 应用”（general-specialized-applicable，GSA，音译为“智萨”）三阶段发展模型，用以描述通用智能的潜在发展路径。该模型认为：通用能力构成特定技能发展的基石，对应人类 0~6 岁的多维能力和价值发展，如运动能力（如行走、抓握）开启环境探索之门，视觉系统引导运动目标定位，听觉功能支持接收养育者反馈。这些通用能力的协同作用，为专项（specialized）能力的学习提供了可能。而应用（applicable）能力又进一步铺垫了融入人类社会应用的基础。GSA 三阶段模型既是对 AGI 发展规律的理论阐释，也是指导测试设计的实践指南。本文致力于探讨 GSA 模型在描述能力空间发展轨迹方面的合理性，也探讨了该模型对 AGI 发展路径的指导价值，为智能发展理论研究提供了新视角。

通智测试是一个已公开的评测系统，本文在这个通智测试体系的基础上提出人工智能发展阶段的概念，发展阶段包括：通用、专项、应用，其中，本文的重点是通用任务，其与人类发展相对应。通过将 GSA 框架与通智测试结合，我们得到了更全面更系统的评价体系，原本的任务群的分层仅有“能力轴”这一维度（如视觉、语言、学习、运动、价值、推理），GSA 框架的加入使得任务的分布空间从单轴变为双轴：从“能力轴”升维至“能力轴、发展阶段轴”，明确特定能力的任务坐落在什么发展阶段（如通用、专项、应用）。对比发现，表 1 中的 AGI 代表性测试与评级平台，除 Tong test 外，在理论与工程部署层面不具备完备性，并不能同时给出定义、分级与可操作使用的测试；即便是给出了可使用的测试，也主要是围绕单项能力的测试，AGI 智能体任务较少。结合了 GSA 框架的通智测试，在理论完备性上有了长足发展，且在工程实践方面提供了一批可测任务与测试结果。

本文针对 GSA 理论框架，展开通用阶段的儿童和大模型对比实证测试，然后在理论层面对专项和应用阶段展开探讨。本研究假设，通用阶段测试任务可映射跨儿童年龄阶段的不同难度，并回答能力的聚类和发展特点。进一步，以此任务集测试大模

表 1 AGI 代表性测试与评级平台

	定义	分级	测试
MMBench 上海人工智能实验室	×	×	三层次视觉语言模型的感知推理测试
OpenAI	各阶段应用角色的语义定义	5 层：各阶段应用角色	×
Google DeepMind	6 项基本原则	6×2：依据性能与通用性分层	×
FlagEval 北京智源人工智能研究院	×	×	语言、文图生成及多语言文图大模型评测
Tong test 北京大学-北京通用人工智能研究院	三个完备性特征	5×6：依据发展水平与能力维度分层	物理社会可交互动态具身平台评测

型可揭示其与儿童年龄阶段对应，并挖掘大模型能力局限性，比如在具身任务中的薄弱点。

2 通用测试

2.1 人类发展启发的通用阶段测试

首先，本文针对 AGI 测试的第一阶段——通用能力阶段展开构建。通用能力阶段大致对应人类 0~6 岁，指智能体与世界互动所需要具备的基础必备能力，保证智能体可以正常进入到人类社会进行进一步的学习和成长。因此，通用能力往往是反映智能体是否“完备”以及是否具备从零到一、举一反三的学习和适应能力。基于这个预期，在通用能力的测试中，可以参照发展心理学中对于儿童发展阶段的定义，将 AGI 与人类不同年龄阶段的能力水平做出对应。通过这种方法，可以实现以人类年龄阶段作为衡量标尺，对通用能力发展水平进行定量评估。此阶段的 AGI 测试旨在回答智能体所对应的人类年龄。

已有大量研究提出人类儿童的认知能力发展理论和不同年龄的能力里程碑。比如，伊丽莎白·斯佩尔克（Elizabeth Spelke）等人提出，人类的心智由独立的、可以分离的核心知识系统构成（Spelke & Kinzler, 2007），灵活的技巧和信念都是基于核心知识系统逐步形成的，得到了人类婴儿研究和动物研究的支持。同时，在发展心理学中，社会基本公认的儿童出生后的发展阶段（大致范围）主要包括婴幼儿期（从出生到 3 岁）、学前期（3~6 岁）、儿童中期（6~12 岁）、青少年期（12~20 岁）和成年早期（20~40 岁）等。美国疾病控制与预防中心（Centers for Disease Control and Prevention，CDC）列举了明确的人类儿童发展里程碑（milestone），包括大多数儿童（75% 或更多）在特定年龄时可以做到的事情（Scharf et al., 2016; Sheldrick et al., 2019）。

然而，并非所有人类儿童的能力和任务都适用于人工智能的衡量标准和测试。在本研究中，我们基于通智测试，提出了百余项适合人工智能测试的综合任务，并展开了对应人类年龄的实验，旨在为 GSA 的通用能力阶段选取合理任务集合。

2.1.1 实验设计与被试

为了研究不同年龄段儿童的行为能力，研究选取了 124 种行为并向监护人发放问卷（附录请见 https://github.com/TongTest-series/GSA）。题目列表见附录。举例而言，一些代表性题目包括：考察空间关系—“您的孩子能够正确描述物体之间的相对位置关系，比如花瓶在桌子上面，椅子在桌子旁边”；考察社会认知（Fan et al., 2022）和“察言观色”—“您的孩子想要获得某样东西，会从您的表情察觉您是否同意，比如当看到您的表情不悦时，会理解为您不同意”；考察价值系统中的助人行为—“想象当您抱着一堆书走向书柜，地上有一滩水，但您看不到，您的孩子看到后，会主动提醒您小心地上的水。”对于每一道题目，评分范围为 -2 到 2（-2：完全不能，-1：几乎不能，0：部分能，1：基本能，2：完全能）。除了 124 个任务评分，问卷中设置两道注意力检测题目，如“请在此题选择完全无法完成”。在两道题目上任意一道答错，问卷作答不计入有效数据。

研究使用见数线上平台，招募了 343 名儿童家长进行线上作答，孩子的年龄范围在 36~72 月之间。

其中有效作答333份，样本量参考大模型测试（Wang et al., 2024）。无效作答的剔除标准包括：注意力检测题目未通过、孩子年龄不在有效范围、答题时间过短。完成有效问卷的家长每人获得11.5元红包奖励。

2.1.2 数据处理与分析

数据处理时，首先按月（1岁以下）或年（1岁以上）计算各行为得分的平均值（范围[-2, 2]），高于0分表示具备一定能力，高于1分表示基本掌握，接近2分则表明能较好完成。同时，研究通过计算不同年龄段的得分差值确定行为能力的增长率（公式1），并综合增长率和原始得分筛选出各年龄段的特征行为。

$$Rate(a,b) = \frac{M(a) - M(b)}{\mathrm{Time}(a) - \mathrm{Time}(b)} \quad （公式1）$$

同时，我们检验了测试集的信度以及探索性因素分析（EFA）。根据Comrey和Lee（1992）的建议，样本数量超过300可满足EFA的要求。本研究共纳入333份有效问卷，测试集的内部一致性信度（Cronbach' s α）为.965，表明测试集整体具有较高的信度。EFA结果显示，其中KMO值为.88，Bartlett球形检验为 $\chi^2(7626)= 18570.86$，$p < .001$，数据适合进行因素分析。在未限制因子个数的情况下，采用平行检验提取因子，分析结果支持3因子结构分别为：认知推理（如猜数游戏，找规律，类别数量计算），运动控制（如动作词理解，开灯，跳房子）和社会互动（如主动安慰，眼里有活，情感对话）。

2.1.3 结果分析与讨论

ANOVA结果显示，任务完成度在年龄上具有显著的主效应，$F(2, 369) = 3.88, p < .05$，$\eta_p^2 = .02$, 95%CI= [.00, 1.00], 可以间接支持任务表现在不同年龄阶段之间的显著差异。如表2所示，不同年龄段（3~4、4~5和5~6岁）中能力增长率排名前十的任务有所不同。在3~4岁阶段，增长率数值整体相比于后续年龄阶段数值更小，可能说明本研究中所涉及任务大多在4~6岁达到最快的发展阶段。其中，3~4岁阶段增长率大的任务包括“集体意识”、“开门”、“占座”等任务，体现了在基本运动、认知和情绪理解方面的增长。在4~5岁阶段，找规律的增长率最高（0.5），其次是“搭积木”、“类别数量计算”、“洗碗”等。整体上这一阶段增长最快的任务表明4~5岁儿童在认知和运动领域的发展较为突出，开始快速掌握大量日常生活中通用的综合任务。最后，进入5~6岁阶段，“猜数游戏”、“意图理解”成为增长最快的任务，显示儿童在复杂社交、团队协作和情绪管理方面进步显著。同时，“自主行为回溯”（.52）、“物品归属”（.48）等任务的增长率其次，代表了这一阶段儿童对于许多涉及高级认知价值（如自主性、行为回溯、归属权）的任务产生快速增长。总体来看，儿童从3岁到6岁的发展呈现从基础能力向高阶社交与实践能力的过渡，是社交情感与综合技能快速发展的关键期，也代表了通用能力整体的快速发展。这些结果与已有的大量研究相吻合，在人类婴儿研究和动物研究中，

表2 基于人类数据的3~6岁发展代表性任务

3~4 岁			4~5 岁			5~6 岁		
任务	*M*±*SD*	增长率	任务	*M*±*SD*	增长率	任务	*M*±*SD*	增长率
集体意识	.93±.98	.46	找规律	.67±1.12	.5	猜数游戏	1.11±.81	.71
开门	1.14±.95	.43	搭积木	.71±1.04	.48	意图理解	1.00±.94	.52
占座	.64±1.08	.42	类别数量计算	.47±1.09	.37	自主行为回溯	.68±.88	.52
隐藏玩具	1.24±.74	.4	洗碗	-.04±1.15	.34	物品归属	1.09±.94	.48
纸盒写字	.18±1.23	.38	摆字母	.84±1.00	.34	安全知识	.61±.92	.47
区分颠倒的符号含义	.50±1.16	.36	拼正方形	1.25±.85	.33	拧瓶盖	1.14±1.05	.46
向前跳	1.34±.74	.34	主动帮助	1.16±.79	.33	通过对话理解心情	1.02±.73	.42
找出按钮的功能	.66±1.12	.33	积木特征归类	.61±1.00	.32	理解对称	1.30±.85	.4
避障	1.29±.63	.33	开灯	1.63±.67	.32	按需求画画	.93±.76	.4
穿反衣服	.61±.93	.32	混乱与整理	.14±1.10	.32	制止不良行为	1.34±.61	.4

灵活的技巧和信念都是基于核心知识系统逐步形成，美国 CDC 列举了明确的人类儿童发展里程碑，包括大多数儿童（75% 或更多）在特定年龄时可以做到的事情（Scharf et al., 2016; Sheldrick et al., 2019），我们的结果与其表格一致：三岁时儿童能注意到其他孩子并加入他们玩耍（社交集体意识），四五岁时儿童在认知上有大的发展如规律的发现、续写故事，五六岁时对数字更加敏感，如数到 10。

2.2 面向通用阶段的通用人工智能具身测试

2.2.1 测试方案设计

鉴于当前 AGI 发展仍在初期阶段，暂时没有智能体可以完全具备通用阶段的所有能力，因此本文的测试以 GSA- 通用阶段为主，对于专项和应用能力测试，仅作必要讨论。

构建通用能力测试，需要依赖 DEPSI 虚拟环境，并将人类真实生活中遇到的各类真实任务在虚拟环境中构建出来。虚拟环境的构建主要分为两个部分：第一为物理与社会模拟，其中物理环境主要负责支持通用智能体在不断探索中逐渐发现并学习物理常识，而社会模拟则可以支持通用智能体体验并学习社会智能现象，第二为人类用户接口，测试系统需要将人类接入的原因是支持人类进行相同任务的具身测试，实现类人化大模型测试。

2.2.2 测试系统构建

为了支持通用能力测试，构建复杂多样的测试任务，我们搭建了物理逼真的虚拟仿真环境，并引入了具身环境中的社会模拟技术，从而构建动态具身的物理与社会交互环境。

在实验中，我们采用了 Unreal Engine 5（UE5）图形引擎作为核心的物理仿真引擎，在其上搭建适合智能体测试的交互式虚拟环境。UE5 提供了高度物理逼真的场景模拟，既可以模拟各种形状的刚性物体，也可以模拟水流、火焰、布料等复杂对象。在本实验中，以虚拟仿真环境中的室内房间作为基本实验环境，内置卧室、厨房、客厅、餐厅、卫生间等人类生活居所必备的空间。各个虚拟房间内设置有丰富的可交互物体，例如可以打开的衣柜、冰箱、微波炉等，可以移动的桌子、椅子、玩具等以及具备特定交互功能的饮水机、电风扇、电灯开关等。基于这些设置，待测智能体可以与环境产生丰富的具身交互，从而实现测试目标。

在社会模拟方面，本实验开发了从家庭环境到社区环境的微型社会模拟系统，支持生成丰富多样的社会智能（social intelligence）测试场景。该社会模拟环境基于已构建的物理逼真的虚拟场景，并加入了不同的社会角色，从而模拟各类人员之间可能发生的社会事件。由于通用能力测试主要针对的是对标人类儿童的智能体，因此我们将测试环境中的待测智能体设置为一个儿童的形象，并可以接收算法模型的驱动指令。在这个儿童角色之外，还设置了不同的家庭成员（如爸爸、妈妈、爷爷、奶奶、兄弟姐妹等）以及可能遇到的社会人员（如学校老师、快递员、医生等）。在这样的设置之上，测试系统可以模拟各类复杂的社会交互事件，测试智能体的社会智能水平。

2.2.3 测试任务

（1）受人类认知启发的大模型综合测试任务集

根据表2的人类儿童的通用能力测试任务集合，我们可以筛选出每个年龄段的经典测试任务。通过测试智能体是否能够通过这些测试任务，可以回答智能体能够对应到人类发展阶段的具体年龄。但是，机器智能和人类智能存在着本质差异，导致大部分人类儿童测试任务并不能直接作为基准任务来测试机器智能。例如，儿童在某个年龄段能够数 10 范围内的数字，但是这对于机器来说却是十分简单。其他的比如直接考察记忆、计算等的任务，对于机器来说并不合适，因为这些能力反而是机器的强项，且无法反映智能发展阶段。因此，我们从人类儿童通用能力测试任务集合中过滤掉不适合作为大模型测试的任务，形成表 3 所示的典型通用任务集合，用于机器测试。 进一步，本研究尝试为所选任务进行文献支撑。由于本研究中所选区的视觉任务大多为综合任务，包含了多维度的能力（如挑选礼物任务包含了社会认知、情感理解、偏好理解等），而经典心理学领域研究大多针对单一能力维度进行研究，本研究基于一些前期文献从不同角度支持当前核心测试任务的对应年龄标准，并将此信息列入附录供参考围绕选择任务。

（2）测试对象与测试方法

我们评测了四款多模态大模型：GPT-4o-2024-08-06（OpenAI 推出的多模态 GPT-4 模型，版本 2024-08-06，后文简称为 GPT-4o）、Claude-3.5-sonnet-20240620（Anthropic 推出的多模态大模型，后文简称为 Claude3.5）、Qwen-vl-max-2025-04-08（阿里巴巴通义千问推出的多模态视觉语言模型，后文简称为 Qwen）以及 Doubao-vision-pro-32k

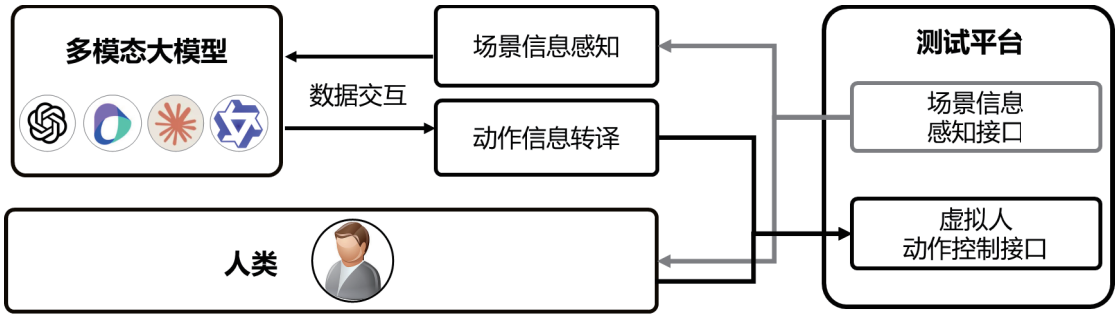


图 1 多模态大模型接入测试平台的方式

（字节跳动豆包推出的多模态大模型，后文简称为 Doubao）。由于多模态大模型本质上还不是一个具身形态的智能体，因此需要再原生大模型之外开发一些必备组件，使得大模型也能以具身形式接入到测试平台之中。针对选取的 4 款大模型产品，我们开发了场景信息感知和动作信息转译模块，联通了大模型与测试平台之间的数据通路，实现了将大模型接入测试的目的（图 1）。

本文选取了 8 个测试任务（图 2），分别是整理行李箱、理解按钮功能、整理桌面、挑选礼物、物品分类计数、搭建积木、拼图和打扫房间，已涉及人类 3~6 岁的各个阶段。如果能够通过该测试，则表明待测试智能体在通用任务表现上与人类相应阶段儿童水平相当。针对这 8 个任务，测试平台提供了对应的测试场景。

测试场景采用了统一的输入和输出接口。对于待测智能体来说，输入数据为任务指令的语言描述、智能体第一视角图像、场景内物体信息的语言描述、可执行动作接口函数的语言描述，输出数据为智能体决策的语言描述、可执行动作序列。部分任务可能仅需要上述输入数据和输出数据的子集。在测试开展过程中，待测智能体需要解析输入数据中的文本和图像信息，并输出代表了模型思考结果的语言信息（即语言反馈或动作描述）。具体的任务设置信息详见表 3。

在测试中，每个模型都会针对每一项任务进行多次重复测试（每个任务重复次数固定为 10 次，每次任务在 0~10 分之间对任务完成度或回答正确率进行量化），然后累加全部测试得分形成百分制的评分结果。换句话讲，每项任务的满分 100 分由 10 道满分为 10 分的随机题目构成，同一个任务内的不同

表 3 通用任务列表

对应年龄段	任务名称	任务描述	输入数据	输出数据	评分标准
3~6	整理行李箱	给定室内房间环境，智能体需要根据指令将合适的物品找到并放置到行李箱中	任务指令描述、第一视角图像、物体信息、可执行动作接口	动作序列	是否将正确的物品放进行李箱
	理解按钮功能	给定室内房间环境，包含各类可以操作的按钮，如灯的开关、空调遥控器、电风扇开关灯，智能体需要通过交互来回答问题	任务指令描述、第一视角图像、物体信息、可执行动作接口	语言	是否答对了按钮功能
	整理桌面	给定杂乱的桌面布局，智能体需要按要求整理收纳成合理的目标状态	任务指令描述、第一视角图像、物体信息、可执行动作接口	动作序列	是否符合桌面整理要求
	挑选礼物	给定任务背景，比如去旅游或去走亲戚，智能体需要根据实际需要选择合理的礼物，符合一般的社会认知	任务指令描述、第一视角图像	语言	是否选对了最佳礼物
	物品分类计数	在室内房间随机生成多样化的不同数量的物品，智能体需要在房间内漫游并统计出各个物品类别的数量	任务指令描述、第一视角图像	语言	是否回答正确
	搭建积木	根据语言描述或者 2D 图片线索，搭建符合要求的积木形状	任务指令描述、第一视角图像、物体信息、可执行动作接口	动作序列	是否搭建成功正确形状
	拼图	给定目标效果图，以及可以自由操作的印有碎片图案的拼图块，智能体需要对拼图块进行移动排放，还原出与目标效果图相同的状态	任务指令描述、第一视角图像、物体信息、可执行动作接口	动作序列	是否还原图像成功
	打扫房间	给定杂乱的室内房间状态，智能体需要对杂乱的物品进行清理收纳	任务指令描述、第一视角图像、物体信息、可执行动作接口	动作序列	是否执行了合理的整理策略

注：表格中将每个任务的具体内容、输入输出数据形式以及评分策略进行了说明。

图 2 对应人类 3 至 6 岁阶段的测试任务

题目设置采用完全随机化配置，保证不同题目的难度分布一致。不同任务之间的类型差异较大，因此各个测试任务的具体评分方法有所不同。但是，需要严格保证不同模型参加同一项任务测试时的评分方法与打分尺度一致。

2.2.4 测试结果和讨论

（1）大模型智能体在与 6 岁儿童的具身任务对比中表现不佳

所选任务为儿童的典型能力任务，如图 3 数据显示，6 岁儿童在相应任务中，能力评价得分的均值均超过 50 分。采用单样本 t 检验，在相同任务条件下，将基于多模态大模型的智能体的得分与 6 岁儿童的任务表现进行比较。结果发现，基于多模态大模型的智能体在具身任务中的表现不佳（图 3）。部分智能体能够在“整理行李箱”任务中超过 50 分（$ps > .05$），与 6 岁儿童水平相似，在“理解按钮功能”、“整理桌面”、“挑选礼物”和“物品分类计数”测试中表现出更低的水平（即在 30 分上下，$ps < .001$），而在“搭建积木”、“拼图”和“打扫房间”测试中甚至进一步降低（$ps < .001$）。这一结果表明，当前多模态大模型在需要空间感知、环境交互和动态调整的具身任务中存在明显局限。原因可能在于大模型主要依赖文本数据训练，缺乏对三维空间的直接感知能力以及物理环境的实时反馈机制，导致其无法有效模拟人类儿童在具身任务中展现的综合任务能力。

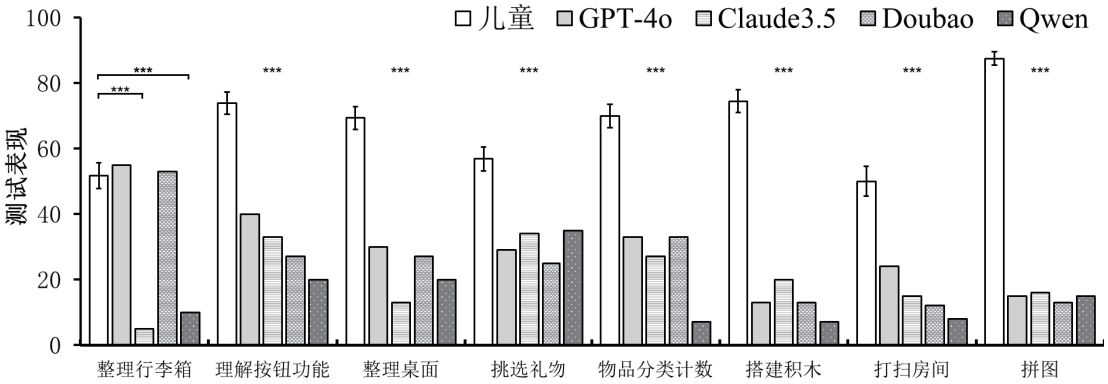
（2）大模型智能体之间的对比显示 GPT-4o 具备最强综合能力

在测试的多个维度中，GPT-4o（OpenAI, 2024）表现出最强的综合能力，在 8 项任务中有 5 项取得最高分数。需要说明的是，该发现为观察性结果，而非统计推断。GPT-4o 作为 OpenAI 推出的端到端统一模型架构的大模型（OpenAI, 2024），将文本、视觉等多种模态信息统一训练，重点提升了跨模态数据的感知处理能力，在环境信息综合感知方面具有超过同类模型的表现，因此其在需要跨领域知识和复杂推理的任务中表现更为出色，在具身交互任务上获得较好的成绩。然而，其在动作输出相关测试的得分仍较低，反映出具身操作能力的普遍瓶颈。上述结果表明，即使是基于性能最优的大模型构建的具身智能体，如果不进行针对性的二次设计与开发，其在环境交互与动态执行方面仍然远达不到预期。

（3）大模型智能体能力受限于具身操作，在偏认知类任务中表现更优

智能体在偏认知类任务（如语言、推理）中的表现明显优于具身操作类任务（如空间感知和动作序列输出）。例如，“挑选礼物”测试和“理解按钮功能”测试只需要智能体输出相对简单的动作序列，主要测试了智能体认知和推理能力，几乎所有智能体在这两个测试中都取得比其他测试更优秀的得分。而在“打扫房间”、“搭建积木”、“拼图”这些需要复杂且长程动作序列输出的任务中，智能体的表现均出现了不同程度的弱化。

这一结果表明，当前阶段大模型的核心设计聚焦于自然语言处理和逻辑推理，其训练数据主要为文本、图片等静态数据，而非可以实时交互与反馈的环境，导致当前大模型架构语言模块强于视觉模块，缺乏对真实世界物理规律和动态环境的建模能力。此外，具身任务需要实时感知和动作反馈，而



模型	整理行李箱	理解按钮功能	整理桌面	挑选礼物	物品分类计数	搭建积木	打扫房间	拼图	平均得分
儿童	52	74	69	57	70	74	50	88	66.7
GPT-4o	**55**	**40**	**30**	29	**33**	13	**24**	15	28
Claude3.5	5	33	13	34	27	**20**	15	**16**	18.4
Doubao	53	27	27	25	**33**	13	12	13	23.8
Qwen	10	20	20	**35**	7	7	8	15	13.4

图 3 大模型多任务测试结果以及与 6 岁儿童能力对比。误差轴为 6 岁儿童任务评分标准误

注：星号标注儿童与大模型任务表现上差异的显著性，星号下连线表示具有显著差异的儿童 – 大模型对比组合，星号下无线条的任务代表相应任务上儿童表现显著高于所有大模型表现，$^{***}p < .001$。

当前模型普遍缺乏与物理世界直接交互的传感器和执行器，导致其在需要动手操作的任务中表现不佳。在不针对所测任务进行专门训练的条件下，基于大模型直接构建出的智能体对大部分具身交互任务均表现不佳。综上，本次测试揭示了大模型在具身任务中的局限性及其能力的异质性。未来研究需进一步探索多模态学习和环境交互机制，以提升模型在具身任务中的表现。

3 讨论

3.1 潜在路线图：基于通用测试的人工智能发展路线图

人类作为通用智能体的代表，人类的发展阶段和心理学理论为 AGI 的发展和测试阶段提供了重要启示。正如经典的智力理论，斯皮尔曼（Spearman, 1923）的二因素理论提出，人类能力由一般因素（G 因素）和特殊因素（S 因素）组成，其中 S 因素代表特定领域的能力。这一理论直接支持了 G 阶段与 S 阶段的区分——通用能力是基础，而专项能力则在此基础上针对特定领域深化发展。专项能力阶段 (S 阶段) 关注 AGI 在特定领域的专长发展，这一构想与发展心理学中的能力分化理论密切相关。

AGI 的最终目标是应用到人类生活的各个方面。因此，评估 AGI 的发展水平应结合其发展阶段进行全面论述。本研究提出一种潜在路线图，AGI 的能力可以主要分为三个阶段，分别为针对通用能力的测试、针对专项能力的测试以及针对行业应用能力的测试。首先，通用能力反映的是智能体在通用能力发展阶段所达到的水平，以人类儿童发展阶段为基础参照。G 阶段的测试可以回答通用智能体达到了人类多少岁的水平。第二，专项能力反映的是智能体在专项任务解决中所表现的智能水平，S 阶段的测试回答通用智能体在专项领域中的能力突破和潜力，为最后一阶段的应用做铺垫。最后，行业应用能力反映的是智能体在具备了通用能力和专项能力后，在投入到人类社会后所能产生的应用效果。A 阶段的测试回答通用智能体在人类社会的融入度，包括安全性、稳定性等一系列维度。一个未经行业应用能力评测和验证的智能体，无法符合人类对 AGI 服务于人类社会、造福人类社会的预期。

具体而言，通用能力已在前文进行相对详细的阐述，而第二阶段的专项能力测试是在不考虑智能体整体性的前提下，以完成某项专门的目标而设立的能力测试。举例来说，擅长下围棋的 AlphaGo 模型就是一个十分典型的专项能力模型（Silver et al., 2016），它在围棋对战中具有巨大的优势，甚至可以轻松战胜人类中的专业选手，但是如果问其他的领域知识或者任务（如运动、社交等领域），就无能为力了。然而，这并不是说专项能力是不重要的，人类生活与生产中的诸多实际场景，恰恰是需要能够针对性地解决问题就可以，而不需要这些模型都变成通才。十分典型的例子有抽象推理能力测试、几何图形推

理能力测试、结构化推理能力测试、社会智能之“察言观色”能力测试、价值取向测试等。

最后，GSA 潜在路线的第三阶段，也是 AGI 的最终发展阶段，可能是行业应用。行业应用能力是智能体作为一个整体进入人类社会后的必要测试环境。为保障安全，在具体实践过程中，不能等智能体完全进入社会后再进行行业应用能力测试，而应该在投入使用前，根据所涉及行业应用的基本需求，对智能体的应用能力进行全面测试。AGI 的第三阶段的行业应用能力测试致力于覆盖无形服务与实体机器人两大领域，形成虚实结合的全场景评估体系。在无形服务层面，重点测试法律咨询的合规推理、音乐创作的跨模态生成、家庭对话的情感共情、手机助手的多轮交互以及数字人（如文旅导览）的拟人化服务能力，通过知识准确性、创意性和用户体验等维度建立评估标准。而在机器人领域，则聚焦蔚来汽车组装的精密操作、机器狗的自主导航、消防抗灾的应急决策等物理场景，结合传感器数据、任务完成效率和环境适应性等指标进行量化评测。既验证核心智能的通用性，又评估垂直场景的落地效能，最终形成可复用的 AGI 行业能力基准。

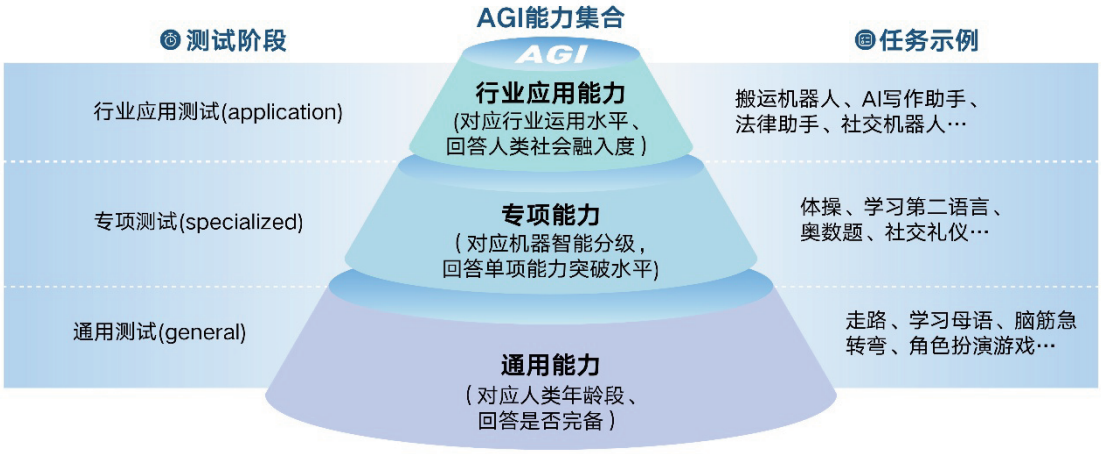


图 4　通用 – 专项 – 行业应用三阶段 AGI 发展和测试示意图

综上，在此潜在路线图中，通用能力、专项能力、行业应用能力评测联合构建起 AGI 评测的完整评测体系，是从实用角度出发形成的具象化技术形态和可操作实践路径（图 4）。本潜在测试路线也为 AGI 的价值对齐机制与安全伦理建设提供了重要启示——通过可解释、可验证的 GSA 测试，能够促进 AGI 与人类价值观的协同，增强其行为透明性与社会责任，从而提升公众信任，助力 AGI 安全、可控地融入人类社会。

3.2　讨论与局限性

本文基于人类认知发展过程启发，展示了对于 AGI 的具身通用能力测试实例，并提出针对衡量 AGI 的“通用 - 专项 - 应用”的三阶段 GSA 评估框架作为潜在路线图，联合构建起 AGI 评测的完整评测体系。与领域内其他知名评级框架相比。GSA 框架更强调能力发展的横向维度与纵向落地的结合，不仅涵盖从通用认知到专业能力的梯度（G-S），还新增了行业应用（A）这一社会价值验证层，弥补了 AI 理论在技术与社会衔接上的模糊性。

本研究仍存在一些局限性。首先，本文在工程实践上以通用任务为主，尚未扩展至对于专项和应用任务的实际测量，仅提出潜在路线图。未来可基于具体数据集的测试，进一步完善通用 - 专项 - 应用的三阶段理论测试以及对于通用人工智能的 GSA-Index 量化方案。其次，本研究基于具身家庭环境构建了针对人类儿童 3~6 岁的代表性任务集，对于儿童的多维度能力有相对全面的表征，但其完备性有待进一步验证。此外，本研究在评估方法上采用了统一的评分标准，并由研究人员根据该标准对大模型完成任务的表现进行手动评分。考虑到研究设计，本研究采取了“10 分 × 10 次”的评分方案，以保证一定的数据量并降低偶然误差。然而，本研究在统计分析中仅报告了均值和标准误，尚未对评分分布及方差进行全面刻画，也未对评分体系的信度与不同任务间的难度差异作深入量化评估。这在一定程度上限制了对评分体系稳定性和解释力的充分判断。未来研究可考虑引入项目反应理论（item response theory, IRT）或 Rasch 模型等量化方法进一

步优化 GSA 指数体系，同时基于儿童真实表现对任务难度进行标定，从而实现大模型能力与儿童发展阶段之间更为精准的映射。最后，本研究对于儿童的能力评价以监护人的主观判断为主，未来可进一步引入客观的行为观察实验，如在实验室环境下儿童观察以及视频标注，将进一步客观验证儿童测试的合理性。

4 结论

本文基于人类智能的认知发展规律，提出了一种“通用-专项-应用”（general-specialized-applicable，GSA）三阶段的 AGI 测试框架。首先，通用阶段回答 AGI 对应的人类年龄，从多模态的能力和价值维度展开测试。其次，专项阶段强调专项领域的自主学习和能力提升，回答 AGI 在专项任务上的能力突破和上限。最后，应用阶段关注 AGI 在真实环境和行业应用中的泛化能力，以验证其是否能高效、安全地融入人类社会，服务于人。本测试框架旨在为AGI的评估提供受人类认知发展规律启发的自动化、客观、系统的标准与测试方法，并推动其向更融入人类社会的高阶智能迈进。

致谢：感谢张驰、李佳琪、郑子隆、牛力兴、范丽凤在测试中的贡献，感谢赵士云、卢宇洁、刘明远在通用任务测试中的贡献，感谢朱爱菊、谢卢彬、韩佳衡在数据收集方面的贡献，傅雨秋、周尚博在数据分析中的贡献，以及陈珍在作图中的贡献。

参考文献

丁贵广 , 陈辉 , 王澳 , 杨帆 , 熊翊哲 , 梁伊雯 . (2024). 视觉深度学习模型压缩加速综述 . *智能系统学报 , 19*(5), 1072–1081.

Antol, S., Agrawal, A., Lu, J. S., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). *VQA: Visual question answering.* Proceedings of the 2015 IEEE international conference on computer vision (pp. 2425–2433), Santiago, Chile, IEEE.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). *Language models are few–shot learners.* Proceedings of the 34th international conference on neural information processing systems (pp. 1877–1901), Vancouver, BC, Canada: Curran Associates Inc..

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Psychology Press.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei–Fei, L. (2009). *ImageNet: A large–scale hierarchical image database*. 2009 IEEE Conference on Computer Vision And Pattern Recognition (pp. 248–255), Miami, FL, USA.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., & Florence, P. (2023). PaLM–E: An embodied multimodal language model. *arXiv*.

Fan, L. F., Xu, M. J., Cao, Z. H., Zhu, Y. X., & Zhu, S. C. (2022). Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research, 1*(2), 144–160.

Feldman, R. S. (2006). *Development across the life span*. Pearson Education New Zealand.

He, Z. Q., Liu, Y. S., Zheng, J. S., Li, X. J., Yao, J. G., Qin, B. W., & Yang, X. (2024). *FlagEvalMM: A flexible framework for comprehensive multimodal model evaluation.* https://github.com/flageval–baai/FlagEvalMM

Kirillov, A., Mintun, E., Ravi, N., Mao, H. Z., Rolland, C., Gustafson, L., Xiao, T., & Girshick, R. (2023). *Segment anything*. Proceedings of the 2023 IEEE/CVF international conference on computer vision (pp. 3992–4003), Paris, France.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. (2014). *Microsoft COCO: Common objects in context.* In Computer Vision–ECCV 2014: 13th European Conference (pp. 740–755), Zurich, Switzerland.

Liu, Y., Duan, H. D., Zhang, Y. H., Li, B., Zhang, S. Y., Zhao, W. B., & Lin, D. H. (2024). *MMBench: Is your multi–modal model an all–around player?* In 18th European Conference on Computer Vision (pp. 216–233), Milan, Italy.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination–free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242.

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., & Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. *Nature, 642*(8067), 451–457.

Morris, M. R., Sohl–Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., & Legg, S. (2024). *Position: Levels of AGI for operationalizing progress on the path to AGI.* In Proceedings of the 41st International Conference on Machine Learning (pp. 36308–36321),Vienna, Austria.

OpenAI. (2024). GPT–4 technical report. *arXiv*.

Oppy, G., & Dowe, D. (2003). The turing test. *Stanford Encyclopedia, 2021,* 1–26.

Peng, Y. J., Han, J. H., Zhang, Z. L., Fan, L. F., Liu, T. Y., Qi, S. Y., & Zhu, S. C. (2024). The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering, 34,* 12–22.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Scharf, R. J., Scharf, G. J., & Stroustrup, A. (2016). Developmental milestones. *Pediatrics in Review, 37*(1), 25–38.

Sheldrick, R. C., Schlichting, L. E., Berger, B., Clyne, A., Ni, P. S., Perrin, E. C., & Vivier, P. M. (2019). Establishing new norms for developmental milestones. *Pediatrics, 144*(6), Article e20190374.

Shu, T. M., Peng, Y. J., Zhu, S. C., & Lu, H. J. (2021). A unified psychological space for human perception of physical and social events. *Cognitive Psychology, 128,* Article 101398.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van Den Driessche, G., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition.* Macmillan.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1), 89–96.

Wang, J. Q., Zhang, C. H., Li, J. P., Ma, Y. X., Niu, L. X., Han, J. H., & Fan, L. F. (2024). *Evaluating and modeling social intelligence: A comparative study of human and AI capabilities.* In Proceedings of the Annual Meeting of the Cognitive Science Society, Rotterdam, Netherlands.

Human Intelligence-Inspired Testing for the Developmental Stages of Artificial General Intelligence: From General to Applicable

Peng Yujia[1,2,3], He Xinyi[2,4], Xie Hongzhao[2], Xiao Xizhi[5], Wang Yuxi[1,2], Zhu Songchun[2,3,4], Zhang Zhenliang[2]

([1] School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, 100871)([2]State Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence, Beijing, 100080)([3] Institute for Artificial Intelligence, Peking University, Beijing, 100871)([4]School of Intelligence Science and Technology, Peking University, Beijing, 100871)([5] Yuanpei College, Peking University, Beijing, 100871）

Abstract The rapid advancement of artificial intelligence (AI) is profoundly reshaping society, presenting unprecedented opportunities for the development of Artificial General Intelligence (AGI). While generative pre-trained models (e.g., the GPT series) demonstrate remarkable generalization in specialized domains, they remain narrow AI systems, still facing gaps in achieving AGI. Our previous work proposed that AGI demands adaptability to dynamic, embodied environments (Dynamic Embodied Physical and Social Interactive, DEPSI), characterized by infinite-task handling, autonomous task generation, and value-driven decision-making. However, translating abstract AGI definitions into practical testing frameworks remains a critical challenge. Here, we proposed a human intelligence-inspired developmental testing framework for AGI to assess its progression from general to applied capabilities.

First, in the general stage, AGI is expected to demonstrate cross-domain foundational cognitive abilities, such as common sense reasoning and adaptive learning, analogous to early childhood intelligence development (ages 0–6). By collecting and analyzing human developmental data, this study establishes a series of general tests to measure an AI system's "cognitive age." Specifically, eight representative tasks were selected and implemented in a UE5-based virtual environment, including organizing a suitcase, tidying a desk, and solving puzzles, which cover the cognitive and motor skills expected of 5–6-year-olds. The environment features realistic domestic settings (e.g., kitchens and bedrooms) with interactive objects (e.g., appliances and furniture) and social agents (e.g., family members and teachers) to assess both physical reasoning and social intelligence. A human-user interface, incorporating VR and motion tracking, enables direct comparisons between AI and human performance. Four multimodal large models (GPT-4o, Claude-3.5, Qwen, and Doubao) were tested after being equipped with perception and action modules to interface with the virtual environment. Each task was repeated 10–15 times, with average scores computed for evaluation.

Key findings reveal critical limitations in current AI systems. A common limitation lies in their constrained embodied performance. While models approached baseline competence (30/100) in simpler tasks, such as understanding button functions, they struggled in complex, physically interactive tasks, including puzzle-solving and room cleaning. GPT-4o emerged as the strongest performer, leading in five tasks, but still exhibited significant shortcomings in motor coordination. Similarly, the models excelled in language-heavy tasks (e.g., selecting gifts) but underperformed in spatial and sequential-action tasks. This reflects their training bias toward static text/image data rather than dynamic, embodied interaction. The study concludes that current large language models, without specialized adaptation, lack the embodied intelligence required for human-like task execution. Future advancements must prioritize real-time sensory feedback, interactive learning, and improved physical simulation to bridge this gap.

Building upon this foundation of general abilities, we introduce a three-phase AGI testing framework: General-Specialized-Applicable (GSA). The specialized phase emphasizes autonomous learning and skill refinement in specific domains (e.g., Go, mathematics), enabling AI to tackle complex problem-solving and knowledge integration, much like human adolescents mastering specialized subjects. It is noteworthy that general and specialized capabilities are not mutually exclusive but exhibit a synergistic, spiral progression in AGI development. General capabilities form the foundational "operating system" of an agent, enabling cross-domain knowledge transfer and adaptive learning. Conversely, advancements in specialized domains refine this system through novel cognitive patterns and problem-solving methods. This bidirectional reinforcement creates a "general-specialized" spiral trajectory of AGI development. Looking back, traditional AI approaches often bypass general capabilities, focusing narrowly on specialized tasks (e.g., chess). To address this, we advocate a "layered development, dynamic balance" strategy: first achieving threshold general competence, then cultivating prioritized specialized skills while establishing feedback mechanisms to generalize domain insights. This approach prevents premature specialization ("ability silos") while ensuring practical utility, enabling continuous breakthroughs in both generality and expertise.

Finally, the applicable phase evaluates AGI's generalization ability in real-world environments and industrial applications (e.g., robotics, autonomous driving), verifying whether it can seamlessly integrate into human society and serve practical needs.

Overall, the GSA framework aims to provide a potential systematic, human development-inspired standard for AGI evaluation, guiding its development toward intelligence that can sufficiently coexist with and benefit humanity. The GSA framework not only proposed a standardized AGI assessment but may also fostered trust by ensuring alignment with human-centric values and practical applicability, which may advance AGI toward safe and meaningful social integration.

Key words artificial intelligence, artificial general intelligence, cognitive development, artificial intelligence evaluation, embodied AI