



人工智能系统测试与 TMMi®

#TMMi 成功实践

中文发布版 1.0

由 TMMi®基金会中国分会负责翻译

编辑：ERIK VAN VEENENDAAL

www.tmmi.org

TMMi 旨在独立于所应用的生命周期或被测试的系统类型。本文档解释了在开展人工智能系统测试时，如何有效应用 TMMi 测试改进模型并发挥其价值。文中逐一论述了 TMMi2 级各过程域及其目标，文中阐述了这些与测试人工智能系统的关系，以及这些实践通常是什么样的。若某项实践对人工智能系统测试无需执行，因为它没有附加价值，文中也会明确指出。本文档后续版本还将对 TMMi 更高级别的过程域进行探讨。TMMi3 级将论述人工智能系统特有的质量特性测试，例如灵活性、自适应能力、自主决策能力及偏见。需要说明的是，本文档并非完整的人工智能系统测试大纲，而仅旨在说明在人工智能系统测试中应如何解读 TMMi 目标，并为进一步学习提供思路与方向。《人工智能系统测试与 TMMi》文档亦不能替代完整的 TMMi 模型，使用本文档时，应始终结合 TMMi 参考模型文档配套使用。

©TMMi®基金会 2011 - 25

版权所有。未经 TMMi®基金会事先许可，不得以任何形式或通过任何方式全部或部分出借、出售、转让、复制或传播本出版物的任何部分，除非以相关许可文件中所述的方式。如果相关许可文件的条款允许任何形式的复制，则本通知必须以任何此类副本的形式复制。

我们有理由相信构成商标的词语已被指定为商标。然而，无论是否存在此类指定，都不应被视为影响任何商标的法律地位。

TMMi®是 TMMi®基金会（英国）的注册商标。

贡献者

Amanda Dean (Australia)

Mattijs Kemmink (The Netherlands)

Vipul Kocher (India)

Poonam Jain (India)

Qin Liu (China)

Srinivas Padmanabhuni (India)

Çiçek Tuna (Germany)

Erik van Veenendaal (The Netherlands)

Kelly Ye (Australia)

修订历史

本节汇总了本文档不同版本间的主要修订内容。

本节内容仅供参考。

版本号	修订记录
1.0	“人工智能系统测试与 TMMi” 文档初始版本。
EN1.0_CN1.0	本文档 “人工智能系统测试与 TMMi” 中文版本地化完成

参与中文本地化的 TMMi®专家

主要内容	翻译专家	评审专家	QA 专家
英文封面-16（封面到 2.2 节结束）	张婷	张楠婕	叶岚
英文 P16-P21（2.3 节开始到 References 结束）	张楠婕	张婷	

目录

1	基于 TMMi 的人工智能系统测试	6
1.1	人工智能	6
1.2	测试成熟度模型集成 (TMMi)	6
1.3	人工智能系统测试与 TMMi	7
1.3.1	人工智能系统测试 vs. 使用人工智能辅助进行测试	7
1.3.2	人工智能系统测试面临的挑战	8
1.3.3	目的与方法	8
2	TMMi 2 级: 已管理级	10
2.1	过程域 2.1 测试方针与策略 (TPS)	10
2.1.1	SG1 建立测试方针	10
2.1.2	SG2 建立测试策略	11
2.1.3	SG3 建立测试绩效指标	12
2.2	过程域 2.2 测试策划 (TPL)	12
2.2.1	SG1 执行风险评估	13
2.2.2	SG2 建立测试方法	14
2.2.3	SG3 建立测试估算	15
2.2.4	SG4 开发测试计划	16
2.2.5	SG5 获得测试计划的承诺	16
2.3	过程域 2.3 测试监督与控制 (TMC)	17
2.3.1	SG1 根据计划监督测试进度	17
2.3.2	SG2 根据计划与期望监督产品质量	17
2.3.3	SG3 管理纠正措施直至结束	17
2.4	过程域 2.4 测试设计与执行 (TDE)	18
2.4.1	SG1 使用测试设计技术执行测试分析与设计	18
2.4.2	SG2 进行测试实施	19
2.4.3	SG3 进行测试执行	20
2.4.4	SG4 管理测试事件至结束	20
2.5	过程域 2.5 测试环境 (TEV)	20
2.5.1	SG1 开发测试环境需求	21
2.5.2	SG2 执行测试环境的实施	21
2.5.3	SG3 管理和控制测试环境	21
	参考文献	22

1 基于 TMMi 的人工智能系统测试

1.1 人工智能

人工智能（AI）处于当今最具变革性的技术革命前沿。作为计算机科学的一个分支，它致力于赋予机器思考、学习与自适应能力，而这些能力以往仅为人类智能所独有。凭借重塑产业、社会与日常生活的巨大潜力，人工智能已成为众多前沿创新背后的核心驱动力。究其本质，人工智能旨在研发能够执行通常需人类智能方可完成的各类任务的计算机系统。这些任务既包括简单运算，也涵盖复杂问题求解、模式识别、语言理解与智能决策。人工智能系统可处理海量数据、识别规律，并基于经验持续优化自身性能，在一个日益由数据洞察驱动的世界中，成为极具价值的工具。

机器学习（ML）作为人工智能的重要分支，是人工智能系统的核心组成部分。它通过在海量数据集上训练算法，使算法能够自主识别模式并做出预测，而无需人为逐条编程实现。这种能力正在改变医疗保健、金融、市场营销、制造等诸多领域，推动实现数据驱动决策与重复性任务的自动化。然而，人工智能在具备巨大潜力的同时，也伴随着一系列独特的挑战。伦理问题、算法偏见、数据质量、隐私安全、可解释性与透明度，都是人工智能系统必须解决的问题。人工智能也引发了关于其对就业与社会发展带来了深远影响，引发重要思考。

1.2 测试成熟度模型集成 (TMMi)

TMMi 框架由 TMMi 基金会开发，作为测试过程改进的指南和参考框架，旨在解决对测试经理、测试工程师、开发人员和软件质量专业人员至关重要的问题。在 TMMi 中，测试被广泛定义为涵盖所有与软件产品质量相关的活动。

TMMi 使用成熟度级别的概念进行过程评估和改进。此外，还明确了过程领域、目标和实践。应用 TMMi 成熟度标准可以优化测试过程，实践证明其对提升产品质量、测试工程效率及缩短周期时间具有显著成效。TMMi 的开发是为了支持组织评估和改进其测试过程。

TMMi 具有用于过程改进的阶段型架构，它包含一系列阶段或级别，组织可以通过它们使组织的测试过程从临时的和未管理的，演变为已管理、已定义、已测量并最终达到持续改进状态，也称为优化。实现每个级别可确保为下一更高级别奠定足够的改进基础。

TMMi 内部结构具备丰富的测试实践，可以通过系统化学习与应用，来支持逐步改进的质量测试过程。TMMi 中有五个级别，规定了成熟度层级和测试过程改进的演进路径。每个级别都有一组过程域，组织需要实现这些过程域以达到该级别的成熟度。TMMi 每个成熟度级别的过程域如图 1 所示。

TMMi 的一个主要基本原则是它作为通用模型，适用于各类生存周期模型与环境。TMMi 定义的大多数目标和实践已被证明适用于顺序和迭代生存周期模型，也包括敏捷与 DevOps 模型。然而，在模型的最底层，很多子实践及示例会因采用的生存周期模型存在差异。请注意，在 TMMi 中，只有目标是强制性的，实践并非强制性的。

TMMi 可在 TMMi 基金会官网免费获取，该模型已被译为西班牙语、法语、中文等多种语言，同时也有正式出版的书籍版本。



图 1: TMMi 成熟度级别与过程域

1.3 人工智能系统测试与 TMMi

1.3.1 人工智能系统测试 vs. 使用人工智能辅助进行测试

在探讨人工智能系统测试与 TMMi 的关联，以及如何在人工智能系统测试场景中应用 TMMi 之前，首先需要明确“人工智能系统测试”与“使用人工智能进行测试”之间的重要区别。二者是人工智能技术在软件测试领域应用中两个截然不同的概念。

人工智能系统测试指对包含人工智能组件的软件应用或系统开展性能、功能及质量评估与验证的过程。其核心目标是保障软件系统内的人工智能算法、模型及功能按预期运行，输出合理结果，并满足预期的质量标准。人工智能系统测试涵盖多项测试活动，包括输入数据测试、模型测试、集成测试、功能测试、性能测试、安全测试、合规性测试，以及针对偏见、公平性等伦理层面的测试。

使用人工智能进行测试指运用人工智能技术与工具来增强并优化软件测试流程本身（及相关活动）。其核心目标是借助人工智能技术支撑并（部分）实现各类测试活动自动化，实现数据驱动决策，从而提升测试效率、准确性与覆盖度。人工智能可应用于测试估算、缺陷预测、测试分析、测试用例生成、测试数据生成、测试执行及测试结果分析等场景，还可基于风险与历史数据确定测试用例优先级排序。

综上所述，“使用人工智能进行测试”旨在保障人工智能驱动软件的质量，而“使用人工智能进行测试”则是借助人工智能提升测试流程本身的效率与效果。本文仅聚焦于人工智能系统测试及其与 TMMi 的关

联，以及 TMMi 所能发挥的支撑作用。使用人工智能进行测试相关内容，或将纳入下一版 TMMi（TMMi2.0）的测试实践范畴予以考量。

1.3.2 人工智能系统测试面临的挑战

应如何对人工智能系统开展测试？尤其是当前最主流的机器学习类系统，会在测试过程中动态改变自身行为。此外，测试工作通常需要明确判定测试通过或失败，但对于人工智能系统而言，“测试通过”究竟意味着什么？在推理逻辑较为简单的场景下，或许可以人工判断测试结果是否合格；但面对海量数据时，人工判定则难以实现。人工智能系统往往缺乏清晰完整的需求规格说明，其预期行为可能随时间发生变化，甚至除系统输出外，我们无法明确何为正确行为。

人工智能系统测试面临的主要挑战包括：

- 人工智能系统多用于处理复杂问题，其可能场景的数量极为庞大（远超传统软件）。
- 数据本身存在偏见，且这类偏见极易被嵌入到人工智能系统中。
- 基于人工智能的系统可能出现模型漂移、数据漂移及概念漂移问题。
- 缺少可判定预期结果的测试结果参照物。
- 人工智能模型的性能指标与传统系统性能指标差异显著，对多数测试人员而言也是全新的。
- 传统测试技术往往不再适用，我们需要重新思考我们的测试技术，以及测试覆盖率的衡量方式。

当前，人工智能系统测试仍极具挑战性，尚无简单方案可全面解决上述问题。一个较为可行的学习切入点，是研究由 AiU United (AiU) 与国际软件测试资质认证委员会 (ISTQB®) 制定的认证大纲。总体而言，组织需要先具备一定的测试成熟度水平，才能有效应对人工智能系统测试带来的各类挑战。

1.3.3 目的与方法

本文旨在阐述将人工智能系统测试与 TMMi 测试过程改进模型相结合，是组织实现更高测试成熟度、并最终达成业务目标的可行路径。文中说明了如何在人工智能系统测试场景中有效应用 TMMi 模型，并提供了经实践验证的、可替代方案，以取代传统的测试方法，这些方案实施了 TMMi 实践，可用于人工智能系统测试。本文阐述了 TMMi 如何为开展人工智能系统测试的组织提供支持，无论组织当前成熟度水平如何，均可借助其提升成熟度与质量能力。例如，可通过提醒组织组织在规模扩张与项目压力增大时忽视这些关键测试实践。本文将逐一讲解各 TMMi 过程域及其特定目标，说明其与人工智能系统测试的对应关系，以及典型测试实践的表现形式。若某项 TMMi 实践对人工智能系统测试无附加价值、无需执行，文中也会明确说明。需要注意的是，本文仅识别并映射了人工智能系统测试实践与 TMMi 实践及目标的对应关系，并未展开详细描述。目前市场上已有如 [Smith] 等优质资料，可提供人工智能系统测试实践的更多详细内容。

由于人工智能系统测试并非简单地测试另一类系统，相关实践调整主要体现在与测试工程活动相关的过程域中，例如测试策划、测试设计、非功能测试等。在 TMMi 模型中，测试工程类过程域主要集中在 2 级（已管理级）和 3 级（已定义级），因此本文仅围绕这两个等级，探讨其在人工智能系统测试中的应用。

本文不可脱离原始 TMMi 模型单独使用，旨在为开展人工智能系统测试领域进行测试过程改进的人员提供指导，帮助其理解各项 TMMi 目标与实践的内涵。本文是对原始 TMMi 模型的补充，应作为辅助文档使用。

2 TMMi 2 级：已管理级

2.1 过程域 2.1 测试方针与策略（TPS）

测试方针与策略过程域的目的是开发并建立测试方针，以及组织级或项目群级的测试策略，其中明确定义各类测试活动（如测试类型与测试级别）。为了衡量测试绩效，体现测试活动的价值并识别改进空间，需引入测试绩效指标。

2.1.1 SG1 建立测试方针

任何组织在启动测试改进项目时，都应首先定义测试方针。测试方针定义了组织的整体测试目的、目标和有关测试和测试专业人员的战略导向。测试方针必须与组织整体业务（质量）方针保持一致，这一点至关重要。测试改进工作应由清晰的业务目标驱动，而这些目标需相应纳入测试（改进）方针并形成文档。测试方针对于在组织内的所有利益相关方之间就测试及其目标达成共识来说是很有必要的。这种共识是使整个组织内的测试（流程改进）活动保持一致所必需的。请注意，测试目的不应孤立存在，其源自建立可用软件、保障产品质量这一高级别目标。

无论所测试的系统类型如何，上述内容适用于任何组织。首先，理解人工智能及其所带来的风险，以及需要采用不同的测试方法是非常重要的。一些主题，如伦理、偏见、数据和法规，可能需要被引入或得到更明确的关注。此外，人工智能系统特有的非功能特性，如透明性、可解释性和可说明性，也需要被引入。因此，测试目的与目标可能需要重新审视，关于测试和测试流程定义的基本观点也同样需要重新审视。虽然其中大部分内容与非人工智能系统大体相同，但也会存在差异。例如，目标可能包含降低人工智能系统引入风险的相关说明；流程定义中可能需要说明人工智能系统对相同输入可能返回不同结果这一特点，并增设集成前的模型测试阶段。而这样的阶段在传统系统中通常并不存在。

伦理层面需要考虑测试人员如何在基于人工智能的系统中使用数据，但这也与系统自身的伦理行为有关。当人工智能系统在现实世界中做出的决策可能对生命与自由产生重大影响时，伦理抉择就会面临风险且至关重要。人工智能领域的伦理问题会存在差异，具体取决于系统是从自身运行结果中持续学习，还是仅依据指定的训练数据集进行学习。

数据将带来诸如偏见、隐私和安全之类的问题，这些是在测试方针中需要考虑的重要事项。完善的数据需求、对群体的理解，正变得至关重要。例如，人脸识别系统需采用能够代表受影响群体的代表性数据进行训练和测试，而非仅使用易于获取的数据集。面向多国市场的产品意味着测试所需的数据范围将远大于仅从本地群体获取的数据。

除了在诸如安全关键系统等领域外，质量通常不是信息系统**受监管**的方面。然而，随着越来越多基于人工智能的系统对个人做出决策，而人们又无法充分理解这些决策，各类监管规范正在推进或已生效，从法律上对组织的质量提出约束。

2.1.2 SG2 建立测试策略

测试策略遵循测试方针，并作为项目中测试活动的起点。测试策略通常在整个组织范围或项目群范围内定义。典型的测试策略基于高层次的产品风险评估，并会包含对将要执行的测试类型和测试级别的描述。仅规定每个项目需执行集成测试与验收测试是远远不够的，还需明确定义集成测试与验收测试的意图、通常如何执行及其主要目标。经验表明，制定并遵循测试策略，可减少各类测试活动间的重叠，提升测试过程效率。同时，由于不同测试类型与测试级别的测试目标和方法保持一致，能有效减少测试遗漏，使测试过程更高效，进而提升产品质量水平。

测试策略对于人工智能系统测试同样是一份至关重要的文档。对许多组织而言，人工智能系统测试仍属于较新的领域，因此，一份通用的组织级测试策略能够为各项目的人工智能测试工作提供指导，避免重复劳动。测试策略还能确保所有参与测试的人员理解整体情况。对于传统系统，测试策略围绕待执行的测试级别与测试类型来建立，并对其进行相互关联的识别与定义。而在人工智能系统测试中，诸多内容均有所不同。

人工智能测试策略首先区分两个主要阶段：离线测试和在线测试。

在**离线阶段**期间，人工智能模型被调优、验证并使用指标（例如准确率或精确率）进行测试，以评估其目标的完成情况。通常除了对机器学习模型开展功能行为测试外，还会单独对模型执行部分适用的非功能测试，例如训练速度、预测速度、计算资源占用、适应性及透明性等。除模型测试外，离线阶段还新增了一类测试级别：输入数据测试。输入数据测试旨在确保系统用于训练和预测的数据具备最高质量。

在**在线阶段**期间，将对经过训练的模型与系统其余部分（包括其他人工智能组件、非人工智能组件及其运行环境）的集成情况进行测试。此阶段会开展功能测试和非功能测试（例如性能测试）。在线阶段可采用传统测试中适用的测试级别与测试类型，来建立测试策略。

- 组件测试适用于任何非人工智能模型组件，例如用户界面和通信组件。
- 开展组件集成测试，以确保系统组件（包括人工智能组件与非人工智能组件）之间正确交互。
- 开展系统测试，以确保由各类组件（人工智能与非人工智能）集成而成的完整系统，在高度贴近生产环境的测试环境中，从功能和非功能角度均能按预期运行。根据系统不同，该测试可采用在预期运行环境中开展现场试验的形式。
- 验收测试用于判断完整系统是否可被客户接受。但部分方面可能需要额外关注。若系统旨在替代人工操作，则可能需要将系统行为与人类行为进行对比，并对照既定的性能标准进行比较。
- 一个同样适用于人工智能系统测试的测试级别，是持续监控系统的性能。对于人工智能系统，经常有必要在实际使用中持续地审查和重新审视系统的质量。在生产环境中监控质量的一个重要原因是偏差检测。有时候在测试环境中很难判定偏差，这是因为训练数据和模型在实际环境中遇到的数据之间可能存在差异，而持续监控有助于解决这一问题。持续监控的另一个重要原因是漂移。对于基于人工智能的系统，模型漂移、数据漂移和概念漂移均为需要长期关注与监控的因素。持

续监控模型性能指标对于及早发现漂移迹象至关重要。漂移指的是这样一种情况：正常工作的模型可能会随着时间推移而性能下降，原因是输入特征和输出之间的关系发生了变化。模型在使用过程中，应用场景可能发生变化，导致模型表现偏离预期目标。

注意，也可能存在这样的情况：人工智能系统不是由组织自身开发的，而是组织应用了外部开发的人工智能系统（作为整体系统配置的一部分），例如由微软、苹果或谷歌开发的系统。这显然会对其人工智能测试策略产生影响，使该策略更接近商用现货（COTS）软件的测试策略，重点聚焦于与其他系统的集成、以及根据组织需求进行定制与系统配置。

2.1.3 SG3 建立测试绩效指标

测试改进的业务目标，如测试方针中所定义的，需要被转化为一系列关键测试绩效指标。测试方针及其配套绩效指标明确了方向，以及一种传递测试绩效预期水平和达成水平的方式。绩效指标必须向利益相关方体现测试与测试过程改进的价值。由于过程改进投入需要管理层长期支持，因此定量衡量改进方案的收益以持续获得管理层动力是至关重要的。请注意，TMMi 这一特定目标是定义数量有限的测试绩效指标（如 2~3 项），并非建立并实施完整的测量体系，而是确定一组核心指标，用以反映测试价值如何随着时间推移以及在不同的交付环境中的变化。

如果人工智能系统的测试是测试过程改进方案的一部分，并且是定期开展的，那么就需要为人工智能系统测试制定专门的测试绩效指标，以展示测试改进方案的结果。由于人工智能系统的本质以及相应测试目标存在差异，传统的绩效指标，例如缺陷

检出率，通常并不适用。人工智能系统的测试绩效指标，将用于衡量经训练或测试后的模型输出结果的准确性与可靠性。需要注意，人工智能系统的测试绩效指标取决于学习类型（无监督或监督）和模型类型（聚类、关联、分类或回归）。例如，对于监督分类型人工智能系统，可采用混淆矩阵（见图 1），并据此计算准确率和/或精确率指标

作为绩效依据。对于监督回归人工智能系

统，均方误差（MSE）可以用作绩效指标[ISTQB]。MSE 是实际值与预测值之间差值平方的平均值。MSE 的值始终为正数，数值越接近 0 表明回归模型越好。更多可用作测试绩效指标的度量示例可以在 AiU 认证测试人员人工智能大纲[AiU]和 ISTQB 认证测试人员人工智能测试大纲[ISTQB]中找到。

		预测结果	
		阴性(N) -	阳性(P) +
实际结果	阴性 -	真阴性(TN)	假阳性(FP) I 型错误
	阳性 +	假阴性(FN) II 型错误	真阳性(TP)

图 1: 混淆矩阵

2.2 过程域 2.2 测试策划（TPL）

测试策划的目的，是基于已识别的风险和已定义的测试策略定义一套测试方法，并为执行和管理测试活动建立和维护有充分依据的计划。

请注意，成功的测试策划的关键在于前期思考（“活动”），而不是定义相关的测试计划（“文档”）。

2.2.1 SG1 执行风险评估

穷尽测试是不可能的，对于基于人工智能的系统测试也是如此，需要始终做出选择并设定优先级。因此该 TMMi 目标同样适用于人工智能系统测试。测试方法将基于对人工智能系统的风险分析来制定，同时包含传统测试以及更专业的测试，以覆盖人工智能组件及人工智能系统特有的属性。

评估基于人工智能系统的产品风险的典型技术包括头脑风暴、专家访谈和检查表。

- 头脑风暴由一次或多次会议推动，其中沟通是关键。在一个项目里，通常无人拥有人工智能系统如何工作的完整图景。具备多元专业技能（尤其是人工智能相关技能）的跨学科团队，对过程质量至关重要，避免项目中某一环节占据主导。团队的目标是针对人工智能系统及其运行的所有方面，提供广泛的视角和看法。
- 专家访谈提供有价值的见解和专业知识，加强风险评估和识别过程。通过使用访谈技术，组织可以直接从人工智能专家那里收集信息，从而更全面地理解潜在风险及其影响。
- 用于风险识别的检查表汇集了早期人工智能项目的经验。理想情况下，清单应聚焦于人工智能系统的通用风险领域。其目的是启发思考那些易被遗忘、但频繁出现的产品风险，这些风险可以被纳入考量，同时提示人工智能系统可能面临的问题类型。

人工智能系统的产品风险示例：

- 输入数据质量过低
- 训练数据量不足
- 运行数据链路存在问题
- 用于模型开发的机器学习工作流程不完全正确
- 机器学习框架、算法、模型、模型参数及 / 或超参数选择错误
- 机器学习功能在生产环境中未达到预期性能
- 用户对机器学习功能的实际表现不满意

人工智能系统的产品风险示例（续）：

- 自学习系统无法提供用户预期的服务
- 用户因无法理解系统决策逻辑而产生不满
- 可用模型因输入特征与输出间关系发生变化而性能逐渐退化（通常称为概念漂移）
- 预训练模型存在未知或未文档化的偏差与缺陷
- 幻觉：可能源于训练数据质量不佳、模型倾向于用看似合理的内容填补信息缺口，或对提示词理解偏差；由于 AI 往往以高置信度输出虚构内容，这类问题会成为重大挑战。

2.2.2 SG2 建立测试方法

定义测试方法，以缓解已识别并划分优先级的产品风险。该测试方法也受已定义的组织测试策略指导（见 2.1.2 节）。应该让参与人员明确工作内容及如何执行。和测试策略一样，详细的测试方法区分了两个主要阶段：应用系统的离线测试和在线测试。需为待执行的每个测试级别和测试类型定义一种测试方法。

离线阶段从数据准备、预处理和输入数据测试开始（初始测试级别）。需要执行的活动是数据操作、数据过滤和数据预处理。数据预处理包含数据插补（即处理缺失值）、通过数据可视化获得整体图景，处理异常点和离群点，进行相关性分析和降维。随后对人工智能模型进行评估与分析，这一过程还包括将可用输入数据集划分为训练集、验证集与测试集。训练集包含用来训练模型的数据。验证集被机器学习算法用于评估训练是否有效。训练后阶段则使用测试集对机器学习模型进行评估。

在线阶段对非人工智能组件、以及人工智能与非人工智能部分之间的交互进行测试。这通常需要对人工智能应用的架构进行详细的研究和理解。最终将对完全集成后的人工智能系统开展测试。除了探索性测试之外，蜕变测试是识别和定义人工智能系统测试场景的一种常用技术。

入口准则，也被称为就绪定义（特定实践 2.3）与出口准则，也被称为完成定义（特定实践 2.4），同样是所定义的测试方法的一部分。某一测试阶段的入口准则，通常为上一测试阶段出口准则的满足情况，同时也与测试准备工作完成和测试环境可用有关。根据测试阶段不同，部分出口准则可沿用传统系统的相关标准，例如测试覆盖率、未闭合缺陷等。同时还有一些可作为定义出口准则依据的度量指标，其与传统系统所使用的出口准则截然不同。

下表显示了各类人工智能模型使用的出口准则（度量指标）示例[AiU]：

学习类型	模型类型	采用的出口准则
无监督	聚类	惯性值 调整兰德指数
	关联	支持度 置信度 提升度
监督	分类	准确率 精确率 召回率 / 灵敏度 特异度 F1 分数
	回归	均方根误差 R 平方误差

表 1: 按人工智能模型类型划分的出口准则与度量指标

特定实践 2.5 “定义暂停与恢复准则”是否适用于人工智能系统测试，取决于所采用的生命周期与方法。当采用更传统的方法时，它是适用的。在敏捷模式下，它可能相关性较低。在敏捷中，相关准则一般会在项目状态与进度会议中隐性执行，但是为每个测试阶段提前明确定义这些准则略显冗余。当出现可能对项目进展构成潜在或实际威胁的阻塞问题时，团队会在会议中进行讨论，并决定需采取哪些行动（如有必要）来解决问题。敏捷项目管理常规流程作为该特定实践的一种替代实践。

2.2.3 SG3 建立测试估算

测试工作量估算是为达成项目或团队测试目标所需开展的测试相关工作总量进行预估。重要的是向干系人明确该估算是基于一系列假设条件得出的，且始终会存在估算误差。对于某类系统（如人工智能系统）测试经验不足的组织而言，这一点尤为突出。

一般而言，有两种可用的估算方法：

基于比例的估算。该方法属于基于度量的技术，通过收集组织内部过往项目的数据，可推导出适用于同类项目的“标准”比例。组织自有项目的比例（例如取自历史数据）通常是估算过程中使用的最佳来源。随后可利用这些标准比例，对新项目的测试工作量进行估算。由于人工智能系统测试对行业和大多数组织来说相对较新，可用的度量指标很少（甚至没有），因此该估算方法并非人工智能系统测试的首选方式。

基于专家的估算，例如宽带德尔菲法。该方法由“专家”基于经验进行估算。基于工作分解结构，每位专家独立地对单个任务的工作量进行估算。收集结果后，如果存在超出约定边界范围的偏差，专家们会讨论各自当前的估算值。规划扑克是宽带德尔菲法的一种变体，常用于敏捷软件开发。由于缺乏可用的度量数据，该方法成为人工智能系统测试工作量与成本估算的首选方式。参与估算的人员可包括项目或团队内部成员，以及组织内曾参与其他人工智能项目的人员。按照流程专家将给出最优估算值，并同步记录

用于估算测试工作量和成本的所有假设。宽带德尔菲法估算会议通常提供以下输入：基于已定义测试方案需开发的测试工作产品清单，以及与测试工作产品相关的待执行测试任务清单。

- 人工智能系统测试的工作量估算通常比传统软件测试更为复杂。以下是一些可以被考虑的人工智能特有因素示例：
- 数据复杂性：人工智能系统 / 应用的测试需要在数据验证、偏差检测和漂移监测方面投入额外工作量，以确保预测结果准确性、公平性。
- 人工智能模型行为：与传统软件不同，人工智能具备学习和演进特性，因此需要持续测试、可解释性校验与调优。
- 持续监控：人工智能需进行实时跟踪、性能检测与模型重训练，以维持性能表现。
- 法规合规性：人工智能测试必须遵循 GDPR《通用数据保护条例》、AI act《人工智能法案》及行业专项规定，确保其符合伦理和法律规范。

2.2.4 SG4 开发测试计划

开发测试计划时应始终考虑两大主要群体：（1）测试人员（需明确工作内容）；（2）干系人（需了解即将开展的测试工作、参与职责、资源承诺、相关风险等）。人工智能系统测试计划的详细程度，同样取决于所采用的生命周期模型。例如，更敏捷的工作方式意味着更轻量化的测试计划。

由于人工智能具备数据驱动、自适应、概率性等特征，其测试计划必须在传统软件测试基础上进行扩展。与传统应用不同，人工智能模型会学习、演进并做出概率性决策，人工智能模型会学习、演进并做出概率性决策，因此需要专业化的测试策略与针对性的测试计划。鉴于人工智能的复杂性与动态演进特性，测试必须是持续且自动化的，并通过流水线实现持续验证。

由于人工智能系统测试对许多组织而言仍较新颖，因此需要形成书面化测试计划，以确保测试人员明确工作内容，同时让干系人充分知情。测试计划可采用文档、演示文稿或思维导图等形式，只要能达成目的的即可。测试计划应记录前期测试策划活动的成果、经优先级排序的产品与项目风险清单，并详细说明离线阶段（数据输入测试、模型测试）的实施方法与活动，以及在线阶段的测试级别和测试类型。

需制定测试进度表。由于测试是开发的一个组成部分，通常不会有单独的测试进度表，而是一份同时包含测试活动的整体进度计划。对于顺序型生命周期，该计划可以采用详细的书面进度表形式；对于敏捷生命周期，则可采用带优先级的任务看板形式。

作为测试规划的一部分，需要制定一份计划来保证具备所需知识和技能的测试人员资源到位，以执行离线和 / 或在线测试。这当然可能具有挑战性，因为掌握此类测试所需专业知识与技能的测试人员往往属于稀缺资源。因此可能有必要在测试规划中定义所需的培训，以确保测试人员能够完成测试工作。

2.2.5 SG5 获得测试计划的承诺

最后，测试计划和已定义的测试方法需要得到干系人的同意。干系人必须理解并同意产品风险的优先级列表以及将要执行的测试缓解措施。例如，可以通过简短的演示并随后进行讨论的方式，使其理解并作出承诺，在讨论中说明产品风险、测试方法及其理由，以及所需预算。针对测试计划与测试方法的讨论，应结合待开发机器学习模型/人工智能系统的建设目的展开，确保与业务优先级保持一致。验收标准显然是需要与干系人充分沟通并达成一致的另一项重要内容。

若采用更偏敏捷的方式制定测试计划与测试方法，该过程本身即为团队协作活动，通常可能由人工智能测试专家（团队成员之一）主导。产品质量是团队的责任。因此，只要团队遵循正确流程，基于共同努力形成的测试方法与测试计划，团队成员自然会达成共识与遵守承诺。当然团队以外的其他干系人可能也需要对测试计划作出承诺。对他们而言，如前所述的演示和讨论环节依然至关重要，必须执行。

2.3 过程域 2.3 测试监督与控制（TMC）

测试监督与控制的目的在于了解测试进度与产品质量，以便当测试进度显著偏离计划或产品质量显著偏离预期时，能够采取适当的纠正措施。

2.3.1 SG1 根据计划监督测试进度

对于测试基于人工智能的系统而言，监控测试进度、开展进度评审以及报告测试进度，与测试传统（非人工智能）系统并无不同。监控测试进度的实施与执行方式，主要取决于所采用的生命周期模型。在顺序式 V 模型类型的生命周期中，测试进度监控通常更为正式，文档要求也更高；而在敏捷类型的生命周期中，测试进度监控通常非正式程度较低，借助任务板进行支持，文档要求也较低（另见[TMMiAgile]）。

2.3.2 SG2 根据计划与期望监督产品质量

在监督产品质量时，重要的是要理解测试基于人工智能系统与测试传统系统有所不同。这一差异也影响到需要监控的产品质量的哪些方面。在传统系统中，“测试”指的是诸如查找缺陷、执行测试和测量性能等多种过程；而在人工智能领域，“测试”则特指针对专用数据集对系统性能进行的统计评估。基本上，所监控和报告的退出准则的类型在本质上有不同（可参考第 2.2.2 节中的示例）。总之，产品质量监控主要是针对已识别的产品风险和已定义的退出准则进行的。缺陷监控则很可能针对系统的非人工智能部分进行。关于产品质量监控的实施方式，请参考上一段的内容。

人工智能系统产品质量监控的另一个不同之处在于，需要对系统的性能进行持续监控。对于人工智能系统，通常需要在实际使用过程中持续审查并重新评估系统的质量。正如前文所述，对于基于人工智能的系统，模型漂移、数据漂移和概念漂移是需要长期考虑和监控的因素。需要持续监控模型性能指标（例如准确率），以便及早发现漂移迹象。发布后仍需持续监控产品质量的其他重要原因还包括：防范对抗性攻击、数据投毒以及自学习演进行为等。

2.3.3 SG3 管理纠正措施直至结束

与监控测试进度（见上文）类似，管理纠正措施直至完成对于测试基于人工智能的系统而言，与测试传统（非人工智能）系统并无不同。然而，问题的类型很可能有所不同。测试人工智能系统时，预期会遇到的问题通常集中在输入数据质量方面。数据质量问题可能对项目产生重大影响，和/或需要较长时间才能修复。

管理纠正措施直至完成的具体实施与执行方式，主要取决于所采用的生命周期模型。在顺序式 V 模型类型的生命周期中，与进度相关的问题通常由测试经理或项目经理负责管理。而在敏捷类型的生命周期中，Scrum Master 通常负责处理进度阻碍事项。这类纠正措施的状态将展示在任务板上，并在每日站会上进行讨论。

2.4 过程域 2.4 测试设计与执行（TDE）

测试设计与执行的目的在于通过建立测试设计规格说明、运用测试设计技术、实施结构化的测试执行过程，以及管理测试事件直至关闭，从而提升测试设计与执行过程期间的测试过程的能力。

尽管目标陈述具有很高的概括性，但定义测试条件、测试用例并执行测试的过程却截然不同，具体如下所述。一旦模型生成（即完成训练、评估和调优），就应使用独立的测试数据集对其进行测试，以确保满足约定的功能性能标准（参见第 2.2.2 节）。

2.4.1 SG1 使用测试设计技术执行测试分析与设计

在测试人工智能系统时，“测试条件”和“测试设计”的概念与传统软件测试有着根本性的不同。其重点不在于推导显式的输入-输出成对，而在于构建和验证能够有效评估训练后模型的行为、准确性和鲁棒性的数据集。

其主要目标变为定义和设计能够代表人工智能模型运行场景范围的数据集。这包括识别用于训练、验证和测试的数据类型、数据源及数据质量。通常，从同一数据源中建立三个独立且具有统计代表性的数据集：

- 训练数据集 - 用于训练人工智能/机器学习模型。
- 验证数据集 - 用于评估和调优模型参数。
- 测试数据集（留出数据集） - 用于评估调优后模型的泛化性能。

在 TMMi 框架下，测试分析与设计阶段涉及识别并优先排序测试数据条件以及源自人工智能系统目标、需求和风险评估的评估指标。在此，测试依据不仅包括规格说明和设计文档，还包括数据收集策略、标注指南以及模型性能标准。

由于合适且无偏的数据通常有限，训练数据集和验证数据集常常从同一个组合数据集中衍生出来，而测试数据集必须保持独立。这种分离确保了模型评估的客观性，并防止数据泄露。数据集的划分没有通用

的固定比例，但一个典型的参考指南是训练集、验证集、测试集按 60:20:20 的比例分配。数据划分应采用随机方式，除非存在特定的分层或代表性约束（例如，数据集较小或存在关键子群体）。

人工智能测试设计侧重于选择和优先处理能够反映真实场景及边缘场景的测试用例与条件。这可能涉及：

- 设计对抗性测试或压力测试数据，以探测模型的薄弱环节；
- 在数据空间中使用等价类划分与边界值测试（例如，特征分布中的边缘值）；
- 应用组合测试技术，如配对测试，因为基于人工智能的系统所关注的参数数量可能极为庞大，尤其是当该系统使用大数据或与外部世界交互时；
- 蜕变测试，这是一种基于已通过的源测试用例来生成测试用例的技术，是另一种可用于人工智能系统测试的技术；
- 将偏差、公平性和可解释性测试纳入测试条件集合；

建立从数据源和模型需求到测试数据集和性能结果的可追溯性。

此目标的产出包括数据集规范、测试数据覆盖范围的文档记录，以及明确的验收标准（例如，准确率阈值、置信区间、偏差限度）。这些产物共同确保对人工智能模型的测试遵循结构化的方法。

2.4.2 SG2 进行测试实施

人工智能系统的测试实施涉及按照测试设计阶段所定义进行，包括获取、准备和最终确定测试数据与测试环境。人工智能测试的重点在于获取和整理用于验证训练后模型的性能、鲁棒性以及需求符合性的数据集。所获取的数据可能呈现多种形式（例如，数值型、分类型、图像型、表格型、文本型、时间序列、传感器数据、地理空间数据、视频或音频），并且在实际用于测试之前，通常需要进行大量的预处理工作。主要的预处理活动包括清洗、转换、增强和采样。

- 清洗：识别不正确、重复或不一致的数据，并进行纠正或删除。缺失数据可通过插补技术进行处理，同时采用匿名化或假名化技术来保护个人或敏感信息。异常值可根据其与人工智能系统运行场景的相关性，决定删除或保留。
- 转换：将数据格式或结构转换为适合被测模型的一致形式（例如，将分类数据转换为数值形式、归一化数值特征，或重新编码图像格式）。这些转换可确保测试执行过程的可比性和稳定性。
- 增强：为了提升覆盖范围和鲁棒性，可以通过生成新的样本或对样本进行修改来人为扩充数据集（例如，旋转图像、改写文本或注入噪声数据）。增强还可以包括旨在测试系统抵御扰动或攻击能力的对抗性样本。

- 采样：从大型数据集中选取具有代表性的子集，以确保测试的可行性，同时保持统计有效性。采样方法（随机采样、分层采样或基于风险的采样）应保留多样性以及在测试设计阶段所识别的关键特征。

所有预处理实施步骤都必须记录在案，并且能够追溯到数据集规范。实施后的测试数据、测试脚本以及环境配置共同构成了测试执行和模型评估所需的测试工作产品。

2.4.3 SG3 进行测试执行

人工智能系统的测试执行涉及运行已准备好的测试数据集和脚本，以收集有关训练后模型在既定验收标准下表现如何的证据。其目标是确保模型按预期运行，达到预期的性能水平，并在早期阶段确定的各种数据条件下可靠运行。只有在人工智能模型的训练、调优和验证工作完成并符合既定标准后，才能开始测试。这可以被视为开始测试执行的入口标准。

在测试人工智能系统时，执行测试通常意味着将测试数据集应用于经过训练和验证的模型，并记录其输出（预测结果、分类结果、分数或动作）。将这些输出与预期结果或参考结果（真实值）进行比较，并使用混淆矩阵以及定量性能指标（如准确率、精确率、召回率或 F1 分数，参见第 2.2.2 节）进行评估。

由于人工智能系统具有固有的随机性，单次测试执行可能会产生略有差异的结果。因此，测试应在受控条件下重复进行，并对结果进行统计分析（例如，性能指标的均值和方差）。这确保了结论建立在可复现且稳定的证据基础之上。

所有测试结果、异常情况以及与预期性能的偏差都必须记录在案，并能够追溯到相应的测试数据和所定义的条件。对测试事件进行分析，以确定其根源是数据质量问题、模型局限性还是实现缺陷。此特定目标的产出包括已执行的测试日志、捕获的结果、性能报告和事件报告。

2.4.4 SG4 管理测试事件至结束

对于与人工智能系统相关的测试事件和缺陷而言，将其管理至闭环的流程和做法并无不同。当然，在更细化的技术层面，所发现缺陷的类型以及缺陷分析等任务的执行方式等方面确实存在差异。例如，详细的数据质量分析可能会揭示与准确率下降、模型偏差以及模型受损相关的缺陷。

事件管理流程在很大程度上也取决于所采用的开发生命周期。在传统的顺序式项目与敏捷环境下的项目之间，事件记录及管理至闭环的方式存在显著差异 [TMMiAgile]。

2.5 过程域 2.5 测试环境 (TEV)

测试环境过程域的目的是，建立并维护一个适当的环境，使已可管理和已重复的方式执行测试成为可能。

通过测试环境过程域，建立并维护一个适当的测试环境（包括通用测试数据），使得能够以已管理且已重复的方式执行测试。基于人工智能的系统可应用于多种多样的运行环境，这意味着测试环境也同样具

有多样性。导致人工智能系统测试环境与传统系统测试环境存在差异的典型系统特征包括：自学习、自主性、多智能体协同以及大数据。此外，在测试基于人工智能的系统时，通常会使用虚拟测试环境。这通常使得测试危险、异常、极端以及时间密集型场景变得更加容易。虚拟测试环境还能提供远高于真实环境的测试可控性。通过虚拟测试环境对硬件进行仿真，使得系统能够在（模拟的）硬件组件上进行测试，而这些硬件组件在实际中可能无法获得。

请注意，数据准备通常是 AI/ML 人工智能/机器学习系统工作流程中的重要组成部分，这部分归属于测试设计和执行过程域（参见第 2.4 节）。数据原则上包含了将用于训练、验证和测试调优后模型的（测试）用例。因此，与传统系统相比，（测试）数据在基于人工智能的系统测试中扮演着非常重要的角色。

2.5.1 SG1 开发测试环境需求

基于人工智能系统的测试环境需求规范应在项目早期进行。这要求制定测试环境需求的人员对所开发的人工智能系统有详细的理解，因为系统的特征（上文列举了一些示例）将决定具体需要什么。需求规范需经过评审，以确保其正确性、适用性、可行性以及对“真实”运行环境的准确呈现。早期的需求规范有助于为获取和/或开发所需的测试环境及组件留出更多时间。这一点对于基于人工智能的系统尤为重要，因为其测试环境通常并非一目了然，且可能相当复杂。

2.5.2 SG2 执行测试环境的实施

通过此 TMMi 特定目标，测试环境需求得以实现，测试环境被提供用于人工智能模型的测试。在适当的情况下，对测试环境组件执行单元测试，并在实现阶段结束时执行可信度测试，以确定测试环境是否已准备好用于人工智能模型的测试。通用测试数据将在测试环境的过程中使用。

2.5.3 SG3 管理和控制测试环境

测试环境的管理和控制贯穿整个人工智能项目。项目团队通常需要自行搭建和维护测试环境，但也可能由团队外部的独立部门来完成。当这些活动由团队自行完成时，这当然意味着团队具备足够的技术知识来执行这些任务。同时这也意味着这些任务成为项目工作的一部分，例如，需要在规划会议中予以考虑、列入任务看板，并在出现问题时于每日站会中进行讨论。对于测试环境过程域，一个主要结论是：其特定目标和实践均适用，且本质上没有变化。不同的地方在于所需的测试环境类型、测试数据及其组件的差异。

参考文献

[AiU] 人工智能联盟（2019），《AiU 人工智能测试认证（CTAI）教学大纲 V1.0》，人工智能联盟

[ISTQB] 国际软件测试认证委员会（2021），《ISTQB 人工智能测试认证（CT-AI）教学大纲 V1.0》，ISTQB

[Smith] A.L. Smith、R. Black、J. Davenport、J. Olszeweska、J. Röβler 和 J. Wright（2022），《人工智能与软件测试-构建值得信赖的系统，BCS 学习与发展出版社

[TMMiAgile] E. van Veenendaal（编）（2020），《敏捷世界中的 TMMi V1.4》，TMMi 基金会