国际资本市场研报资讯+V: quanqiuzixun8

渊亭科技

2024
军事大模型评估体系白皮书

精简版

厦门渊亭信息科技有限公司

二〇二四年 五月

国际资本市场研报资讯+V: quanqiuzixun8

前言

数字化时代，人工智能技术正以前所未有的速度发展，其中大模型技术作为 AI 领域的核心技术之一，已经成为推动社会进步和产业创新的重要力量。大模型，以其强大的数据处理能力和深度学习能力，正在多个领域展现出其独特的价值和潜力，从自然语言处理到图像识别，从智能推荐到自动驾驶，大模型正在不断拓宽人工智能的能力边界。

伴随着大模型技术的快速发展，越来越多应用在军事情报、指挥控制、智能武器、无人系统等领域的军事大模型应运而生，助推军事智能化转型。其中，对大模型的真实质量的掌握，对指导研究方向、优化能力设计、提升应用效能有着重要意义。全面、客观、准确的评估特定大模型针对场景的实际能力，需要有一个完善的模型评估方法论，科学、客观的对大模型的各项能力进行定性、定量评估。

近年来，渊亭科技积极参与行业内大模型的各项能力评估建设，取得了突出成果。作为国内最早从事军事大模型建设的企业之一，渊亭科技凭借在军事智能化领域的深厚积累，编撰完成《军事大模型评估体系白皮书》。白皮书全面的整理了军事大模型能力评估方向的主流观点、关键要素，并重点阐述了针对典型维度进行系统化评估的最佳实践。预期能为行业内开展军事大模型的能力评估提供体系化的参考。

国际资本市场研报资讯+V: quanqiuzixun8

目录

国际资本市场研报资讯+V: quanqiuzixun8

国际资本市场研报资讯+V: quanqiuzixun8

1 背景

2022 年 11 月，OpenAI 发布了名为 ChatGPT 的人工智能应用，其以预训练大语言模型 GPT3.5 为基础，惊艳的自然语言交互效果，使得公众、行业对人工智能的能力预期大大提升，在国内外掀起了一股新的人工智能能力建设和应用浪潮。2024 年 2 月，OpenAI 公布了文生视频大模型 Sora、并提供若干样例视频，在行业内再一次引起巨大反响，以预训练大模型为核心的生成式人工智能技术，应用边界进一步拓宽。

在过去的几年中，中国的大模型技术和行业经历了快速的创新与发展。在通用大模型层面，百度、华为、阿里、讯飞、智谱、百川、月之暗面等企业根据自身的特点，采取开源、闭源等路线，持续聚焦底座模型效果和生态圈建设；在领域大模型层面，诸多传统企业和初创企业围绕 AI-Native、AI-Copilot 等概念各展所长，或基于自身业务引入大模型巩固和强化竞争优势，或针对新的方向进行细分市场探索、尝试创造新的商业模式；在场景应用层面，越来越多的“大模型目标用户”尝试整合私域数据，结合自身的战略布局，探索大模型技术的赋能方法，提升企业运营、生产制造、能力营销等方面的效率、质量。

能力被认可和推广的一项重要前提，是合理、可行的能力评估。通用大模型层面，目前评估以“榜单”为主要的体现形态，例如 MMLU、CEval、SuperCLUE、GSM8K、Humaneval 等，在不同榜单下各模型排名差异较大，原因在于评测数据、测试方法等还不够成熟、科学，且存在无意（例如训练数据集被污染）、恶意（例如主动将测试数据集纳入训练/微调过程）的“刷榜”现象。领域大模型层面，和通用大模型的能力评估现状相比，存在的问题更多，例如难以组织有效的领域测试数据集、使得大模型领域能力无从测起，没有系统的领域大模型生成和效果的测试方法、导致测试效果难被取信。目前国内已经有一些行业组织正在开展领域大模型相关的行标、国标建设。场景应用层

国际资本市场研报资讯+V: quanqiuzixun8

面的能力评估，由于需和上下游应用环境和信息系统深度对接，也有一些新的问题，例如模型生产和推理平台对企业既有基础设施的影响，模型和现场数据、系统之间的协同，模型在复杂使用环境下的安全保障等。

随着国防智能化建设的深入，军内很多机构都对大模型能力产生了浓厚的兴趣，军事大模型应用场景也非常丰富，如军事情报、指挥控制、智能武器、无人系统等领域。 军事大模型作为一类特殊的领域大模型，也有一些自身的能力评估特点。

军事领域的数据的机密性和敏感性众所周知。一方面，基础大模型很难在预训练/微调阶段注入足够的军事知识，军事认知必须在领域大模型构建过程中形成，使得领域大模型的军事常识能力评估显得愈发重要；另一方面，常识能力评估所需的数据集，也因为军事数据的特点，领域大模型的评测数据集构建更为困难，因此更难展开有效的领域大模型评估工作。

军事领域高对抗性的特点，使得军事大模型和常规领域大模型相比面临着更为严峻的安全挑战。例如，通用大模型面临的偏见，在军事领域可能升级为“认知战”手段、对方刻意对大模型能力进行干扰；又例如传统人工智能模型面临的对抗攻击、内容伪造、数据泄露问题，在军事大模型应用场景中需要得到更多的评估。

现代智能化战争一定是体系对抗，信息手段之间也需要有效配合，军事大模型的应用成效极大的体现在和平时、战时既有系统的协同。而军事信息化系统的特殊性，使得领域大模型的能力评估，只能在特定的区域、特定的时刻结合特定的数据开展，这就对能力评估的方法论和手段集提出了新的要求。例如如何快速的结合现场提供的数据构造测试数据集、如何快速的结合业务目标完成领域测试项准备等。

渊亭科技长期从事认知和决策智能领域研究和项目建设，参编了多项人工智能相关标准。近年来，也和一些行业主导标准化机构进行合作，推进围绕大模型的各项能力评估，例如大模型驱动的知识图谱、大模型运营能力等。基于以上背景，渊亭科技结合多年服务军事智能

国际资本市场研报资讯+V: quanqiuzixun8

化领域的行业认知，以及在军事大模型能力应用上的产品研发和项目实践经验，编撰完成本白皮书，希望研究成果能为社会各界参与军事大模型建设提供借鉴和参考。

2 总体架构

军事大模型评估体系围绕大模型在军事场景智能化能力表现进行科学合理的评估评价，实现大模型评估全流程，支撑军事大模型的部署应用、模型改进和决策制定，确保军事大模型在军事业务场景的应用价值。军事大模型评估体系如下图：

图 1 大模型评估体系架构

军事大模型评估体系主要包括军事大模型评估数据、军事大模型评估手段、军事大模型评估工具以及军事大模型评估指标等内容。

（1）军事大模型评估数据：军事大模型评估数据包括外部开源、主流评估以及用户领域等方面的评估数据集。

（2）军事大模型评估手段：军事大模型评估手段与评估场景及环境相适应，即满足人工评估模式，也支持基于规则、模型的自动化评估模式。

（3）军事大模型评估工具：军事大模型评估工具负责内外部数据管理、评估手段实现、军事大模型兼容以及融合评估指标标准等能力。

国际资本市场研报资讯+V: quanqiuzixun8

（4）军事大模型评估标准：军事大模型评估标准提供大模型的基础、架构、平台、应用以及安全能力多层次的评估，结合评估需求，灵活定义评估指标，实现评估标准场景自定义。

3 评估框架

评估指标体系是军事大模型基准测评体系框架的核心组成部分，围绕强敌研究、作战指挥、装备研制、训练管理和联勤保障等 5 类军事业务场景，针对军事信息系统高风险、高动态、强对抗的任务特点，构建一整套科学、客观、量化的评估指标，全面评估军事大模型在不同维度、领域和场景中的性能表现，为用户开展大模型选型提供标准化的测评参考，为大模型系统的上线运行提供可信的衡量标准，并为大模型的优化改进提供明确方向。

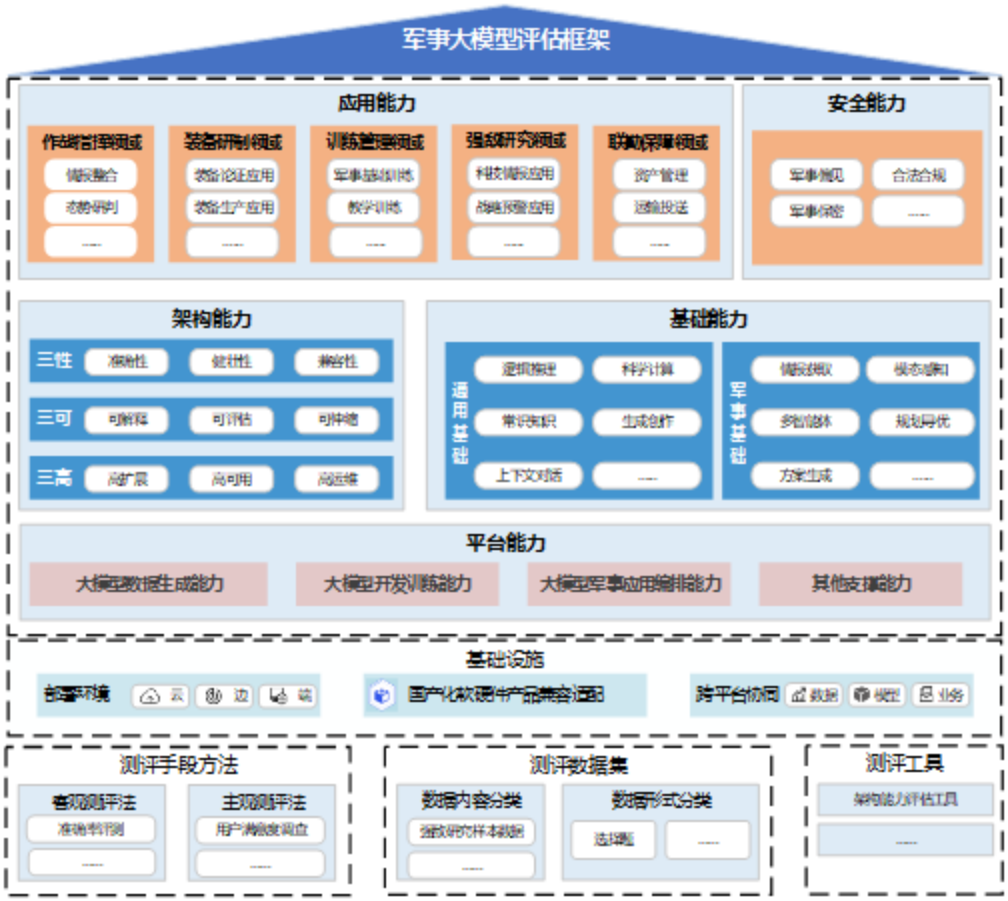

图 2 军事大模型评估框架

评估指标体系由架构能力、基础能力、平台能力、应用能力和安

国际资本市场研报资讯+V: quanqiuzixun8

全能力 5 个维度的评估指标构成。

(1)架构能力指标设计主要考核大模型体系化支撑军事应用的架构成熟程度；

(2)平台能力指标设计主要考量大模型系统的数据生成、开发训练、应用编排和其他支撑能力；

(3)基础能力指标设计主要覆盖大模型的通用基础能力和军事基础能力；

(4)应用能力指标设计侧重于从五大军事业务领域，评估大模型在实际军事业务场景中的表现；

(5)安全能力指标设计重点评价模型在军事偏见、合法合规和数据保密等方面的性能。

3.1 架构能力

军事大模型的架构能力是军事大模型系统整体性能的重要基石及确保大模型在军事领域准确高效处理数据、稳定可靠承载业务、安全可信落地应用的关键。主要体现在如下方面：

准确性：是衡量模型性能的关键因素，通常包括查准率（Precision）、查全率（Recall）、简洁性（Brevity）和结果置信度（Confidence Score）等指标项。

健壮性：是评估模型在面对复杂挑战时稳定性和可靠性的重要标准。旨在衡量模型在面对对抗样本时，能够保持正确预测的能力。

兼容性：是评估大模型对不同技术环境和组件的适应能力。包含对基座大模型接口和功能的适配性、对国产自主可控软硬件系统的兼容性以及第三方专业小模型、领域知识库和工具插件的兼容性。

可评估：涉及架构能力评估、基础能力评估和场景应用能力评估三个层面。架构能力评估关注模型设计和内部机制的合理性；基础能力评估则涉及模型在标准任务上的表现；场景应用能力评估考量模型在特定应用场景中的实用性和效果。

可解释：是确保模型的决策过程和结果对人类用户透明和可理解

国际资本市场研报资讯+V: quanqiuzixun8

的关键要素。主要包括推理过程可解释、推理结果可解释、数据来源可解释、推理流程可视等指标项。

可伸缩：衡量的是模型在不同规模硬件部署环境下的适应性和灵活性。包括模型部署运行尺寸的可伸缩性，即模型能够在不同计算能力和资源条件下运行；不同参数量的部署可伸缩性，意味着大模型能够根据实际需求调整参数规模等。

高扩展：用于衡量大模型能否适应未来技术发展和应用需求的变化，包括对基座大模型版本升级、专业小模型、领域知识库和工具插件的扩展升级及系统功能扩展和二次开发能力的支持。

高可用：是衡量大模型系统在实际应用中的稳定性和响应能力的重要标准。包括系统的可靠性、平均无故障时间、平均响应时间、内容生成速度等指标项。

高运维：体现了模型在运维管理方面的高效性和便捷性。该指标主要考核大模型是否配备了专门的运维平台，该平台能否支持大模型的部署、监控、权限管理、版本管理、故障排查和日志管理等运维活动。

3.2 基础能力

军事大模型基础能力的测评包括通用基础能力、军事基础能力两部分指标体系，前者面向通用基座大模型的基础能力的测试，后者面向军事业务领域大模型需要具备的共性能力的测试。

3.2.1 通用基础能力

语言理解与抽取：是衡量大模型处理自然语言的核心能力，包括对文本进行语义分析，识别出关键的实体和它们之间的关系，以及对文本进行情感倾向的判断。

上下文对话：重点评估大模型在对话系统中的表现，特别是在理解用户意图和维持对话连贯性方面，能够跟踪对话的上下文，确保多轮对话的内容一致。

国际资本市场研报资讯+V: quanqiuzixun8

生成与创作：重点评估大模型在创造性写作方面的潜力，包括生成新闻文章、故事、诗歌等。

常识与知识：是大模型理解世界的基础，涉及对广泛常识的掌握以及对特定领域知识的深入理解。大模型需具备进行基于常识的推理，回答知识库中的问题的能力。

多模态：是大模型处理和理解多种类型数据的能力，如文本、图像和声音。模型需能够理解图像内容，识别语音转换及根据文本内容生成相应图像。

科学计算：是评估大模型在执行数学和逻辑运算方面的能力。大模型需具备解决复杂的数学问题并进行逻辑推导分析数据的能力。

工具使用：模型需能够集成和使用外部 API，从数据库或互联网检索信息，并模拟使用特定软件或工具。

3.2.2 军事基础能力

信息获取：考核大模型从复杂军事战场环境中筛选、定位、整合信息的能力。主要包括信息获取准确性、信息获取速度、复杂信息抗干扰等指标项。

理解分析：重点关注大模型对军事信息理解的准确性、上下文关联广度、理解分析速度。主要包括语义理解准确度、上下文关联、理解分析速度、推理与预测等指标项。

知识推理：重点关注大模型根据已有知识库进行逻辑推断推理的能力，评估大模型在态势研判、战术分析、作战决策等方面的推理水平。包括推理准确性、推理速度、知识库丰富度等指标项。

方案生成：重点关注军事大模型根据任务需求提出解决方案的能力，重点评估大模型生成方案的创新性、实用性和可行性。包括方案创新性、方案实用性、方案可行性、方案调整灵活性等指标项。

规划寻优：重点关注大模型在规划军事行动、资源配置等方面的优化能力、规划速度和环境任务适应性。包括寻优准确性、寻优速度、环境任务适应性等指标项。

国际资本市场研报资讯+V: quanqiuzixun8

模态感知：重点关注大模型和对多种信息模态的融合感知能力、感知准确性与实时性。包括多模态融合、感知准确性、实时性、模态适应性等指标项。

专项智能：重点关注大模型在特定军事任务或领域内展现出的智能化分级水平、任务完成度、稳定性与可靠性等能力。包括任务完成度、智能化分级水平、稳定性与可靠性等指标项。

多智能体：指标设计重点关注多个模型和智能体之间相互配合、协同工作的能力，包括协作效率、信息共享程度、协同任务完成度、协同决策等指标项。

3.3 平台能力

3.3.1 大模型数据生成能力

向量知识库管理：指标设计旨在通过将非数值型数据（如文本、图像等）转换成数值型向量表示，构建、维护和使用这些向量集合提高信息处理的效率和准确性。

数据生成：旨在帮助用户实现数据增强，解决数据集分布不合理、数据集量过少的问题。

数据回流：旨在对大模型多轮问答答案进行数据溯源准确性能力进行测试。

3.3.2 大模型开发训练能力

数据管理：旨在对大模型军事领域源数据进行自动审核标注、任务分发、数据集版本等进行管理。

模型微调：旨在测试大模型对预训练模型进行的再训练或调整过程中是否能适应特定的应用场景或任务，使得模型更好地泛化到新的数据上。

模型交付：将训练完成的模型通过适当的集成和部署流程，转化为可在生产环境中运行的应用程序或服务的过程。包括模型的测试、

国际资本市场研报资讯+V: quanqiuzixun8

验证、封装、优化以及与现有系统的对接，确保模型的稳定性、可扩展性和安全性。

模型服务：指标设计涉及模型的部署、封装为 API 服务、以及与前端应用程序的集成，以便用户或系统可以方便地访问模型的预测能力。

资源管理：旨在确保资源得到高效利用，以满足军事环境特定的业务目标和项目需求。包括需求分析、资源分配、优先级排序、风险管理、成本控制和进度规划等关键活动。

3.3.3 大模型军事应用编排能力

基础插件管理：涉及对用于支持模型测试和评估过程的各种软件组件和工具的集中控制和维护。确保测试环境的稳定性和一致性，支持自动化测试流程，允许快速迭代和持续集成，同时简化复杂测试任务的执行。

军事机理插件库管理：是针对军事场景定向创建的预制插件库，提供武器装备插件、火力打击方案规划插件、军事考评出题专家插件等。

应用编排：涉及对模型测试和评估过程中涉及的多个应用、服务和工作流程进行自动化管理和调度的过程。指标设计旨在实现测试流程的自动化和标准化，提高测试效率，确保测试的可重复性，并能够快速响应测试需求的变化。

提示工程：通过设计和优化输入提示词（prompts），引导和调整大模型的输出结果，以满足特定的测试评估需求。

3.3.4 其他支撑能力

其他支撑能力是指除上述功能要求以外的平台能力，提高模型生产质量、效率，降低成本，提升用户体验和模型服务应用价值。包括会话管理、对话交互、用户反馈、专题场景会话、自定义指令等。

国际资本市场研报资讯+V: quanqiuzixun8

3.4 军事大模型的应用能力

3.4.1 强敌研究领域

科技情报应用指标设计：旨在评估军事大模型对于科技情报信息的广泛搜集、深度理解、逻辑分析以及报告撰写和内容生成能力，通过构建技术预警、情报整编、报告撰写等典型的科技情报领域具体应用场景，对军事大模型信息搜集信息来源的权威性和广泛性，情报理解分析的专业化程度与准确性，内容生成的规范性和独创性等方面给出主观和客观评价标准。

战略预警应用指标设计：旨在衡量军事大模型在威胁分析、形势预测、专题生成和对抗策略制定方面的应用效能。核心指标项包括威胁分析的全面性、形势预测的精确度、专题生成的时效性和对抗策略的创新性。

军事理论应用指标设计：旨在评估军事大模型在规则认知、作战概念发展、战法生成和法规条令遵循等方面的应用效果。核心指标项涉及规则认知的深度、作战概念的创新性、战法生成的实用性和法规条令的适用性。

3.4.2 作战指挥领域

情报整合评估指标设计：旨在全面评价军事大模型在科技情报领域的信息搜集广度、情报分析深度、逻辑推理严密性以及报告撰写和内容生成的专业度。通过设定技术预警、报告撰写、研究脉络和情报整编等关键应用，为情报专业人员提供一个标准化的评价体系，帮助用户选择和优化科技情报领域的大模型应用。

态势研判评估指标设计：旨在评估军事大模型在目标意图识别、COP 生成、战场态势解析等方面的应用能力。

任务规划评估指标设计：旨在全面评价军事大模型在 COA 生成、甘特图生成、冲突消解、火力规划等关键任务规划环节的策略制定能力和资源优化水平。

国际资本市场研报资讯+V: quanqiuzixun8

行动控制评估指标设计：旨在精确衡量军事大模型在无人集群协同、火力精准打击、战场人机融合等行动执行环节的控制能力和协同效率。

复盘评估指标设计：目的在于系统评估军事大模型在战例评估分析、体系效能评估、复盘推演分析等关键复盘环节的深入分析能力和学习改进效果。

认知作战评估指标设计：旨在深入评价军事大模型在认知对抗识别、认知对抗生成、群体认知分析等关键认知作战环节的洞察力和策略制定能力。

3.4.3 装备研制领域

装备论证评估指标设计：旨在全面评价军事大模型在标准撰写、标准贯彻、可行论证、型号对比等关键论证环节的逻辑推理能力和决策支持水平。

装备生产评估指标设计：旨在精确衡量军事大模型在模型生成、自动排产、工艺优化、质量管理等关键生产环节的创新能力和生产效率。

装备试验评估指标设计：目的在于系统评估军事大模型在试验生成、实装分析、仿真推演、交叉比对等关键试验环节的分析能力和试验效果。

装备评估指标设计：旨在深入评价军事大模型在智能评估、评估建模、体系贡献分析、效能评估等关键评估环节的评估精度和决策支持效果。

3.4.4 训练管理领域

军事基础训练评估指标设计：目的在于全面评价军事大模型在体能分析、作战知识学习、靶场训练、综合评估等关键训练环节的教学支持能力和训练效果。

教学训练评估指标设计：旨在精确衡量军事大模型在计划生成、

国际资本市场研报资讯+V: quanqiuzixun8

知识问答、模拟训练智能助手、考核评估等关键教学环节的教学互动性和学习效果。

模拟训练评估指标设计：目的在于系统评估军事大模型在想定生成、智能体生成、计算机生成兵力、复盘评估等关键模拟环节的创新能力和模拟效果。

实战演训评估指标设计：旨在深入评价军事大模型在想定生成、平行演习、智能蓝军、复盘评估等关键实战演训环节的实战模拟能力和决策支持效果。

3.4.5 联勤保障领域

资产管理评估指标设计：目的在于全面评价军事大模型在战备统筹、仓储优化、补给预测、计划生成等关键管理环节的统筹能力和管理效率。

运输投送评估指标设计：旨在精确衡量军事大模型在运筹优化、路线优化、智能投送、精准保障等关键投送环节的优化能力和投送效率。

检测维修评估指标设计：目的在于系统评估军事大模型在故障检测、维修预测、寿命预测、检修助手等关键维护环节的智能诊断能力和维护效果。

3.5 军事大模型的安全能力

3.5.1 军事偏好

军事偏好评估指标设计旨在评估军事认知力是否存在对不同作战单位的亲和/反亲和。这包括但不限于在资源选择、方案规划、决策取舍时对空军、陆军、海军等军兵种职能职责或武装设施的处理。

3.5.2 合法合规

合法合规评估指标设计重点关注模型是否遵守了相关的法律法

国际资本市场研报资讯+V: quanqiuzixun8

规和军事操作准则。确保模型的设计与应用遵循法律、符合军事行动的规范和道德标准等。

3.5.3 军事保密

军事保密评估指标设计确保模型在处理敏感信息时的安全性和保密性。指标设计旨在确保所有敏感数据在传输和存储时都经过加密。

3.5.4 对抗攻击

对抗攻击评估指标设计评估模型在面对对抗性样本时的表现，评估模型抵御恶意攻击的能力。

3.5.5 算法加固

算法加固评估指标设计关注提升模型的安全性和抵御攻击的能力。确保模型在各种输入条件下的稳定性和预测性，提高模型的可靠性。

3.5.6 伪造检测

伪造检测评估指标设计评估模型识别和阻止伪造内容的能力。确保模型能够区分真实和伪造的输入数据、验证数据来源的真实性和可信度、检测和标记异常行为或异常模式。

3.5.7 数据防泄露

数据防泄露评估指标设计确保模型在处理数据时不会泄露敏感信息。能够对敏感数据进行匿名化或去标识化处理，将敏感数据与其他数据隔离存储以及定期进行数据泄露风险评估，并采取相应措施。

国际资本市场研报资讯+V: quanqiuzixun8

4 评估标准

4.1 评分标准

（1）通用能力

语言理解与信息抽取：评估模型在海量文本中精准提炼核心信息与细节的性能，及其在复杂叙述中理解实体关系、情感色彩和隐含意义的能力。

上下文对话：考察模型维护连贯对话、依据前期对话内容有效回应，以及根据用户反馈灵活调整对话策略的水平。

生成与创作：检验模型产出内容的原创性、关联性及与军事规范的契合度，及其根据不同情境调整文本风格的能力。

常识与知识运用：评价模型掌握军事专业知识的深度与广度，及其在此基础上进行合理判断与策略建议的能力。

科学计算辅助：衡量模型在处理军事数据统计与预测时的准确度与效率，以及在量化分析决策中的辅助作用。

逻辑与推理：测试模型识别因果关系、进行情报分析的能力，及其基于逻辑推理提出有效军事策略的效能。

工具使用与系统集成：评估模型与现有军事系统兼容性及操作军事软件工具的效能，强化技术与平台的整合能力。

多模态能力：评价模型跨图像、语音、文本等多媒体信息处理的统一性，及在不同媒介间建立关联进行综合分析的效能。

（2）专项能力

任务规划：评价模型在多任务场景下优先级排序、资源优化配置、风险管理和应急响应的策略性与效率。

运筹优化：测试模型在复杂环境中的路径规划、资源调度灵活性、时间与成本效益的最大化策略制定能力。

仿真模拟：评估模型创建逼真战场环境、预测行动影响、支持交互式演练及确保模拟数据真实性的能力。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/575214302201011233