DOI 10.15302/J-SSCAE-2026.04.051

本体增强行业智能体的关键技术及应用

庾朝富，林传文*，闫洲，陈希，何军权，孟凡颖

（合肥综合性国家科学中心数据空间研究院，合肥 230088）

摘要：行业智能体是大语言模型与行业知识、数据系统及业务工具深度融合的产物，正由面向知识问答的辅助应用向面向复杂任务的执行系统演进。随着行业数据来源不断扩展，业务规则日益复杂，外部工具持续增加，仅依赖大语言模型难以维持任务理解、资源调用与结果判断之间的语义一致。本体增强通过形式化描述领域对象及其关系，并将业务条件、操作能力和运行状态纳入统一语义框架，可为行业智能体提供稳定的知识基础与可验证的任务依据。本文界定了本体增强行业智能体的基本内涵及适用范围；构建了相应的本体语义模型与分层技术体系；梳理了本体构建、知识获取、工具调用、任务规划及结果校验等关键技术；并以某大型制造企业的生产数据为对象开展应用验证。研究表明，本体增强能够通过显式表达业务对象、关系路径、指标口径、时间条件和状态规则，减少多约束查询中的语义偏移，为行业智能体连接异构业务数据提供可验证的语义支撑。

关键词：本体增强；行业智能体；语义模型；工具调用；状态语义

中图分类号：TP315　　**文献标识码：**A

Technologies and Applications of Ontology-Enhanced Domain-Specific Agents

Yu Chaofu, Lin Chuanwen*, Yan Zhou, Chen Xi, He Junquan, Meng Fanying

(Institute of Data Space, Hefei Comprehensive National Science Center, Hefei 230088, China)

Abstract: Domain-specific agents emerge from the deep integration of large language models (LLMs) with domain knowledge, data systems, and business tools. They are evolving from auxiliary applications for knowledge question answering into execution systems for complex tasks. As sources of domain data diversify, business rules grow more complex, and external tools proliferate, LLMs alone struggle to maintain semantic consistency across task understanding, resource invocation, and result assessment. Ontology enhancement formally represents domain objects and their relations and incorporates business conditions, operational capabilities, and runtime states into a unified semantic framework. It thus provides domain-specific agents with a stable knowledge foundation and a verifiable basis for task execution. This study first defines the concept and scope of ontology-enhanced domain-specific agents, develops an ontology semantic model and a layered technical framework, and reviews key technologies for ontology construction, knowledge acquisition, tool invocation, task planning, and result verification. It then conducts an application validation from a large manufacturing enterprise. The findings indicate that ontology enhancement can reduce semantic deviations in multi-constraint queries by explicitly representing business objects, relation paths, metric definitions, temporal conditions, and state rules, thereby providing verifiable semantic support for domain-specific agents accessing heterogeneous business data.

收稿日期：2026-04-29；**修回日期：**2026-07-30

通讯作者：*林传文，合肥综合性国家科学中心数据空间研究院研究员，研究方向为本体论、数据要素创新、数据资产化、数据基础设施；E-mail: lincw11@foxmail.com

资助项目：中国工程院咨询项目“国家数据资源管理调度体系”(2024-JZ-03)；国家自然科学基金项目(62441230)

本刊网址：sscae.engineering.org.cn

Keywords: ontology enhancement; domain-specific agents; semantic model; tool invocation; state semantics

一、前言

大语言模型在自然语言理解方面的进展，为人工智能系统处理开放任务提供了新的技术基础。随着模型的任务理解与规划能力不断增强，大语言模型开始由内容生成器转变为任务控制与执行中枢，智能体由此成为其重要的应用形态。此类系统能够感知任务目标，连接外部知识与业务能力，并依据环境反馈调整后续行为[1]。人工智能的作用也由提供信息逐步扩展至参与任务执行。

智能体向行业领域的深入应用，对系统可靠性提出了更高要求。开放域问答主要关注输出内容的质量，而行业任务往往具有较强的专业性、规范性和过程性，不仅要求系统正确理解任务，还要求其决策与执行适应具体的业务环境。特别是在持续时间较长、参与环节较多的任务中，局部偏差可能在后续过程中逐步累积，从而影响任务完成质量乃至实际业务结果。

从运行机制看，大语言模型主要依据概率生成完成语义理解与方案推演，行业信息系统则强调业务对象、规则条件和责任边界的明确表达。二者在知识表示与执行控制方式上的差异，使通用模型难以直接满足行业任务对准确理解、规范执行和过程可控的要求。具体而言，跨系统信息缺少统一的语义基础，领域知识难以转化为稳定的行动依据，任务运行中的状态变化也难以得到连续管理。上述问题表明，行业智能体需要建立贯通知识组织与任务执行的语义支撑机制[1]。

本体的核心作用在于提供显式、共享且可计算的领域语义，使不同系统能够围绕同一业务对象形成一致理解[2]。行业智能体需要将自然语言任务转化为面向具体业务环境的行动过程，其运行效果取决于语义能否从任务理解贯通至知识获取和执行控制。本体语义由此成为连接模型认知与业务行动的基础，使智能体能够在统一概念体系下理解任务，并依据领域关系与业务规则组织后续行为。面向行业智能体，有必要进一步分析本体语义在运行链路中的作用方式、适用范围及增量价值。

本文采用文献调研、技术比较、体系分析和案例验证方法开展研究。首先，从行业智能体由自然语言理解走向业务执行过程中面临的语义衔接问题出发，分析其对形式化语义的现实需求，进而界定本体增强的作用定位及其与相关技术的关系。其次，依据行业智能体的运行过程，提炼本体语义模型应具备的功能维度，构建相应的技术体系，并阐释各层功能及其关系。随后，围绕该体系的构建与运行，系统梳理相关技术的发展现状、适用条件和主要局限。进一步以某大型制造企业的50条生产数据查询问题为对象，对比仅提供字段说明与加入本体语义两种条件，检验本体语义对查询生成的增量作用。最后，总结研究的主要认识、适用范围和证据边界，并展望本体增强行业智能体的发展方向。

二、行业智能体的运行特征与本体增强定位

行业智能体面向具有明确业务规则和系统边界的任务，需要将自然语言需求转化为可执行过程，并根据运行反馈进行调整。以下从其运行链路中的语义衔接问题出发，分析本体增强的作用定位及其与相关技术的关系。

（一）运行链路中的语义失效

行业智能体通常从用户需求解析出发，随后获取外部知识，选择工具并执行；工具返回结果后，系统还需完成结果校验，并依据环境变化更新任务状态和后续行动。各环节之间具有明显的依赖关系，前序阶段形成的语义偏差会被后续过程继承，进而在多步任务中逐渐放大。

这一运行过程横跨自然语言、领域知识与业务系统。用户需求以开放表达进入系统，知识资源按照各自的模式进行组织，业务接口则依据预设条件执行操作。因此，行业智能体不仅需要生成可行的任务方案，还要保证任务对象、业务含义和规则依据在整个运行过程中保持一致。

第一类问题表现为语义接地与跨源对齐失效。相同的业务对象在不同系统中可能采用不同名称、标识或分类层级，同一指标也可能因统计范围和计算口径不同而具有差异化含义。以企业数据查询为例，用户给出的企业简称需要被准确映射为唯一主体，并进一步关联相应的数据目录和查询服务。若

缺少统一的对象标识与口径映射，智能体可能在主体层级或数据范围上发生偏移，导致后续检索和调用围绕不同的语义对象展开。

第二类问题表现为领域约束难以转化为稳定的执行条件。现有接口描述通常能够规定输入格式，却难以完整表达参数的业务含义和动作的适用范围。对于涉及多个业务环节的任务，计划能否执行还取决于权限是否有效、步骤之间是否满足依赖关系，以及当前状态是否允许继续操作。若这些条件未被显式表达并及时更新，即使工具调用在语法上成立，也可能偏离业务规则，或者继续沿用已经失效的状态信息。

上述问题的本质，在于领域语义尚未贯通智能体的认知链路与执行链路。检索、规划和工具调用等单项技术能够改善局部性能，却难以独立消除跨组件之间的语义断裂。行业智能体因此需要建立共享的语义基础，对任务对象及其关联进行统一描述，并将业务约束、可执行行为和运行状态纳入同一框架，为后续技术体系的构建提供依据。

（二）本体增强的作用定位

文中所称本体增强，是指以形式化本体组织行业对象及其相互关系，并将数据资源、业务规则、工具能力和运行状态纳入统一的语义框架。该框架在自然语言需求与底层数据、业务操作之间建立可解释的对应关系，使任务理解、资源调用和结果校验基于同一语义基础展开，并为执行过程及其结论提供可追溯依据。

本体增强建立在现有智能体技术体系之上，不改变语言模型、检索、规划与业务系统各自的基本功能，而是通过共享语义基础加强不同技术环节之间的协同。借助统一的业务概念、数据口径和规则表达，模型生成的候选方案可以与领域知识及业务条件相互校核。运行过程中形成的新知识和新关系，应在证据充分、逻辑自洽并经专业人员确认后，纳入正式语义体系。

（三）本体增强与相关技术的关系

表 1 从主要功能及本体带来的增量作用两方面，比较本体增强与相关技术的关系。

从技术层次看，本体工程和语义网技术提供形式化表达，数据库、文档库与知识图谱承载数据和事实，大语言模型及检索方法处理开放输入与候选信息，规划器、工具和业务系统完成行动，规则与溯源组件负责验证和记录。本体增强研究关注这些层次之间的语义衔接，下文将分别讨论其在体系架构和关键技术中的具体实现。

三、本体增强行业智能体的技术体系

既有大数据知识工程研究将数据知识化、知识体系化和知识可推理作为技术演进的基本主线。行业智能体在此基础上还需连接数据资源、外部工具和动态业务状态。研究表明，主要问题并非某一技术环节的能力不足，而是业务对象、数据口径、适用条件、执行动作和运行状态在不同技术环节的衔接中难以保持一致。以下以对象语义、关系语义、约束语义、状态语义、动作语义为主轴，分析其在数据与工具资源层、本体语义模型层、智能体能力增强层、行业应用层中的功能位置及相互作用，并进一步讨论相应的关键技术。

（一）运行问题与语义需求

五类语义来源于行业智能体运行过程中出现的

表 1　本体增强与相关技术的功能关系

技术方向	主要功能	本体增强的增量作用
大语言模型	意图理解、内容生成、任务分解和开放推理	提供规范概念、业务边界及候选结果的校验依据
知识图谱[3]	存储实体、属性、关系及事实，支持关联查询	以模式层连接事实、约束、动作和状态
检索增强生成[4]	召回外部证据并支持生成	规范查询对象、限定关系路径并过滤不适用证据
工具使用与函数调用	选择接口、生成参数并利用返回结果	连接任务对象、接口参数、输入输出与执行条件
智能体规划与强化学习	任务分解、路径生成、策略学习和动态调整	给出可审查的动作空间、前后置条件与安全边界
规则引擎与神经符号方法	条件判断、逻辑推理和约束校验	统一规则变量、领域对象和运行事实的语义接地
本体工程	概念、关系、属性、公理和约束的形式化建模	为上述组件提供共享模型，并由运行反馈持续维护

典型问题。业务对象在不同系统中的表达与对应关系不一致，要求系统明确任务涉及的实体和资源，并利用关系语义描述其跨系统联系。工具调用受到权限、业务条件和数据规范的共同约束，因而需要约束语义提供判断依据。接口参数、动作依赖及其预期效果直接影响任务计划的可执行性，需要由动作语义加以描述。随着多步任务持续推进，对象和任务的实际条件也会发生变化，状态所对应的时间、有效期限和版本又会影响规则与数据的适用性，因此还需对状态语义作出专门表达。

从传统本体构成看，对象语义主要落实类、实例和属性，关系语义组织对象属性及跨源映射，约束语义利用公理、规则或形状约束规定合法条件。动作和状态也可由类、关系与公理形式化，将其单列有助于突出行业智能体的外部执行和动态环境特征。结果溯源属于贯穿运行过程的能力，其内容由对象、关系、约束、动作、状态及相应运行记录共同表达。

在功能位置上，底层资源提供事实与行动能力，本体语义模型统一其业务含义，智能体能力层消费语义服务，应用层再依据任务风险和业务需求组合相应能力。本体构建与运行维护、安全评价与治理贯穿各层。由此形成的四层架构描述功能依赖，不限定具体系统的物理部署方式。

（二）总体技术架构

数据与工具资源层汇集行业运行所需的数据、知识及外部工具。本体语义模型层对其中涉及的业务对象、相互关系、规则条件、操作方式和运行状态进行统一表达，在底层资源与智能体运行之间建立共享语义基础。基于上述语义基础，智能体能够对任务涉及的对象和条件形成规范化理解，并进一步完成资源获取、工具调用、行动规划与结果校验。行业应用层则结合具体业务场景及其风险特征，对相关能力进行组织和应用。

数据与工具资源、本体语义模型与智能体能力之间的应用关系如图1所示。

1. 数据与工具资源层

数据与工具资源层是行业智能体获取知识、执

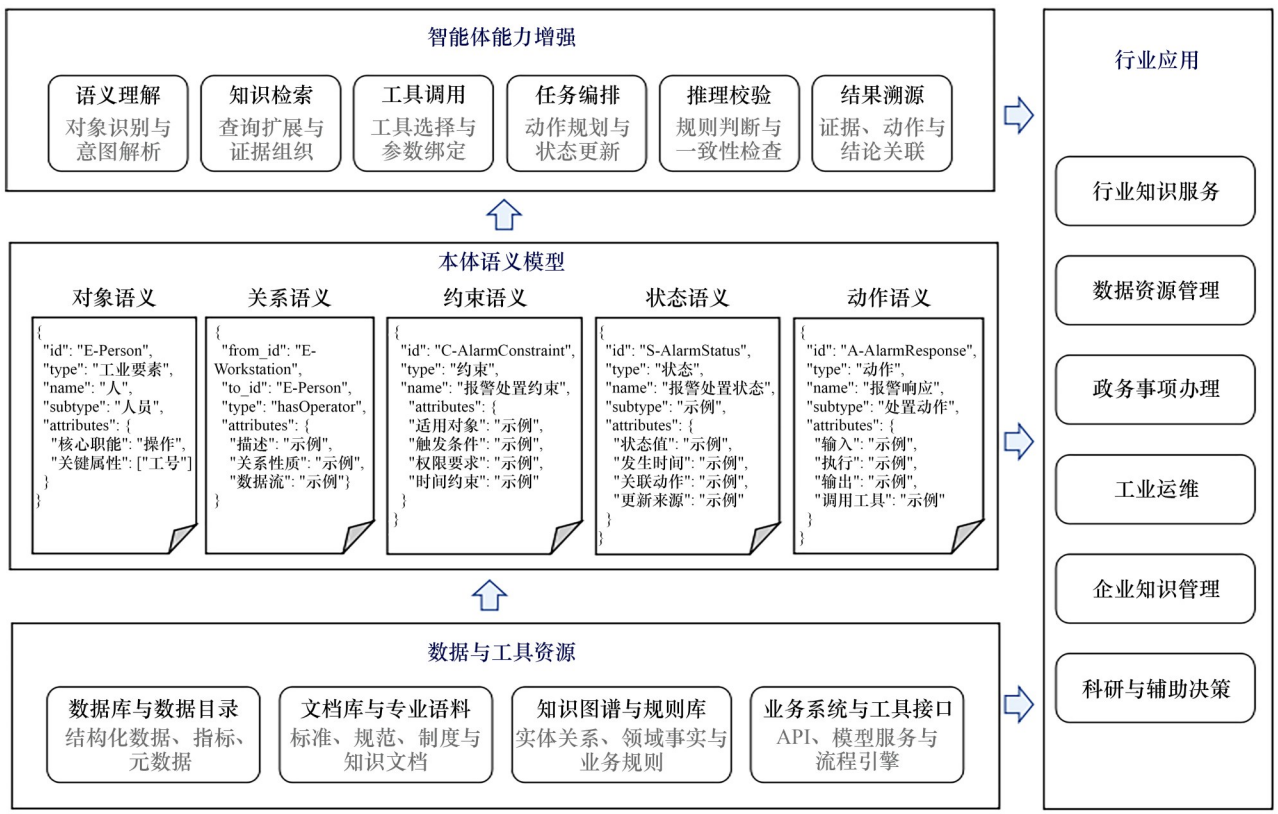


图1　数据与工具资源、本体语义模型与智能体能力的应用关系

注：API为应用程序编程接口。

行任务和改变外部状态的基础，主要包括结构化数据库、数据目录、文档库、知识图谱、业务系统接口、算法模型、计算工具和流程引擎等。数据库以表、字段和代码值组织业务数据，文档库以自然语言文本承载制度和专业知识，知识图谱以实体关系形式表示关联事实，业务接口则通过参数和返回值提供查询或操作能力。不同资源在数据模式、访问协议和管理主体等方面存在差异。

该层保留既有资源的存储、管理与执行方式，并通过语义映射将资源元素关联至本体概念。表、字段、指标和文档片段对应领域对象，数据集和数据服务关联适用任务、数据口径与访问条件，接口参数和返回结果对应规范化业务概念。智能体由此可以从自然语言任务出发，按照业务语义定位底层资源，而不必直接依赖各系统内部名称和技术结构。

数据资源可作为对象语义中的资源对象，并通过关系语义关联服务对象、访问方式、发布主体、时间范围、更新周期和版本信息[5]。语义映射保留不同系统的自治性和真实口径差异，为跨系统理解与调用建立必要的共同语义。

2. 本体语义模型层

本体语义模型层负责对行业智能体运行涉及的对象、关系、约束、动作和状态进行形式化描述。既有研究系统梳理了本体学习及从文本中构建领域本体的主要技术路线、评价方法和关键问题[6-7]。本研究扩展的是领域本体在行业智能体中的建模范围和介入环节：数据资源、工具接口、业务动作、权限条件及运行过程均被纳入共享模型，仍通过类、属性、公理和规则等形式机制进行表达[8]。

本体语义模型的关键在于形成机器可理解、系统可映射和规则可执行的共享语义结构。该层向下连接数据和工具资源，将异构资源转换为统一语义对象；向上提供概念查询、实体消歧、关系检索、约束判断、动作发现和结果校验等服务。在具体系统中，这些能力可以通过图查询、语义检索、规则服务或统一语义接口提供，无需把完整本体持续写入大语言模型上下文。

本体语义模型以对象、关系、约束、动作和状态五类语义为核心：对象语义界定领域实体、数据与工具资源，关系语义描述对象之间的层级、关联与跨源映射，约束语义规定取值、权限、流程和时效等条件，动作语义表征工具和业务操作的适用对象、输入输出及预期效果，状态语义记录对象与任务在特定时间下的状态、有效期、版本和变化。

3. 智能体能力增强层

智能体能力增强层以大语言模型为主要交互和规划组件，在本体语义模型支持下形成五类能力。语义理解与对象识别将自然语言映射为规范概念、实体和属性；知识与资源检索利用概念层级、关系路径和资源条件定位适用证据；工具发现与参数绑定依据工具功能、适用对象和输入输出完成语义匹配；动作规划与安全执行利用动作依赖、状态转移和业务约束检查步骤完整性、顺序及权限；推理校验与结果溯源核对概念、事实、规则、调用结果及其来源。

在这一层，大语言模型主要处理非结构化输入、开放意图和候选方案生成，本体及规则组件保存相对稳定的概念定义、关系模式和确定性约束。系统可由大语言模型生成语义解析或行动计划候选，再由本体语义服务进行匹配、过滤和校验，从而兼顾开放理解与业务确定性。

4. 行业应用层

行业应用层体现本体增强技术在具体行业任务中的应用形态，主要包括行业知识问答、数据资源管理、专业知识服务和辅助决策等场景。不同应用对语义支撑的需求有所差异。知识问答重在规范领域概念、关联多源知识并组织证据；数据资源管理则更关注数据对象、指标口径与资源之间的语义关联，以提高数据理解和利用的一致性。

本体增强在不同任务中的作用程度也有所不同。对于以信息获取和知识服务为主的任务，其价值主要体现在概念规范、跨源检索和证据组织；对于涉及业务规则、外部操作和状态变化的任务，则需进一步发挥条件约束、过程校验和结果追溯作用。由此，行业应用层依据不同任务的业务特征，有针对性地组织和运用相关语义能力。

四、本体增强行业智能体的关键技术

围绕本体增强行业智能体所需的关键技术，依次讨论领域本体构建与维护、语义理解与知识检索、工具语义建模与调用、动作规划与安全执行、推理校验与结果溯源。按照技术问题、主要方法、本体介入点、适用条件和现有局限展开。Toolformer、

GraphRAG、ReAct等相邻技术用于说明行业智能体的实现基础。

（一）领域本体构建与维护

领域本体构建是本体增强行业智能体的基础，其主要任务是从行业文档、数据资源、业务规则和系统接口中识别具有稳定业务含义的概念、关系、属性与约束，并将其组织为形式化、可共享和可推理的语义模型。面向行业智能体，建模范围除领域概念和实体关系外，还需覆盖数据资源、工具接口、业务动作、权限条件和运行状态，并处理资源及业务变化带来的维护需求。

1. 专家驱动与文本驱动的本体构建

传统领域本体主要采用专家驱动的构建方式。知识工程人员通过访谈、标准分析和业务流程梳理并确定本体范围，领域专家负责定义核心概念、关系、属性和规则，随后使用形式化语言完成编码，并通过推理和专家评审进行验证。该方式能够较好地保证概念体系与行业共识、业务规范和管理制度一致，适用于医疗、工业、政务等专业性和规范性较强的领域。然而，专家驱动方法需要领域专家和知识工程人员长期协作，构建周期较长；当数据、规则和工具接口频繁变化时，依靠人工维护也难以及时覆盖新增知识。

为降低知识获取成本，本体学习研究尝试从自然语言文本和网络资源中自动或半自动提取本体要素。典型流程包括领域术语抽取、术语聚类与概念化、上下位关系发现、非分类关系抽取和公理学习。OntoLearn从领域文档中识别术语，对复杂术语进行语义解释，并借助通用词汇资源构建概念层级[9]；Text2Onto进一步通过概率本体模型保存候选概念、关系及其置信度，使系统能够在具体本体语言确定之前组织学习结果，并支持用户对候选内容进行筛选[10]；OntoLearn Reloaded则利用图结构和词义消歧方法，从术语网络中自动归纳概念分类体系[11]。这些研究形成了从语料中发现术语和分类关系的基本技术路线，为大规模行业文本的本体化处理提供了基础。

文本驱动方法对术语和显性关系具有较好的发现能力，但文本共现和语言模式并不必然对应严格的概念关系。例如，两个术语在业务文档中频繁共同出现，不能直接说明二者具有稳定的语义关系；由语言模式识别出的上下位关系，也可能混入组成关系、功能关系或实例关系。政策规范和业务制度中的适用主体、条件、例外、时态和责任要求通常跨越多个句子，单纯依赖局部模式难以转换为完整公理。因此，文本学习结果一般只能作为候选本体要素，需要与行业标准、数据模式和专家知识进行交叉验证。

2. 数据与知识图谱驱动的本体学习

除自然语言文本外，数据库模式、数据实例和知识图谱也是领域本体的重要知识来源。数据库中的表结构、字段类型、主外键关系和代码项可以用于发现候选概念、属性和关系；知识图谱中的实例分布、关系路径和类型结构则可以支持概念边界归纳、类别层次补充和规则发现。相较于文本驱动方法，数据驱动方法能够直接利用实际业务系统中的结构和实例，有利于建立本体概念与底层数据资源之间的映射。

基于描述逻辑的概念学习方法通常给定一组正例和反例，从已有知识图谱中学习能够区分两类实例的类表达式。这类方法可以将实例中的共同特征归纳为机器可解释的概念定义，用于补充类别、公理和实例分类。近年来，相关研究逐步将符号搜索、进化计算和神经表示学习结合起来，以提高复杂类表达式的学习能力。Ontolearn框架集成了多种符号和神经符号类表达式学习方法，可以在大规模资源描述框架（RDF）知识图谱上学习Web本体语言（OWL）类表达式，并将形式化表达转换为便于人员理解的自然语言描述[12]。

数据与知识图谱驱动的方法适合在已有系统和实例数据基础上扩展本体，但其结果高度依赖数据质量和样本覆盖。当历史数据本身存在字段口径不一、类别标注错误或业务流程缺失时，自动归纳出的概念和规则可能固化原有数据问题。部分隐含规则并不会直接反映在数据分布中，如尚未发生的禁止行为、低频异常处理和管理责任边界，仍需从制度规范和专家知识中获取。因此，数据驱动方法更适合用于发现候选模式和补充已有本体。

3. 大语言模型辅助的本体构建

大语言模型增强了对复杂自然语言和长文本的理解能力，为本体构建提供了新的自动化手段。相较于传统模式匹配和分类模型，大语言模型能够根据任务说明，从行业标准、政策文件、业务文档、数据字典和接口说明中生成候选术语、概念定义、

上下位关系、对象关系和规则结构；也可以根据少量示例，将自然语言规则转换为结构化表示，并为已有概念生成定义、同义表达和使用说明。

从本体构建流程看，大语言模型可以参与多个环节：在需求分析阶段辅助生成或整理本体能力问题，明确本体需要回答的业务查询；在知识获取阶段识别领域术语、对象、属性、条件和动作；在概念化阶段归并同义词、提出候选分类体系和非分类关系；在形式化阶段将候选知识转换为类、属性、三元组或规则模板；在评价阶段解释逻辑冲突、生成测试问题和比较不同版本。面向行业智能体，还可以利用大语言模型分析 API 说明、流程文档和调用记录，生成工具功能、参数语义、前置条件和异常状态等动作本体候选。

LLMs4OL 研究将大语言模型应用于术语类型识别、分类体系发现和非分类关系抽取等本体学习任务，并在词汇语义、地理和医学等不同知识领域对多类模型进行了评价[13]。研究结果表明，大语言模型具有利用语言知识辅助本体学习的潜力，但未经专门适配的基础模型尚不足以稳定处理需要较强逻辑推理和领域知识的本体构建任务；通过领域数据适配并将专家置于验证环节，更适合作为缓解知识获取瓶颈的辅助工具，而非独立完成本体设计。

大语言模型辅助本体构建存在输出不确定性的问题。对于同一文档和提示，模型可能生成不同的概念划分和关系结构；模型也可能根据一般语言知识补充原文中不存在的术语、关系和规则。自然语言上合理的概念关系并不一定满足形式逻辑要求，模型生成的本体还可能出现概念循环、定义重复、关系方向错误和公理冲突。因此，较为可行的路线是建立“大语言模型生成候选 – 形式化工具验证 – 领域专家审核”的人机协同机制。大语言模型负责处理大规模非结构化材料和提出候选结构，约束验证工具负责发现结构和逻辑问题；专家则负责确定概念边界、规则效力和高风险动作的约束。

4. 本体对齐、融合与复用

行业智能体通常需要连接多个部门和业务系统，不同系统往往已经形成各自的术语体系、数据模式或领域本体。因而，领域本体构建不能完全依赖从零开始建模，还需要通过本体对齐和融合发现不同语义模型之间的概念对应关系，并在保留来源差异的基础上形成共享语义。

传统本体对齐方法综合使用字符串相似度、同义词资源、概念描述、上下文关系和图结构等信息生成候选映射。对于语义表达较丰富的本体，还需要在映射后检查新引入的公理是否导致概念不可满足或整体逻辑不一致。LogMap 将词汇匹配同逻辑推理和冲突诊断结合，可以在处理大规模本体的同时识别和修复部分不一致映射[14]；BERTMap 利用上下文语言模型判断概念文本的语义对应关系，并进一步使用本体图结构和逻辑约束扩展、筛选映射[15]。这类方法体现了本体对齐由单一词项匹配向“语言语义 – 结构关系 – 逻辑校验”协同演进的趋势。

大语言模型进一步提供了利用自然语言定义和背景知识进行对齐的能力。OLaLa 通过零样本和少样本提示，让大语言模型判断不同本体概念之间的对应关系，并研究了候选生成、提示设计和结果过滤等环节[16]。相关实验说明，在提供适当示例和概念描述时，大语言模型能够获得具有竞争力的匹配结果；但模型选择、提示方式、候选规模、调用成本和结果可复现性仍会显著影响实际应用。

面向行业智能体，本体对齐不仅涉及不同领域概念之间的同义映射，还包括本体概念与表字段、指标、接口参数、工具返回值之间的跨层映射。此类映射很难简单归结为概念等价关系，还可能表现为上下位、组成、转换、近似或条件对应。例如，一个系统中的“企业编号”可能对应另一个系统的内部主键，而不能直接等同于统一社会信用代码。因而，本体融合应保留映射类型、适用条件、来源和置信度，避免为追求统一而消除真实存在的业务口径差异。

现有研究表明，文本、数据和大语言模型能够补充专家建模并降低候选知识获取成本，本体对齐也可综合词汇、结构和逻辑信号复用既有资源；这些自动化路线尚不能独立保证领域正确性、逻辑一致性和业务适用性。对于规则密集的行业智能体，人机协同的价值主要体现在扩大候选覆盖，核心概念、高风险规则和关键工具映射仍需可复核的确认与持续维护。

（二）语义理解与知识检索

语义理解与知识检索是行业智能体连接自然语言需求和外部知识资源的基础环节。用户通常以非

规范、上下文依赖的自然语言描述任务，而数据库、知识图谱、文档库和业务系统则按照不同的数据模式、术语口径和访问方式组织信息。因而，智能体不仅需要判断用户“说了什么”，还需要确定其表述对应何种领域概念和业务对象、涉及哪些属性与关系、受到何种时间和权限条件限制，以及相关证据存储于哪些资源。

传统检索主要根据关键词匹配、统计相关性或向量相似度，从文档集合中召回与查询相关的文本。稠密段落检索通过分别编码查询与文档，在向量空间中完成大规模候选召回，增强了不同语言表达之间的语义匹配能力[17]。但行业任务中的正确检索还依赖对象身份、概念层级、关系类型、数据口径和适用条件，仅凭文本相似度难以稳定区分名称相近但业务含义不同的对象，也难以组织跨文档、跨数据库和跨接口的关联证据。

本体引导检索先将自然语言问题转换为由领域概念、实体、关系、属性和约束构成的语义查询，再以该查询组织关键词、向量和图结构等检索通道。其介入点包括查询对象规范、关系路径限定以及候选证据的口径与适用条件过滤，从而把相关性召回与业务适用性判断分开。

1. 领域表达的本体语义映射

行业语义理解的关键在于将自然语言中的业务表达对应到本体所规定的概念、对象及其相互关系。本研究将这一过程称为本体语义映射。其内容既包括领域术语的规范化和具体对象的识别，也包括类别、属性与关系的确定；对于时间范围、数量要求和业务条件等信息，还需将其组织为结构化的语义表达。通过上述过程，自然语言任务才能与领域知识、数据资源和业务能力建立明确联系。

现有实体链接研究主要关注自然语言中的名称表达与知识库具体对象之间的关联。BLINK 采用双编码器分别表示实体提及的上下文和候选实体描述，从大规模实体集合中检索候选对象，再利用交叉编码器进行精细排序[18]。ELQ 进一步将实体提及识别与实体链接相结合，能够在一次编码过程中发现问题中的多个实体，并将其关联至知识库中的相应对象[19]。这类方法形成了由对象识别、候选检索到上下文消歧的基本技术路径，适用于人员、机构、设备和事项等具体对象的识别。

除具体对象链接外，概念规范化研究还关注自然语言表述与标准概念之间的对应。例如，同一疾病、药物或临床现象通常存在多种同义表达，相关方法通过术语体系或领域本体将这些表达归并到统一概念。BioSyn 综合利用词形特征和稠密语义表示，并通过同义词边际化机制提高生物医学实体规范化的准确性[20]。SapBERT 则利用统一医学语言系统中的同义概念进行自对齐预训练，使具有不同语言形式但语义相同的术语在表示空间中相互接近[21]。相关研究表明，本体语义映射的目标不仅是确定自然语言所指向的具体对象，还包括明确其所属的规范概念。

知识图谱问答研究进一步将映射范围扩展到类别和关系。研究人员提出将自然语言问题转换为结构化查询，在查询构造过程中处理问题表达与知识库实体、类别及关系之间的对应[22]。EARL 将实体链接和关系链接作为相互关联的任务，利用知识图谱中的连接结构对候选实体和候选关系进行联合判别[23]。后续关系链接研究还尝试利用句法或语义结构识别自然语言中显式及隐含的关系表达[23]。这类研究使“属于何种对象”“对象之间存在何种联系”等信息能够映射到知识图谱或本体中的类别与关系，而不再局限于名称相似度层面的对象匹配。

在面向数据库的自然语言查询中，类似问题通常被表述为模式链接，即识别用户问题所涉及的数据表、属性和关联结构。IRNet 首先在自然语言问题与数据库模式之间建立链接，再通过中间语义表示生成结构化查询[24]。RAT-SQL 利用关系感知机制同时表示数据库内部结构以及问题表达与数据模式之间的联系，从而加强自然语言与数据表、字段及其关联关系的对应[25]。从本体增强行业智能体的视角看，模式链接可以理解为将业务对象及其属性进一步关联至底层数据资源，是概念语义向数据表示延伸的重要环节。

对于包含时间、空间、比较、排序和统计要求的复杂任务，仅识别概念和对象仍不足以形成完整的任务语义。语义解析研究通过构造查询图或逻辑形式，将对象、关系与限制条件组织为可执行的结构。研究人员采用分阶段查询图生成方法，逐步确定问题涉及的主题实体、关系路径和约束结构，使自然语言问题能够转化为知识库可处理的查询表示[26]。这表明，本体语义映射还需要在概念和对象识别的基础上，进一步表达各语义要素之间的组合关系。

现有研究已经分别覆盖具体对象链接、标准概

念规范化、类别与关系链接、数据模式链接及结构化语义解析等问题，但不同方法大多面向特定知识库、术语体系或查询任务展开。行业智能体通常同时涉及业务概念、数据资源、外部工具和运行条件，需要将上述技术统一到领域本体所规定的语义框架中。

2. 本体引导的查询解析与语义扩展

完成领域表达的本体语义映射后，还需根据任务目标明确查询涉及的核心对象、目标信息及其关系条件，并形成相应的检索表达。用户问题通常较为简短，难以覆盖知识资源中的同义表述和关联概念。查询扩展通过补充与原始意图相关的表达，缓解用户用语与资源表述之间的差异。传统方法主要利用词典、语料统计和相关反馈生成扩展内容，本体则能够依据已定义的概念结构和语义关系确定扩展范围，并为扩展结果提供明确的语义依据[27]。

从扩展所利用的语义信息看，本体引导的查询扩展可概括为三种基本方式。词汇扩展依据规范术语及其同义表达，处理简称、别名和缩略语等表述差异；概念扩展利用类属结构调整查询的抽象层次，并可由概念延伸至相应的具体对象；关系扩展则沿本体中已经定义的业务联系发现与任务相关的对象。既有研究先后考察了同义关系、上下位关系及知识库实体链接在查询扩展中的作用，表明扩展内容既可以来源于词汇语义，也可以来源于实体属性和显式关系结构[27]。本体扩展需要在信息覆盖与语义专指性之间保持平衡。已有研究表明，不同语义关系及其组合对检索效果的影响存在差异，扩展关系增加并不必然带来性能提升，候选内容的选择仍需结合查询语境进行控制[27]。因此，查询扩展应综合考虑关系类型、语义距离和候选权重，避免扩展范围持续扩大所造成的语义偏移。

大语言模型能够生成查询改写、同义表达和相关概念，为复杂任务提供较为丰富的候选内容。然而，模型知识不足或原始问题存在歧义时，生成的扩展可能包含错误信息，也可能过早限定查询含义，进而降低检索效果[28]。知识感知查询扩展研究开始将语言模型与知识图谱结合，利用结构化关系限定候选内容。研究人员提出面向文本与关系条件的知识感知扩展方法，在生成扩展表达时引入知识图谱中的文档关系，并通过关系筛选保持扩展结果与查询结构要求的一致性[29]。

3. 基于本体的跨源查询与资源发现

行业知识资源通常分布于关系数据库、知识图谱、文档系统和数据目录等不同载体。集中索引可以统一检索入口，却难以消除各类资源在对象定义、属性表达和统计口径方面的差异。基于本体的数据访问及虚拟知识图谱研究以领域本体作为统一的概念层，通过语义映射连接既有数据资源，使用户能够依据业务概念提出查询，而不必直接了解各系统的存储结构。

经典的本体驱动数据访问体系由领域本体、数据源模式和语义映射构成。用户基于本体提出查询后，系统综合本体公理和映射规则对查询进行改写，再将其转换为底层数据系统能够执行的查询表达。关系数据库到RDF映射语言（R2RML）标准对关系数据库与RDF数据之间的映射作出了规范，可用于生成RDF数据，也可在不改变原有数据存储方式的条件下提供虚拟化访问[30]。

查询改写能力建立在轻量本体语言和数据库查询技术的结合之上。DL-Lite描述逻辑在保持基本概念建模能力的同时，支持将合取查询转化为关系数据库可以处理的查询形式，并构成OWL 2 QL的重要理论基础[31-32]。Ontop等系统进一步实现了从SPARQL到结构化查询语言（SQL）的转换，使关系数据库中的数据能够通过本体概念进行访问[33]。对于分布在多个图数据端点中的信息，SPARQL则提供了针对RDF图数据的标准查询语言，可以通过图模式匹配查询必选关系、可选关系及其组合[34]。

相关研究已经在能源和制造业场景中开展了应用验证。Statoil的研究以本体连接多个业务数据库，将面向领域概念的查询自动转换为数据查询，并通过可视化方式支持专业人员获取勘探数据。西门子股份公司的相关研究进一步将本体驱动访问扩展到静态数据与流式数据的联合分析，以支持设备监测和故障诊断。博世集团的制造数据研究则采用虚拟知识图谱整合不同机器和生产环节的数据，为跨设备分析提供统一的语义视图。这些研究表明，本体能够降低专业人员理解异构数据结构的负担，并为跨系统分析建立相对稳定的概念基础[35-37]。

面向行业智能体，跨源查询还需由结构化数据访问延伸至文档和服务资源。在文中构建的体系中，可通过本体建立业务对象与指标、数据集、知识图谱实体、文档内容和数据服务之间的语义联

系。智能体依据任务涉及的对象和信息需求确定候选资源，再根据资源形态分别采用关系查询、图查询或文本检索。资源发现还需要以规范化元数据描述资源的基本特征。数据目录词汇表（DCAT）提出了面向数据集和数据服务的统一目录模型，可支持不同目录之间的元数据交换和联合发现；PROV-O用于描述资源的来源及其生成过程；数据质量词汇表可以表示质量指标、评价结果和适用性信息；开放数字权利语言（ODRL）则为资源的使用许可、禁止事项和相关义务提供语义表达。基于这些元数据，智能体可以综合发布主体、更新时间、质量评价和使用政策选择适合当前任务的资源，而不仅依据文本相似度确定检索顺序[5,38-40]。

除关系数据库和知识图谱外，近年来相关研究开始将本体用于文档知识的组织与检索。OG-RAG依据领域本体将文档中的事实组织为具有概念层级和关系约束的超图，并围绕用户问题选取能够覆盖相关事实的局部结构，以此为大语言模型提供检索上下文[41]。该研究表明，本体驱动的资源发现可以由结构化数据访问进一步延伸至非结构化文档，使检索范围的确定不仅依赖文本相似性，也能够参考领域概念及其关系结构。不过，这一方向目前仍处于探索阶段，其适用范围和实际效果尚需更多行业场景验证。

（三）基于本体约束的工具发现与调用

外部工具使行业智能体能够访问数据资源、调用专业计算能力并参与业务过程。Toolformer、ToolLLM等研究表明，大语言模型可以依据工具说明识别调用需求，选择相应工具并生成调用参数[42-43]。然而，行业任务中的工具使用不仅取决于接口形式是否匹配，还涉及工具能力与业务对象、任务条件和数据含义之间的对应关系。功能表述相近的工具可能面向不同业务对象，同一参数形式也可能承载不同的行业含义。因此，工具调用的关键在于将自然语言任务与外部能力建立准确的语义联系。

1. 工具能力与业务动作的语义描述

现有函数调用机制通常以工具名称、自然语言说明和参数模式描述外部能力。这类描述能够约束参数结构并减少调用形式上的错误，但对工具在具体业务情境中的含义表达有限。本体增强的工具描述需要将接口所提供的技术能力与相应的任务建立联系，明确其作用对象、输入输出所表达的领域含义，以及执行可能产生的业务影响。

这一描述过程与前文提出的五类语义相互衔接。工具所作用的领域实体由对象语义加以确定，输入输出与业务对象之间的联系由关系语义表示，工具使用的适用条件由约束语义说明，可实施的操作及其预期结果则构成动作语义的重要内容。工具执行后形成的实际变化进一步反映在状态语义中。由此，工具语义建模不再被理解为对接口字段的简单补充，而是将外部能力纳入统一领域语义体系的过程。

语义Web服务研究较早探索了服务能力与接口形式的分离描述。OWL-S通过服务概况、过程模型和服务绑定刻画服务能力、执行过程与具体访问方式，并利用输入、输出、前置条件和执行效果描述服务行为[44]。SAWSDL则允许将Web服务描述语言（WSDL）及其数据结构中的相关要素关联到外部语义模型，为异构服务的语义标注、发现和匹配提供基础[45]。这些研究表明，接口的技术描述可以与领域语义模型分别维护，再通过语义标注和映射建立联系。对于行业智能体而言，这一思路可用于将面向大语言模型的工具说明与领域本体相连接，使工具能力能够按照业务概念进行理解和组织。

工具说明的表达质量同样会影响大语言模型的使用效果。EASYTOOL将冗长且形式不一的工具文档转化为结构较为统一的工具指令，实验结果表明，精练后的描述有助于模型识别工具功能和参数要求[46]。该研究主要表明了工具说明规范化的作用，而本体在此基础上进一步建立工具描述与领域概念之间的语义联系，使来自不同系统的工具能够依据相对一致的业务含义进行表示。

2. 工具能力的语义匹配与调用衔接

随着可用工具规模扩大，从工具集合中识别与用户任务相关的外部能力，已经成为大语言模型工具使用的重要环节。ToolLLM配置神经网络检索器，根据用户指令从大量应用程序编程接口中推荐候选工具[43]；近期提出的TOOLRET基准进一步将工具检索作为独立研究任务，比较了稀疏检索、稠密检索及生成式检索等方法。相关研究主要依据用户请求与工具说明之间的文本关联形成候选，并表明通用信息检索模型在工具检索任务上的表现仍存在明显局限[47]。

工具选择还可以利用服务能力的形式化描述。

OWL-S 以输入、输出、前置条件和执行效果表征服务功能及其可能引起的状态变化[44]；SAWSDL 通过语义标注将接口操作及其输入输出结构关联至外部本体，并支持接口数据与语义模型之间的转换映射[45]。在此基础上，OWLS-MX 将逻辑推理与文本检索结合，根据服务请求和服务描述中的输入输出语义进行匹配[48]。这些研究表明，工具发现可以同时参考文本描述和形式化语义，使候选选择建立在工具能力、任务对象及其所需信息之间的对应关系之上。

工具调用还涉及任务语义与接口表示之间的转换。SAWSDL 提出的语义标注与数据映射机制，可以将接口中的消息结构关联至本体概念；OWL-S 的服务绑定则进一步描述语义层输入、输出与具体接口消息之间的对应关系[44-45]。借鉴这一思路，自然语言任务中识别出的业务对象及其条件，可以依据语义映射转换为工具所需的输入；工具产生的结果也可重新关联至相应的领域对象、属性或状态。

（四）基于动作语义的任务规划

复杂行业任务通常需要多个外部能力相互配合，ReAct 通过交替生成推理内容、工具动作和环境反馈，使模型能够根据执行结果调整后续行为[49]。ToolPlanner 进一步利用路径规划和反馈信号加强多轮工具使用中的任务完成与指令遵循能力[50]。ToolSandbox 则将隐含的状态依赖、交互过程和中间目标纳入评价，表明工具使用研究已由单次调用逐步转向有状态的连续任务执行[51]。

上述研究主要依靠语言模型生成和调整行动序列，能够适应自然语言任务的开放性，但对业务对象、行动条件和状态变化的表示通常服务于特定工具环境。行业任务中的动作规划还需要明确不同操作作用于何种对象，在何种条件下可以实施，以及执行后会产生何种业务变化。因而，在规划过程中，本体需要为行动及其相关条件提供统一的领域语义，使不同工具能够在共同的业务概念和状态描述下参与任务组织。

语义 Web 服务研究为本体与自动规划的结合提供了直接依据。OWL-S 利用输入、输出、前置条件和执行效果描述服务行为，使服务能力能够转化为规划过程可以利用的形式化信息[44]。Sirin 等进一步将 OWL-S 服务描述转换为 SHOP2 分层任务规划模型，并通过任务分解与动作选择生成服务组合方案[52]。这说明，当工具能力及其适用条件具有明确的语义描述时，自动规划可以利用这些信息确定服务之间的组合关系和执行次序[52]。

基于这一思路，行业智能体可以将任务目标表示为需要达到的业务状态，将工具能力表示为具有适用条件和预期结果的行动，而对象类型、业务关系和状态条件由本体提供。近年来，大语言模型与形式化规划工具的结合进一步验证了这种分工的可行性。研究人员将自然语言中的复杂约束转化为形式化可满足性问题，再交由具有完备求解能力的工具生成和检验计划[53]。相关结果表明，在约束较多的任务中，由语言模型完成需求理解和形式化转换，再由专门求解工具判断可行性，比完全依赖模型直接生成行动序列更有利于获得可验证的结果[53]。

（五）形式化校验、事实核验与结果溯源

行业智能体的输出可能直接用于业务判断，也可能进一步影响外部业务过程。其可信性不仅表现为生成内容的语言质量，还涉及结论的有效性、推理过程的可解释性以及相关责任和依据的可追溯性。美国国家标准与技术研究院（NIST）提出的人工智能风险管理框架也将有效可靠、可解释以及透明负责等方面作为可信人工智能的基本特征[54]。因此，行业智能体需要从形式化一致性、事实证据和形成过程三个方面对运行结果进行检验[54]。

形式化校验用于判断事实、规则和状态之间是否保持语义与逻辑一致，事实核验用于判断生成主张是否得到有效证据支持，结果溯源则记录数据、工具、规则及相关处理活动与最终输出之间的联系。三者分别面向内部推理关系、外部证据依据和过程可复核性，作用有所区别、又相互补充。

1. 形式化推理与一致性检验

本体公理可以用于推导显式数据中未直接给出的概念归属和对象关系，也可以检查概念定义、实例类型及其相互关系是否存在逻辑矛盾。HermiT 以 hypertableau 演算为基础，支持符合 OWL 2 直接语义的本体推理，为复杂概念体系的分类与一致性判断提供了代表性工具[55]。

行业任务中的判断依据还包括授权条件、业务资格和操作要求等规则。语义 Web 规则语言（SWRL）将 OWL 中的类和属性与 Horn 形式规则相结合，可依据显式事实和条件推导相应结论[56]。形状约束语

言（SHACL）则面向RDF数据图定义结构和取值条件，用于验证已有数据是否符合规定的形状约束[57]。由此，本体公理侧重表达概念含义及逻辑蕴含，业务规则用于形成条件性判断，数据约束则用于检验显式记录的规范性[56-57]。

本体推理与数据验证具有不同的语义基础。OWL采用开放世界假设，知识库中未明确表达的事实可能只是未知，不能直接判定为不成立；数据质量检查则往往需要判断必要属性是否缺失或取值是否符合规定。因此，行业智能体需要根据任务性质区分知识推导与完整性验证，避免将信息缺失等同于事实否定[8]。

已有行业研究表明，逻辑推理、数据约束和专业审核通常需要结合使用。OpenPVSignal研究将药物警戒信号报告组织为知识图谱，利用HermiT检验本体一致性，设置SHACL规则验证数据结构，并由药理学和医学专家对实例及概念关系进行复核[58]。DTSafe-O将OWL 2本体、SWRL规则和SHACL约束与传感器、建筑信息模型及监管文件相连接，用于安全要求的形式化表达和合规性验证[59]。这类案例说明，行业校验通常不是依靠单一推理机制完成，而是需要在统一概念体系下综合运用逻辑推导和专业确认。

2. 生成结果的事实与证据核验

大语言模型生成的回答可能同时包含已有事实、推导性结论和缺乏依据的陈述。对整段文本作出统一的正确性判断，难以识别具体问题。事实核验研究通常将生成内容分解为可验证主张，并分别检索和判断相应证据。

FEVER将主张划分为证据支持、证据反驳和证据不足三类，同时要求系统检索形成判断所需的证据[60]。这种任务设置区分了“已有证据表明结论不成立”与“现有证据不足以判断”两种情形。FActScore进一步将长文本分解为原子事实，并以获得可靠知识源支持的事实比例评价长文本的事实性[61]。

证据链接本身并不能保证主张得到充分支持。ALCE从内容正确性、引用完整性和引用准确性等方面评价带引用的长文本回答，表明引用生成与证据支持质量需要分别考察[62]。在医疗问答领域，MedCite进一步围绕医学文本的引用生成和评价建立专门框架，并引入医学专业人员对引用支持关系进行验证[63]。这些研究说明，行业证据核验需要结合领域知识和专业判断，通用文本相似性不足以单独确定证据的适用性。

生成后的证据补充与评价也形成了相应方法。RARR通过检索外部资料、比较生成内容并修订缺乏支持的陈述，为已有文本补充归因依据[64]；RAGAs则将检索上下文和生成回答的评价加以区分，用于识别偏差产生于信息检索还是内容生成[65]。这些方法可以补充最终结果层面的评价，但行业智能体还需进一步检查主张与业务对象、规则条件及当前状态是否一致。

3. 结果归因与过程溯源

事实核验关注结论能否获得证据支持，过程溯源则进一步说明结论由何种数据和处理活动形成。对于包含知识检索、工具调用、规则判断和人工参与的行业任务，仅保留最终答案难以满足后续解释与复核要求。

PROV-O以实体、活动和主体为核心要素描述信息的来源及其生成过程，可表示某项结果由何种活动产生、使用了哪些输入以及相关活动由何种主体实施[38]。这一通用模型还允许根据具体领域扩展相应的对象和关系，为异构系统之间交换溯源信息提供了标准化基础。

近期研究已将标准化溯源扩展到机器学习。Schlegel和Sattler提出符合W3C PROV的数据模型，并开发MLflow2PROV工具，持续提取模型开发过程中由MLflow和Git产生的流水线活动、数据、模型及软件工件之间的关系[66]。该案例说明，仅依赖一般的模型管理记录往往难以形成完整且可查询的过程链，标准化溯源模型能够加强不同阶段和不同类型工件之间的关联。

OpenPVSignal和DTSafe-O等案例进一步说明，在专业性和风险要求较高的行业中，通常需要将形式化方法与领域专家复核结合使用。本体在其中主要提供共同的概念、关系和条件表达，使不同校验方式围绕一致的业务对象展开。

五、本体增强的制造业生产数据查询验证

（一）案例数据

为进一步考察本体语义信息在真实行业数据查询任务中的适用性，本研究选取某大型制造企业生

产管理场景开展原型测试。测试数据来自企业资源计划（ERP）、产品数据管理（PDM）和制造执行（MES）等系统，涉及生产排程、生产报工、库存、出入库、包装、称重、产品规格和客户主体等业务对象，共计数据库表 12 张，涉及 1353 个字段，10 578 681 条数据（见表 2）。案例测试数据包含 50 个由企业业务人员围绕日常生产管理需求提出的问题，覆盖产量趋势、库存分析、出入库分析、计划与任务、批次追溯、包装发运和称重计量等主题。该类问题类型通常同时包含业务对象识别、跨表关系选择、指标口径判定、时间范围解析和聚合排序要求，适合用于观察本体语义在结构化数据查询中的增量作用。

（二）本体构建

本案例的领域本体，采用兼顾标准规范性的标准锚定方法结合数据真实性数据激活的两阶段递进方式构建，兼顾语义完备性与场景适用性。

第一阶段是“标准锚定”，以 ISA-95/IEC 62264、ISA-88/IEC 61512、ISO 22400、ISO 10303 STEP、OPC UA 与 ISO 8000 六项标准为概念来源，对制造领域的术语体系作出规范定义，涵盖产品、工艺、生产、物流、设备与数据质量的完整语义链条，据此抽取其中的对象类型、对象关联与对象特性，分别映射为本体的概念、关系与属性，经跨标准语义对齐后合并为一个覆盖面宽泛的候选本体。该候选本体不追求最小性，其作用在于为当前领域提供一个语义完备且与具体企业无关的假设空间。基于上述标准集共形成 77 项类别、47 项属性、70 项关系。

第二阶段是“数据激活”，以上述大型制造企业实际业务数据（12 张表、1353 个字段、约 1058 万条记录）作为证据源对上述假设加以检验：遍历候选本体中的每一个元素，考察其能否在表结构、字段语义中找到对应者，可对应者判定为被激活；遍历结束后对未获任何证据支持的元素执行剪枝，剪枝后的结果即为满足当前数据场景的企业制造本体。其中，数据中存在而标准未覆盖的部分另以企业局部命名空间作受控扩展。最终，形成面向企业的制造本体，包括 47 个类、63 条关系、243 项属性，如图 2 所示。

（三）实验结果

为区分不同层次语义信息对查询生成的影响，测试设置了三类对照条件。原始结构条件仅提供底层数据表名称和字段名称，不提供字段业务含义、指标口径及关系路径，用于表征模型直接依据物理数据结构生成查询的情形；字段说明条件在原始结构基础上补充表和字段的中文含义，用于考察字段级语义对查询生成的支撑作用；本体增强条件则在字段说明基础上进一步加入指标口径、对象关系、状态规则和禁止性约束等本体语义，用于检验本体约束对复杂业务查询生成的增量作用。三类条件均要求模型输出只读 SQL，并通过数据库执行计划检查 SQL 可执行性。评价指标包括语义组件命中率、整题全通过率和 SQL 执行计划通过率。其中，语义组件命中率用于统计生成 SQL 对数据表、关联路径、时间条件、统计字段、聚合方式和排序规则等必要查询要素的覆盖比例；整题全通过率用于衡量

表 2　某大型制造企业生产管理场景数据集概况

表名	表名含义	字段数	数据量
存储资源实体	记录企业存储资源单元的基础信息（包括编码、名称、层级、状态等）	46	15 670
产品标识主数据	记录产品的主数据定义（包括客户产品信息、工艺路线、包装规格等）	167	10 572
制造指令明细	记录制造任务清单的详细条目（包括工序、产线、计划时间、物料需求等）	110	283 067
作业执行记录	记录制造工序的完工汇报信息（包括产出数量、质量指标、操作人员等）	127	881 320
制造计量记录	记录制造过程中的称重数据（包括毛重、净重、样品重、皮重等）	86	669 668
工艺流转追溯	记录物料在多道工序间流转的追溯卡片（包括工序进度、质量结果、标记状态等）	158	826 188
库存结存记录	记录库存物资的收发存数据（包括期初、入出库、期末的数价及批次维度）	106	116 155
入库作业记录	记录物料入库作业的详细流水（包括来源、去向库位、物料、数量、价格等）	150	1 950 315
出库作业记录	记录物料出库作业的详细流水（包括领用方向、发料库位、物料、数量、价格等）	149	2 167 026
包装作业记录	记录产品包装作业明细（包括装箱方式、包重、作业员、箱码等）	93	115 401
批次追溯记录	记录生产批次的全过程追溯信息（包括批次标识、关联客户、工序进度、质量数量等）	114	886 011
库存期间结存	记录按月份核算的库存期初期末结存（包括结存日期、批次、物料、数量等）	30	2 657 288

单个问题所需语义组件是否全部满足，以反映查询结果的整体业务可用性；SQL执行计划通过率用于检验生成查询能否被数据库解析并形成执行计划。

如图3所示，随着语义信息层次的逐步丰富，两种模型的查询生成质量均呈现持续提升趋势。相较于仅提供原始物理结构，补充字段说明后，两种

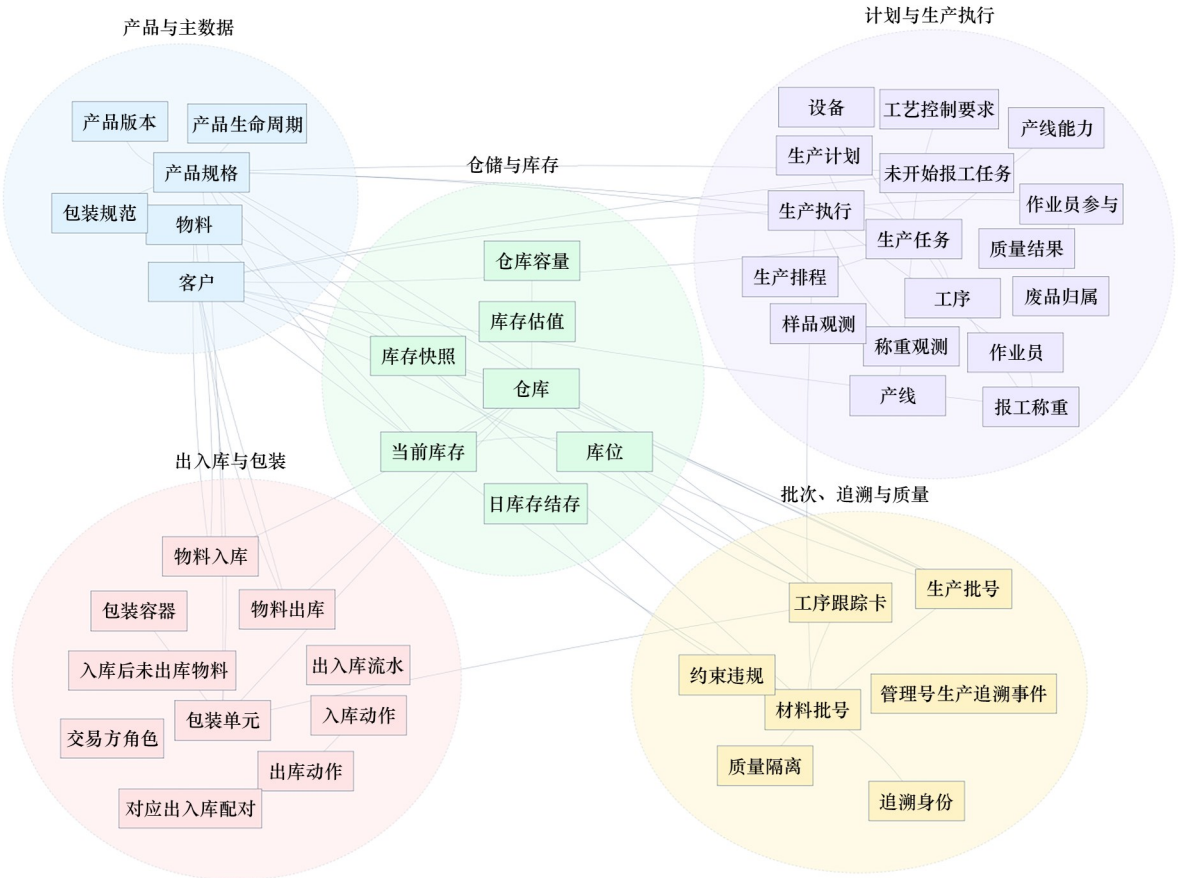


图2　某大型制造企业生产管理场景本体图

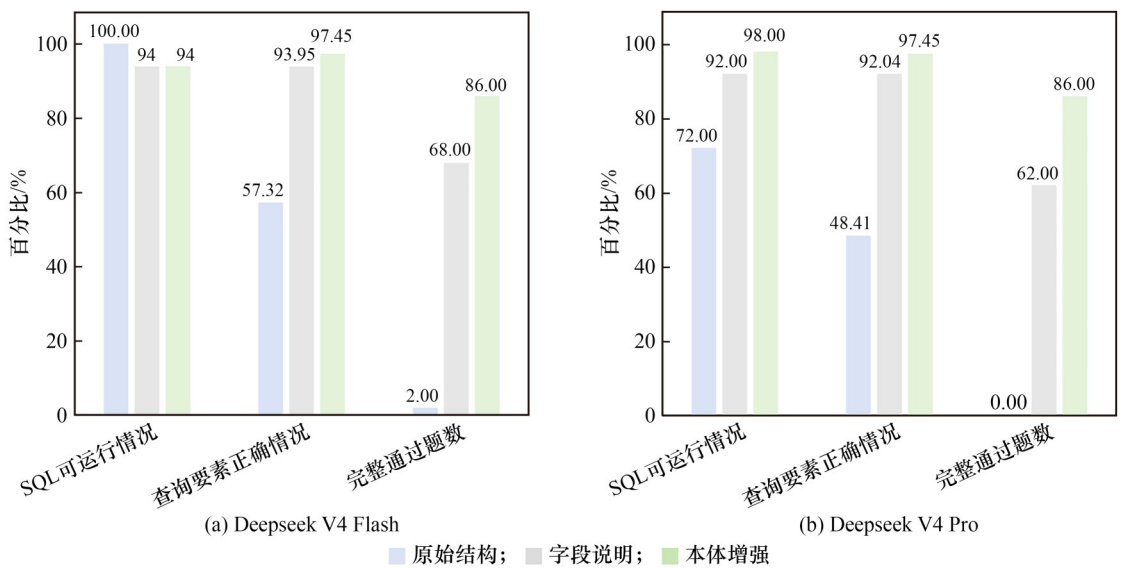


图3　本体增强条件下制造业生产数据查询生成结果

模型的语义组件命中率和整题全通过率均大幅提高，说明表名和字段含义等浅层语义是模型生成可用查询的基础。在此基础上进一步加入本体语义后，查询生成质量继续提升：DeepSeek V4 Flash 的整题全通过率由 68% 提高至 86%，DeepSeek V4 Pro 由 62% 提高至 86%；对应的未完整通过问题数量分别由 16 题降至 7 题、由 19 题降至 7 题。相比语义组件命中率的小幅提升，整题全通过率的改善更能体现本体增强的作用特征。真实业务查询通常需要数据对象、关联路径、指标口径、时间范围和状态规则同时正确，任一环节发生偏移都可能导致查询结果不可用。因此，本体语义的增量价值主要体现为对复杂查询中多项语义约束的协同限定，从而减少局部语义偏移并提高查询结果的整体业务可用性。

总体而言，该应用验证表明，本体增强在制造业数据查询中的作用主要表现为语义契约化和查询约束化。对于跨系统生产数据，查询错误往往并非源于 SQL 语法本身，而是源于业务对象、指标口径、状态集合、时间粒度和关系路径的不一致。通过将这些隐含规则显式化，本体语义能够在查询生成和结果校验环节提供稳定依据，从而提升自然语言查询在真实业务数据环境中的可控性和稳定性。

六、结语

行业智能体正在由知识问答向面向真实业务过程的任务执行演进，其可靠运行取决于自然语言需求、行业知识、数据资源与外部能力之间能否形成稳定的语义联系。本体增强的核心作用是在大语言模型、检索系统和业务工具之上建立共享的领域语义基础，使任务对象、业务条件、可执行操作及运行状态在不同技术环节之间保持一致。本文从行业智能体运行中的语义问题出发，构建了由对象、关系、约束、动作和状态组成的本体语义模型，并在此基础上提出分层技术体系，分析了本体在知识获取、工具调用、任务规划和结果核验过程中的作用。制造业生产数据查询验证表明，在相同字段说明基础上加入本体语义后，两种受测模型的回答效果均显著提升。这些结果说明，本体增强能够通过显式表达业务对象、关系路径、指标口径、时间条件和状态规则，减少多约束查询中的局部语义偏移。

未来，本体增强行业智能体研究应更加重视语义模型的持续演化，以及本体与大语言模型、规划工具和业务系统之间的稳定协同。尤其需要完善动作与状态的形式化表达，建立面向真实行业任务的对照实验与长期运行评价，并逐步形成可复用的建模规范和技术平台。随着相关方法不断成熟，本体有望由静态知识组织工具进一步发展为连接行业数据、业务规则与任务执行过程的基础设施，为复杂业务环境下行业智能体的可信运行和规模化应用提供支撑。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: April 29, 2026; **Revised date:** July 30, 2026

Corresponding author: Lin Chuanwen is a research fellow from Institute of Data Space, Hefei Comprehensive National Science Center. His major research fields include ontology, data element innovation, data assetization, data infrastructure. E-mail: lincw11@foxmail.com

Funding project: Chinese Academy of Engineering project “National Data Resource Management and Scheduling System” (2024-JZ-03); The National Natural Science Foundation of China Project (62441230)

参考文献

[1] 郑逸宁，余镇，李不凡，等. 大语言模型的工具使用综述[J]. 自动化学报，2025，51(11)：2371-2386.
Zheng Y N，Yu Z，Li B F，et al. Survey of tool use in large language models[J]. Acta Automatica Sinica，2025，51(11)：2371-2386.

[2] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition，1993，5(2)：199-220.

[3] Pan S R，Luo L H，Wang Y F，et al. Unifying large language models and knowledge graphs：A roadmap[J]. IEEE Transactions on Knowledge and Data Engineering，2024，36(7)：3580-3599.

[4] Lewis P，Perez E，Piktus A，et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Advances in Neural Information Processing Systems 33，2020：9459-9474.

[5] Albertoni R，Browning D，Cox S J D，et al. Data catalog vocabulary（DCAT）—Version 3[EB/OL].（2024-08-22）[2026-07-21]. https://www.w3.org/TR/2024/REC-vocab-dcat-3-20240822/.

[6] 杜小勇，李曼，王珊. 本体学习研究综述[J]. 软件学报，2006，17(9)：1837-1847.
Du X Y，Li M，Wang S. A survey on ontology learning research[J]. Journal of Software，2006，17(9)：1837-1847.

[7] 任飞亮，沈继坤，孙宾宾，等. 从文本中构建领域本体技术综述[J]. 计算机学报，2019，42(3)：654-676.
Ren F L，Shen J K，Sun B B，et al. A review for domain ontology construction from text[J]. Chinese Journal of Computers，2019，42(3)：654-676.

[8] Motik B，Patel-Schneider P F，Parsia B. OWL 2 web ontology language：Structural specification and functional-style syntax（second

edition)[EB/OL]. (2012-12-11)[2026-07-21]. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.

[9] Navigli R, Velardi P. Learning domain ontologies from document warehouses and dedicated web sites[J]. Computational Linguistics, 2004, 30(2): 151-179.

[10] Cimiano P, Völker J. Text2Onto[M]. Berlin: Springer, 2005: 227-238.

[11] Velardi P, Faralli S, Navigli R. OntoLearn reloaded: A graph-based algorithm for taxonomy induction[J]. Computational Linguistics, 2013, 39(3): 665-707.

[12] Demir C, Baci A, Kouagou N J, et al. Ontolearn—A framework for large-scale OWL class expression learning in Python[J]. Journal of Machine Learning Research, 2025, 26(63): 1-6.

[13] Babaei Giglou H, D'Souza J, Auer S. LLMs4OL: Large language models for ontology learning[C]//The Semantic Web-ISWC 2023, 2023: 408-427.

[14] Jiménez-Ruiz E, Cuenca Grau B. LogMap: Logic-based and scalable ontology matching[C]//The Semantic Web-ISWC 2011, 2011: 273-288.

[15] He Y, Chen J Y, Antonyrajah D, et al. BERTMap: A BERT-based ontology alignment system[C]//The AAAI Conference on Artificial Intelligence, 2022: 4184-4191.

[16] Hertling S, Paulheim H. OLaLa: Ontology matching with large language models[C]//The 12th Knowledge Capture Conference 2023, 2023: 131-139.

[17] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering[C]//The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 6769-6781.

[18] Wu L, Petroni F, Josifoski M, et al. Scalable zero-shot entity linking with dense entity retrieval[C]//The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 6397-6407.

[19] Li B Z, Min S, Iyer S, et al. Efficient one-pass end-to-end entity linking for questions[C]//The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 6433-6441.

[20] Sung M, Jeon H, Lee J, et al. Biomedical entity representations with synonym marginalization[C]//The 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3641-3650.

[21] Liu F Y, Shareghi E, Meng Z Q, et al. Self-alignment pretraining for biomedical entity representations[C]//The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 4228-4238.

[22] Yahya M, Berberich K, Elbassuoni S, et al. Natural language questions for the web of data[C]//The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 379-390.

[23] Dubey M, Banerjee D, Chaudhuri D, et al. EARL: Joint entity and relation linking for question answering over knowledge graphs[C]//The Semantic Web-ISWC 2018, 2018: 108-126.

[24] Guo J Q, Zhan Z C, Gao Y, et al. Towards complex text-to-SQL in cross-domain database with intermediate representation[C]//The 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4524-4535.

[25] Wang B L, Shin R, Liu X D, et al. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers[C]//The 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7567-7578.

[26] Yih W T, Chang M W, He X D, et al. Semantic parsing *via* staged query graph generation: Question answering with knowledge base[C]//The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1321-1331.

[27] Massai L. Evaluation of semantic relations impact in query expansion-based retrieval systems[J]. Knowledge-Based Systems, 2024, 283: 111183.

[28] Abe K, Takeoka K, Kato M P, et al. LLM-based query expansion fails for unfamiliar and ambiguous queries[C]//The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025: 3035-3039.

[29] Xia Y, Wu J D, Kim S, et al. Knowledge-aware query expansion with large language models for textual and relational retrieval[C]//The 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025: 4275-4286.

[30] Das S, Sundara S, Cyganiak R. R2RML: RDB to RDF mapping language[EB/OL]. (2012-09-27)[2026-07-21]. https://www.w3.org/TR/2012/REC-r2rml-20120927/.

[31] Calvanese D, De Giacomo G, Lembo D, et al. Tractable reasoning and efficient query answering in description logics: The DL-lite family[J]. Journal of Automated Reasoning, 2007, 39(3): 385-429.

[32] W3C. OWL 2 web ontology language profiles: Second edition [EB/OL]. (2012-12-11)[2026-07-21]. https://www.w3.org/TR/2012/REC-owl2-profiles-20121211/.

[33] Calvanese D, Cogrel B, Komla-Ebri S, et al. Ontop: Answering SPARQL queries over relational databases[J]. Semantic Web, 2016, 8(3): 471-487.

[34] Harris S, Seaborne A. SPARQL 1.1 Query language [EB/OL]. (2013-03-21)[2026-07-21]. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[35] Kharlamov E, Hovland D, Skjæveland M G, et al. Ontology based data access in statoil[J]. Journal of Web Semantics, 2017, 44: 3-36.

[36] Kharlamov E, Mailis T, Mehdi G, et al. Semantic access to streaming and static data at Siemens[J]. Journal of Web Semantics, 2017, 44: 54-74.

[37] Kalaycı E G, Grangel González I, Lösch F, et al. Semantic integration of Bosch manufacturing data using virtual knowledge graphs[C]//The Semantic Web—ISWC 2020, 2020: 464-481.

[38] Lebo T, Sahoo S, Mcguinness D. PROV-O: The PROV ontology [EB/OL]. (2013-04-30)[2026-07-21]. https://www.w3.org/TR/2013/REC-prov-o-20130430/.

[39] W3C. Data on the web best practices: Data quality vocabulary [EB/OL]. (2016-12-15)[2026-07-21]. https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/.

[40] Iannella R, Villata S. ODRL information model 2.2 [EB/OL]. (2018-02-15)[2026-07-21]. https://www.w3.org/TR/2018/REC-odrl-model-20180215/.

[41] Sharma K, Kumar P, Li Y Q. OG-RAG: Ontology-grounded retrieval-augmented generation for large language models[C]//

The 2025 Conference on Empirical Methods in Natural Language Processing，2025：32962-32981.

[42] Schick T，Dwivedi-Yu J，Dessi R，et al. Toolformer：Language models can teach themselves to use tools[C]//Advances in Neural Information Processing Systems 36，2023：68539-68551.

[43] Qin Y J，Liang S H，Ye Y N，et al. ToolLLM：Facilitating large language models to master 16000+ real-world APIs[C]//The Twelfth International Conference on Learning Representations（ICLR），2024.

[44] Martin D，Burstein M，Hobbs J，et al. OWL-S：Semantic markup for web services[EB/OL].（2004-11-22）[2026-07-21]. https://www.w3.org/submissions/2004/SUBM-OWL-S-20041122/.

[45] Farrell J，Lausen H. Semantic annotations for WSDL and XML schema[EB/OL].（2007-08-28）[2026-07-21]. https://www.w3.org/TR/2007/REC-sawsdl-20070828/.

[46] Yuan S Y，Song K T，Chen J J，et al. EASYTOOL：Enhancing LLM-based agents with concise tool instruction[C]//The 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics：Human Language Technologies，2025：951-972.

[47] Shi Z L，Wang Y H，Yan L Y，et al. Retrieval models aren't tool-savvy：Benchmarking tool retrieval for large language models[C]//The Association for Computational Linguistics：ACL 2025，2025：24497-24524.

[48] Klusch M，Fries B，Sycara K. OWLS-MX：A hybrid semantic web service matchmaker for OWL-S services[J]. Web Semantics：Science，Services and Agents on the World Wide Web，2009，7（2）：121-133.

[49] Yao S Y，Zhao J，Yu D，et al. ReAct：Synergizing reasoning and acting in language models[C]//The Eleventh International Conference on Learning Representations（ICLR 2023），2023：1-21.

[50] Wu Q Z，Liu W，Luan J，et al. ToolPlanner：A tool augmented LLM for multi granularity instructions with path planning and feedback[C]//The 2024 Conference on Empirical Methods in Natural Language Processing，2024：18315-18339.

[51] Lu J R，Holleis T，Zhang Y Z，et al. ToolSandbox：A stateful，conversational，interactive evaluation benchmark for LLM tool use capabilities[C]//The Association for Computational Linguistics：NAACL 2025，2025：1160-1183.

[52] Sirin E，Parsia B，Wu D，et al. HTN planning for web service composition using SHOP2[J]. Journal of Web Semantics，2004，1（4）：377-396.

[53] Hao Y L，Chen Y C，Zhang Y，et al. Large language models can solve real-world planning rigorously with formal verification tools[C]//The 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics：Human Language Technologies，2025：3434-3483.

[54] Tabassi E. Artificial intelligence risk management framework（AI RMF 1.0）：NIST AI 100-1[R]. Gaithersburg：National Institute of Standards and Technology，2023：1-42.

[55] Glimm B，Horrocks I，Motik B，et al. HermiT：An OWL 2 reasoner[J]. Journal of Automated Reasoning，2014，53（3）：245-269.

[56] Horrocks I，Patel-Schneider P F，Boley H，et al. SWRL：A semantic web rule language combining OWL and RuleML[EB/OL].（2004-05-21）[2026-07-21]. https://www.w3.org/submissions/2004/SUBM-SWRL-20040521/.

[57] Knublauch H，Kontokostas D. Shapes constraint language（SHACL）[EB/OL].（2017-07-20）[2026-07-21]. https://www.w3.org/TR/2017/REC-shacl-20170720/.

[58] Chytas A，Gavriilides G，Kapetanakis A，et al. OpenPVSignal knowledge graph：Pharmacovigilance signal reports in a computationally exploitable FAIR representation[J]. Drug Safety，2025，48（4）：425-436.

[59] Arya V，Colace F，Lombardi M，et al. DTSafe-O：A semantic digital twin ontology for construction safety[J]. International Journal on Semantic Web and Information Systems，2025，21（1）：1-36.

[60] Thorne J，Vlachos A，Christodoulopoulos C，et al. FEVER：A large-scale dataset for fact extraction and VERification[C]//The 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2018：809-819.

[61] Min S，Krishna K，Lyu X X，et al. FActScore：Fine-grained atomic evaluation of factual precision in long form text generation[C]//The 2023 Conference on Empirical Methods in Natural Language Processing，2023：12076-12100.

[62] Gao T Y，Yen H，Yu J T，et al. Enabling large language models to generate text with citations[C]//The 2023 Conference on Empirical Methods in Natural Language Processing，2023：6465-6488.

[63] Wang X，Tan M J，Jin Q，et al. MedCite：Can language models generate verifiable text for medicine?[C]//The Association for Computational Linguistics：ACL 2025，2025：18891-18913.

[64] Gao L Y，Dai Z Y，Pasupat P，et al. RARR：Researching and revising what language models say，using language models[C]//The 61st Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers），2023：16477-16508.

[65] Es S，James J，Espinosa Anke L，et al. RAGAs：Automated evaluation of retrieval augmented generation[C]//The 18th Conference of the European Chapter of the Association for Computational Linguistics：System Demonstrations，2024：150-158.

[66] Schlegel M，Sattler K U. Capturing end-to-end provenance for machine learning pipelines[J]. Information Systems，2025，132：102495.