LDP：面向多智能体大语言模型系统的身份感知协议

Sunil Prakash[1]

[1]印度商学院，印度海得拉巴，sunil_prakash_pgpmax2026@isb.edu

arXiv:2603.08852v1 [cs.AI] 9 Mar 2026

摘要

随着多智能体人工智能系统日益复杂，连接它们的协议越来越制约其能力。当前的智能体间协议（例如谷歌的 A2A 和 Anthropic 的 MCP）并未将模型级属性作为一等协议原语公开——智能体通过名称、描述和技能列表来标识，但有效委派所必需的属性（模型身份、推理特征、质量校准、成本特征）仍不属于协议接口。我们提出**LLM 委托协议**（LDP），这是一种 AI 原生通信协议，可将这些属性提升为一等要素。LDP 引入了五种关键机制：(1) 携带模型系列、质量提示和推理配置的丰富委托代理身份卡；(2) 支持自动协商与回退的渐进式载荷模式；(3) 支持多轮委派和持久上下文的受治理会话；(4) 追踪置信度和验证状态的结构化来源信息；以及 (5) 在协议层强制执行安全边界的信任域。我们将 LDP 实现为 JAMJET 智能体运行时的插件适配器，并使用本地 Ollama 模型和 LLM-as-judge 评估，将其与 A2A 及随机基线进行初步实证研究。在实验中，身份感知路由通过更好的受托智能体专业化，使简单任务的延迟降低了 $\sim$12$\times$；尽管在我们规模较小的委托池中，它并未比技能匹配提高总体质量。语义框架载荷将令牌数量减少了 37%（p=0.031、d=$-$0.7），且未观察到质量损失；与无状态重新调用相比，受治理会话在 10 轮对话中消除了 39% 的令牌开销。最值得注意的是，噪声溯源会使综合质量降至无来源信息基线以下，这表明未经验证的置信度元数据有害。对信任域和回退链的模拟协议分析显示，它们在攻击检测（96% 对 6%）和故障恢复（完成率 100% 对 35%）方面具有架构优势，不过这些结果反映的是协议设计属性，而非实证攻击场景。本文贡献了一种协议设计、一套参考实现，以及初步证据，表明 AI 原生协议原语能够实现更高效且可治理的委派。

关键词： 智能体协议、多智能体系统、大语言模型委派、载荷协商、AI 互操作性、信任域

1 引言

大语言模型（LLM）在生产系统中的快速部署，催生了多智能体架构。在此类架构中，专用 AI 代理协作解决复杂任务 [Wu 等人， 2023，Hong 等人， 2023，Li 等人， 2023]。这些系统需要通信协议，使代理能够彼此发现、协商交互条款、委派任务并综合结果。

目前已出现两种主流范式。谷歌的智能体对智能体协议（A2A） [谷歌， 2025]提供面向服务的接口，代理可通过该接口公开“代理卡”，其中包含名称、描述和技能列表。Anthropic 的模型上下文协议（MCP） [Anthropic， 2024]专注于工具级集成，使主机能够调用工具服务器上的函数。这两种协议都刻意省略了模型级元数据，这一设计选择旨在兼顾简洁性和通用性。

然而，这种不透明性丢弃了有效委派所需的关键信息。当路由器必须在一个针对分类进行优化的30亿参数模型和一个80亿参数

模型（具备深度推理能力）之间做出选择时，仅知道它们的技能名称（“分类”“推理”）是不够的。路由器无法评估质量与成本之间的权衡、协商通信格式、验证输出的来源信息，也无法维护受治理的多轮次上下文。

我们认为，*AI 原生*的智能体间协议——即暴露模型身份、协商通信丰富度、实施治理并追踪来源信息——能够实现比不暴露这些属性的协议更高效、更可治理的委派。为探讨这一论点，我们提出**LLM 委托协议**（LDP），一种附带参考实现与初步实证研究的协议设计。LDP 围绕以下三项观察，从第一性原理出发进行设计：

1. **AI 委托方拥有丰富且可操作的身份**。模型系列、参数数量、推理特征、上下文窗口和成本特征，都是影响委派决策的属性。以结构化身份卡暴露这些属性，可实现质量感知路由。

2. **通信开销是一阶成本**。词元消耗直接决定延迟与货币成本。协商紧凑的负载格式（结构化语义帧与冗长的自然语言）可带来显著的效率提升。

3. **治理需要协议支持**。多轮会话、来源追踪、信任边界和策略执行，无法可靠地事后附加到无状态、不透明的协议之上。

LDP 被实现为 JamJet 智能体运行时 [JamJet, 2024], 作为外部插件，与现有的 A2A 和 MCP 适配器一同注册，无需对宿主系统进行任何修改。我们围绕六个研究问题，将 LDP 与 A2A 和随机路由基线进行比较，委托方使用本地 Ollama [Ol-lama, 2024] 模型，并使用 Google Gemini [谷歌DeepMind, 2023]作为大语言模型评审器。我们的评估结合了实证实验（路由、载荷、来源信息、会话）与模拟协议分析（安全性、回退）。

贡献。
• LDP 协议规范：原生适用于 AI 的身份卡片、支持协商与回退的渐进式载荷模式、受治理会话、结构化来源信息以及信任域（§3）。
• 作为 JamJet 协议适配器的参考实现（§4）。
• 一项初步实证研究及消融分析表明，身份元数据和结构化载荷可提升路由质量与通信效率（§6），并对安全性和回退特性进行模拟协议分析。

2 背景与相关工作

2.1 智能体通信协议

智能体通信在多智能体系统中有着悠久的历史。知识查询与操纵语言（KQML）[Finin 等人, 1994] 以及 FIPA 智能体通信语言（FIPA- ACL）[FIPA, 2002] 在 20 世纪 90 年代和 21 世纪初为软件代理之间基于言语行为的消息传递奠定了基础。这些协议定义了言语行为（告知、请求、提议），但早于大语言模型时代，且未涉及模型特有的属性。

现代基于大语言模型的智能体协议包括：

智能体对智能体协议（A2A）。谷歌的 A2A [谷歌, 2025] 定义了一种面向服务的协议，用于实现智能体互操作性。智能体会发布代理卡，其中包含名称、描述、版本、URL 和技能列表。通信采用无状态方式：客户端发送任务并接收结果。A2A 提供了简洁、通用的接口，但有意不涵盖模型级元数据、质量信号、会话支持、来源信息和信任域强制执行。

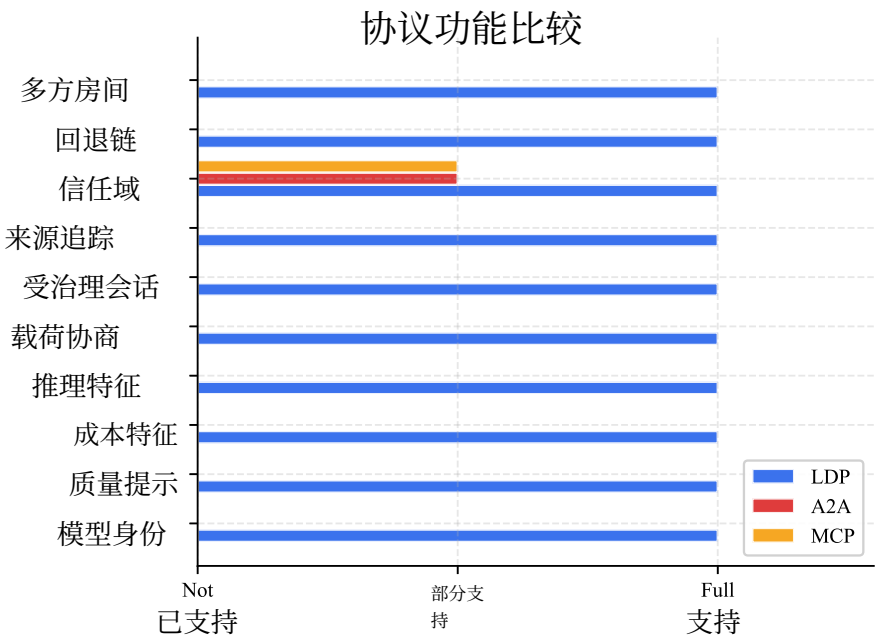

图 1：各智能体通信协议的功能比较。LDP 完全支持所有原生 AI 功能，而 A2A 和 MCP 仅提供部分传输层安全支持。

模型上下文协议（MCP）。 Anthropic 的 MCP [Anthropic， 2024] 提供了一种标准化接口，用于将 AI 助手连接到外部工具和数据源。MCP 运行于工具层——主机调用服务器上的函数——而非智能体层。它与 A2A 和 LDP 互为补充，但不涉及智能体间委派。

智能体网络协议（ANP）。ANP [智能体网络协议， 2024]侧重于代理之间的发现和网络通信，提供目录层，但通信语义有限。

LDP 基于 A2A 模式（服务发现、任务提交、结果检索）构建，并通过原生 AI 感知能力对其进行扩展（§3）。

2.2 多智能体框架

若干框架负责编排多智能体协作。AutoGen [Wu 等人， 2023]提供了一种具有灵活智能体拓扑结构的基于对话的框架。MetaGPT [Hong 等人， 2023] 为大语言模型智能体分配软件工程角色。CAMEL [Li 等人， 2023] 研究用于智能体协作行为的角色扮演。CrewAI [CrewAI， 2024] 强调带有任务委派的基于角色的智能体团队。

必须区分多智能体系统中的三个层次：协议（智能体如何通信）、框架（智能体如何被编排），以及路由器（选择哪个智能体）。这些框架定义了编排模式，但在协议层依赖临时通信（通常是函数调用或字符串传递）。LDP 提供了此类框架所缺乏的通信层——其中任何一个都可以将 LDP 用作其智能体间协议，从而获得身份感知路由、载荷协商和治理。一个自然的问题是，是否可以通过自定义元数据字段扩展 A2A，而不采用新协议；我们将在§7.4中直接讨论。

2.3 大语言模型路由与委派

将查询路由至适当的模型是一个活跃的研究领域。FrugalGPT [Chen 等人， 2023] 提出通过级联大语言模型调用，在保持质量的同时降低成本。RouterLLM [Ong 等人， 2024] 训练一个轻量级路由器，根据查询难度将查询分发给强模型或弱模型。这些方法

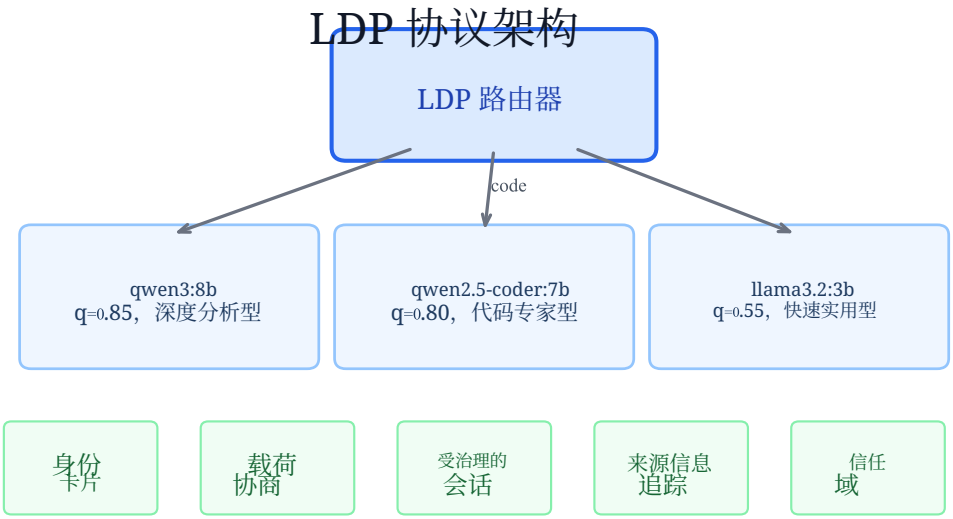

图 2：LLM 委托协议的架构。LDP 路由器利用丰富的身份卡片（模型系列、质量提示、推理配置）进行智能路由决策——将简单任务发送给轻量级模型，将困难任务发送给能力更强的模型。五种协议机制协同运作：身份卡片、载荷协商、受治理会话、来源追踪和信任域。

优化应调用哪个模型，但没有解决选定模型后如何与其通信的问题。LDP 的身份模型提供了路由器所需的元数据（质量提示、成本概况、能力清单），并将其作为协议原语，而非独立系统。

2.4 面向大语言模型智能体的结构化通信

Du 等人 [Du 等人， 2023] 发现，多智能体辩论能够提升事实准确性。Liang 等人 [Liang 等人，2023] 发现，多样化的智能体角色能够提升推理质量。这些研究表明，智能体 之间的通信方式会影响输出质量——这支持了 LDP 的核心论点：更丰富的协议语义（结构化载荷、身份上下文）能够带来更好的结果。

3 LLM 委托协议

LDP 的设计基于一个核心前提：智能体通信协议应认识到，其参与者是具有可测量且彼此不同属性的人工智能模型。下面介绍关键的协议机制。

3.1 AI 原生身份模型

A2A 的代理卡公开了七个字段：名称、描述、版本、URL、技能、身份验证方案和支持的协议。LDP 通过**委托身份卡**扩展了这一机制，其中包含 20+个字段，分为四类：

核心身份。 委托代理_ID、主体_ID、模型_系列、模型_名称、模型_版本、运行时_版本、权重_指纹、端点_地址。

信任与安全 安全性。 信任_域名、公钥_、管辖权、数据_处理_策略 .

功能。 上下文_窗口、支持的模态_、支持的语言_、分词器_指纹
每个能力条目都包含质量_提示（0–1 的连续得分）、延迟_提示_毫秒_p50，以及成本提示元数据。_

行为特征。 推理_档案（定性描述符，例如“深度分析型”或“快速实用型”）以及成本_档案（“低”“中等”“高”）。

这种丰富的身份信息支持元数据感知路由：路由器可以将简单的分类任务交给快速、低成本的30亿参数模型，并将困难的推理任务留给速度较慢但能力更强的80亿参数模型——而借助 A2A 的仅技能名称匹配，这类决策是不可能实现的。

3.2 渐进式载荷模式

令牌数量是大语言模型系统中的首要成本因素。LDP 定义了六种效率逐步提高的载荷模式：

- **模式 0 — 文本**。 自然语言。通用兼容，最易于审计，令牌数量最高。
- **模式 1 — 语义帧**。 带有显式字段的类型化结构化 JSON（任务_类型、指令 n, 预期_输出_格式）。在保持人类可读的同时减少冗长。
- **模式 2 — 嵌入提示**。 作为检索或路由信号附加的向量表示。
- **模式 3 — 语义图**。 用于规划与形式推理的结构化关系表示。
- **模式 4 — 潜在胶囊**。 兼容模型之间经压缩的机器原生语义数据包。
- **模式 5 — 缓存切片**。 在高度兼容的模型实例之间传输执行状态。

模式 2–3 已完整定义，但尚未经过实证评估。模式 4–5 需要模型层级的访问权限，而当前应用程序接口并非都能提供这种权限。我们将模式 2–5 作为未来工作，并将评估重点放在模式 0–1 上；这两种模式足以证明载荷效率这一论点。

协商与回退。 在会话建立期间，委托方会协商双方共同支持的最高级模式。如果较高级模式在交换过程中失效（例如，出现模式验证错误），协议会自动回退：模式 $N \to$ 模式 $N-1 \to \cdots \to$ 模式 0。每个委托方**都必须** 至少支持模式 0（文本），以确保通信不会彻底中断。

3.3 受治理会话

A2A 是无状态的：每次任务提交都是独立的。对于多轮委派（迭代精炼、渐进式分析、验证链），每次请求都必须重新传输上下文，因此会产生二次方级的令牌开销。

LDP 引入了**受治理会话**——经过协商、具有明确生命周期的持久化上下文：

1. HELLO：调用方声明身份。
2. CAPABILITY_MANIFEST：被调用方回复支持模式和约束。
3. SESSION_PROPOSE：调用方提出会话参数（载荷模式、延迟目标、成本预算、隐私约束、审计级别）。
4. 会话_接受：被调用方确认，也可能会调整参数。

5. TASK_提交 / 任务_更新 / 任务_结果：会话上下文中的任务交换 .

会话会在服务器端保留上下文，无需重新发送对话历史。预算跟踪、隐私范围界定和质量等级协商会在各轮次之间持续保留。

3.4 结构化来源信息

每个 LDP 任务结果都包含结构化溯源元数据：

```
{
  "产生者_": "delegate:qwen3-8b", "模型_版本": "qwen3-8b-2026.01", "所


}
```

这使下游使用者能够根据来源可靠性为输出赋予不同权重——合成器综合三位委托方的意见时，可以优先采用置信度较高且状态已验证的意见。A2A 除任务完成状态外不提供任何来源信息。

3.5 信任域

LDP 将信任域定义为安全边界，在这些边界内，身份、策略和传输保障均得到认可。安全性通过多个层级实施：

1. **消息级别：** 逐条消息的签名、随机数和重放保护。

2. **会话层：** 会话建立期间进行信任域兼容性检查。

3. **策略层：** 策略引擎根据可配置规则（能力范围、司法管辖区合规性、成本限制）验证每项任务。

A2A 仅依赖传输层身份验证（持有者令牌），无法区分信任域、检测能力提升，也无法强制执行逐任务策略。

3.6 多方房间模型

除了点对点委派，LDP 还规定了一种**房间模型**，并定义了正式角色：主席（协调）、记录员（状态管理）、验证者（输出检查）、仲裁者（冲突解决）和参与者（专业代表）。这使得结构化的群组协作模式成为可能，而 A2A 的点对点拓扑无法表达这些模式。房间模型属于协议规范的一部分，但本文未对其进行评估；多方协调的实证验证将留待未来工作。

4 实现

我们将 LDP 实现为 JAMJET 智能体运行时 [JamJet 的外部插件， 2024], 这是一个基于 Rust 构建的生产级系统，用于编排人工智能智能体工作流。JAMJET 提供了 ProtocolAdapter 特征，其方法涵盖发现、调用、流式处理、状态检查和取消。现有适配器实现了 MCP 和 A2A。

4.1 插件架构

LDP 通过一次函数调用在运行时注册：

```
let mut registry  =default_协议_registry(); register_ldp(&
```

基于 URL 的分派会将以 ldp:// 为前缀的地址路由至 LDP 适配器。无需修改 JAM-JET 的得分；LDP 是纯外部依赖。

4.2 身份卡扩展

JAMJET 的 AgentCard 结构包含一个 labels: HashMap<String, String> 字段。LDP 身份字段存储为 ldp.* 标签（例如，ldp.model_family、ldp.reasoning_profile、ldp.cost_profile），因此无需更改架构即可通过 JAMJET 现有的路由和调度基础设施访问这些字段。

4.3 会话管理

会话生命周期（协商、缓存、拆除）封装在适配器中。从 JAMJET 的角度来看，每次 invoke() 调用都是无状态的。在内部，适配器会维护一个以 (url, session_config) 为键的会话缓存，按需创建新会话，并在多轮交互中复用已有会话。

4.4 模型适配器

为支持实验评估，我们为 JAMJET 的 ModelAdapter trait 实现了两个新的模型适配器：用于本地推理的**Ollama 适配器**（支持通义千问、Llama、Gemma 和 Phi 模型系列），以及用于基于云端的评审评估的**Google Gemini 适配器**。两者均支持结构化输出、OpenTelemetry 跟踪和基于前缀的模型路由（例如，qwen* → Ollama，gemini-*→ Google）。

5 实验设置

我们围绕六个研究问题，将 LDP 与基线进行比较评估。所有实验均使用本地 Ollama 模型进行委托推理（API 成本为零），并使用 Google Gemini 2.5 Flash 作为大语言模型评审。

5.1 研究问题

RQ1 — 路由质量。 与技能匹配和随机选择相比，AI 原生身份是否能提升委派路由质量?

RQ2 — 载荷效率。 语义框架载荷能否在保持质量的同时降低通信成本?

研究问题3 — 来源信息价值。 结构化来源信息能否提升多源综合中的下游决策质量?

RQ4 — 会话效率。 与多轮委派中无状态的重新调用相比，受治理会话能否降低令牌开销?

RQ5 — 安全边界（模拟）。 信任域能否检测出持有者令牌认证未能发现的未授权委派尝试?

RQ6 — 回退可靠性（模拟）。 载荷模式回退是否能在通信故障下提升任务完成率?

5.2 委托池

三个本地 Ollama 委托方模拟了一个具有不同质量、成本和专业化程度的异构池：

• qwen3:8b — 高质量（q=0.85），推理与分析专家，中位延迟 5 秒。

• qwen2.5-coder:7b — 中等质量（q=0.80），代码专家，中位延迟 4 秒。

• llam a3.2:3b — 较低质量（q=0.55），快速分类与提取，中位延迟 1 秒 .

所有委托方均通过 Ollama 在单台 Apple Silicon 机器（36GB 内存）上运行，确保零 API 成本与可复现条件。

5.3 基线

所有条件使用相同的委托池、任务集和评审器。这些条件在两个维度上有所不同：路由策略（如何选择委托方）以及提示条件化 （委托方接收何种上下文）。我们明确说明这一点，是因为收益可能来自其中之一或两者兼有；消融研究（§6.7）将分离这些效应。

LDP 基线。 使用身份卡字段（质量提示、推理特征、成本配置）进行元数据感知路由。系统提示包含委托者身份上下文。代表完整的 LDP 协议。

A2A 基线。 仅进行技能名称匹配（无质量/成本元数据）。通用系统提示（“你是一名 AI 助手”）。代表标准的 A2A 基于能力的发现。

随机基线。 使用通用提示进行均匀随机委托方选择。为路由质量提供下限。

5.4 LLM 评判评估

任务输出由 Gemini 2.5 Flash 从三个维度进行评估：质量 （清晰度和实用性）、正确性（事实准确性）和完整性 （任务覆盖度）。每个维度的得分为 1–10 分；总得分为加权平均值：

$$\text{score}_{\text{overall}} = 0.3 \times \text{quality} + 0.4 \times \text{correctness} + 0.3 \times \text{completeness} \tag{1}$$

评审器会接收原始任务提示、委托方输出，以及（可选的）参考答案。我们采用广泛使用的 LLM 评判方法 [Zheng 等人， 2023]。

5.5 任务生成

对于**RQ1**，在三个难度级别（简单、中等、困难；每个级别 10 个）下生成 30 个任务，涵盖分类、推理、分析、编码和数学领域。对于**RQ2**，每个条件下设置 20 个任务，用于测试三种载荷类型：推理交接、上下文传递和验证任务。对于**研究问题3**，设置 15 个多源综合任务，要求综合三位委托方的意见。对于**RQ4**，在基于会话（LDP）和无状态（A2A）两种条件下，分别测试 3、5 和 10 轮的 10 个多轮委托场景（共 n=60 个）。**RQ5** 模拟四种攻击类型下的 100 个安全注入场景。**RQ6** 模拟四种故障类型下的 40 个通信故障。

5.6 统计方法

我们报告均值、标准差和 95% 置信区间。统计显著性通过曼-惠特尼 U 检验进行评估（该检验为非参数检验，适用于有序 LLM 评判分数）。效应量报告为 Cohen’s d。我们认为 $p < 0.05$ 具有统计显著性。

表 1：不同协议条件下的路由质量比较（RQ1）。质量评分为 LLM评审评分（1–10）。所有委托方均通过 Ollama 在本地运行。

条件	质量	延迟（秒）	令牌	n
LDP	6.80 ± 3.62	85.6	1379	30
A2A	7.43 ± 3.54	67.4	1227	30
随机	6.95 ± 3.21	42.2	826	30

表 2：载荷模式效率比较（RQ2）。与原始文本相比，语义帧使令牌数量减少 37%，延迟降低 42%，且未观察到质量损失（$p = 0.96$）。令牌减少具有统计显著性（$p = 0.031$， $d = -0.7$）。

Mode	令牌	延迟（秒）	质量	信息保留
文本（模式 0）	1215 ± 751	24.1	5.54 ± 4.05	5.85 ± 4.26
语义框架（模式 1）	765 ± 510	14.0	5.70 ± 2.93	5.50 ± 3.03
A2A JSON	1128 ± 804	23.4	5.47 ± 3.98	5.75 ± 3.89

6 结果

6.1 RQ1：路由质量

表 1 展示了总体路由质量。A2A 达到最高的总体质量（7.43 ± 3.49），其次是随机（6.95 ± 3.22）和 LDP（6.80 ± 3.60）。**两两比较的差异均不具有统计显著性**（p=0.56，LDP 与 A2A 之间的 Mann-Whitney U 检验）；鉴于 LLM 评审分数固有的高方差、中等样本量（n=30）以及较小的委托池（n=3），这一结果并不意外。随机基线的竞争力表现反映了委托池较小这一特点：随机选择有 $1\,3$ 的概率选中最优委托代理。

LDP 的总体质量略低于 A2A，这是一个重要且诚实的结果。这表明，在我们的设置中，身份感知路由并不会自动提升输出质量。不过，按任务难度考察结果，可以发现情况更加复杂。

图 3 展示了不同难度级别下的质量。三种条件在简单任务（∼9.1–9.6）和中等难度任务（∼7.7–8.6）上的表现相当。在困难任务上，所有条件都表现不佳（3.3–4.1），其中 A2A 略优于 LDP（4.1对比 3.3）。这些按难度划分的差异均未达到统计显著性，每个单元格的样本量为 n=10。

质量结果表明，在我们的小型委托池设置中，LDP 的元数据感知路由并未比 A2A 更简单的技能匹配提升输出质量。我们坦诚地报告这一负面发现，并将其归因于两个因素：(1) 委托池过小（n=3），路由策略的复杂程度无关紧要——所有方法选择的委托方都相似；(2) 对于这些任务，身份信息增强提示相较于通用提示可能没有足够的优势。

LDP 路由能带来明显价值的方面：延迟效率。在我们的实验中，身份感知路由的主要优势是基于专业化的延迟降低，而非质量提升。

图4 展示了延迟维度。LDP 的简单任务延迟为 **2.9秒 vs. 34.8秒**（相对 A2A ∼12×更快），因为 LDP 将简单任务路由到轻量级 `llama3.2:3b` 模型，而 A2A 的技能匹配会选择更重的模型。这一延迟优势是元数据感知路由的直接结果，且独立于提示条件化——它纯粹源于该协议将任务难度与委托代理成本特征相匹配的能力。

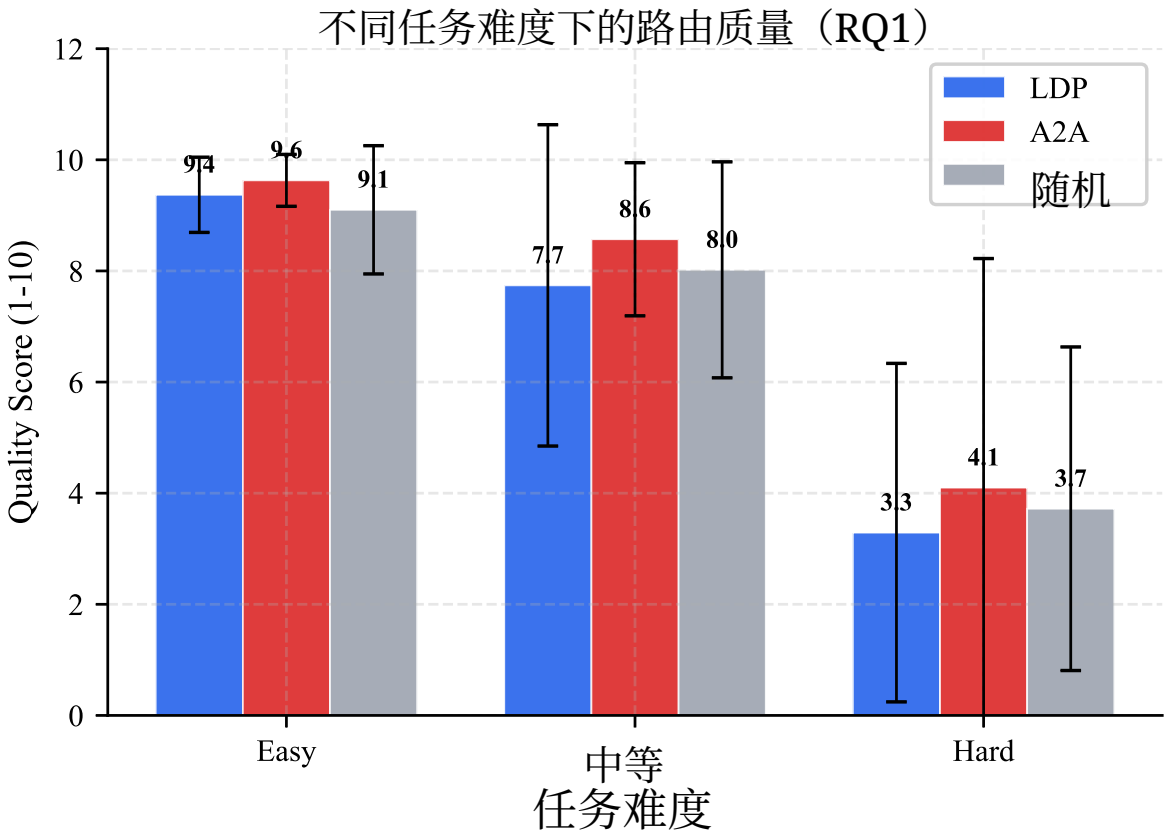

图 3：不同任务难度下的路由质量（**RQ1**）。所有条件的表现相近；总体而言，A2A 略优于 LDP，但差异均不具有统计显著性（每个单元格 n=10）。误差线表示 ±1 标准差。

6.2 RQ2：载荷效率

表 2 与图 5 表明，语义帧（模式 1）可将令牌数量减少 **37%** （相比原始文本：765 vs. 1,215 令牌），并相比 A2A JSON 减少 32%（765 vs. 1,128）。这一降幅具有统计显著性（p=0.031，Mann-Whitney U检验；d=−0.7，大效应）。

延迟与令牌数量相关：语义帧快 42%（14.0 秒 vs. 24.1 秒）。质量相当或略高（文本为 5.70 vs. 5.54；p=0.96，不显著），这表明结构化提示有助于模型集中注意力，同时不会损失信息。

A2A 的 JSON格式相比原始文本带来的收益很小（令牌减少 7%），因为它缺乏类型化语义框架的结构紧凑性。关键区别在于，语义框架使用类型化字段（任务_类型、指令、预期_输出_格式），从而消除了冗长的自然语言表述；而 A2A 的 JSON 只是在 JSON封装中包裹了同样冗长的文本。

6.3 研究问题3：来源信息价值

表 3 和图 6呈现了一个细致的发现。准确溯源与无来源信息带来了相近的决策质量（7.65 vs. 7.85；p=0.47，不显著）——无论合成器是否了解委托方置信度，或是平等对待所有来源，其表现都相当。

然而，**噪声溯源会显著降低质量**： 6.85 ± 2.66，其方差几乎是准确溯源的两倍（±1.21）。当某个委托代理的置信度被人为提高到 0.99 并标记为已验证时，合成器会过度重视其输出，从而做出比完全没有来源信息时更差的决策。

这一 论证支持 LDP 的结构化验证字段：来源信息只有在可信时才有价值 *t*-

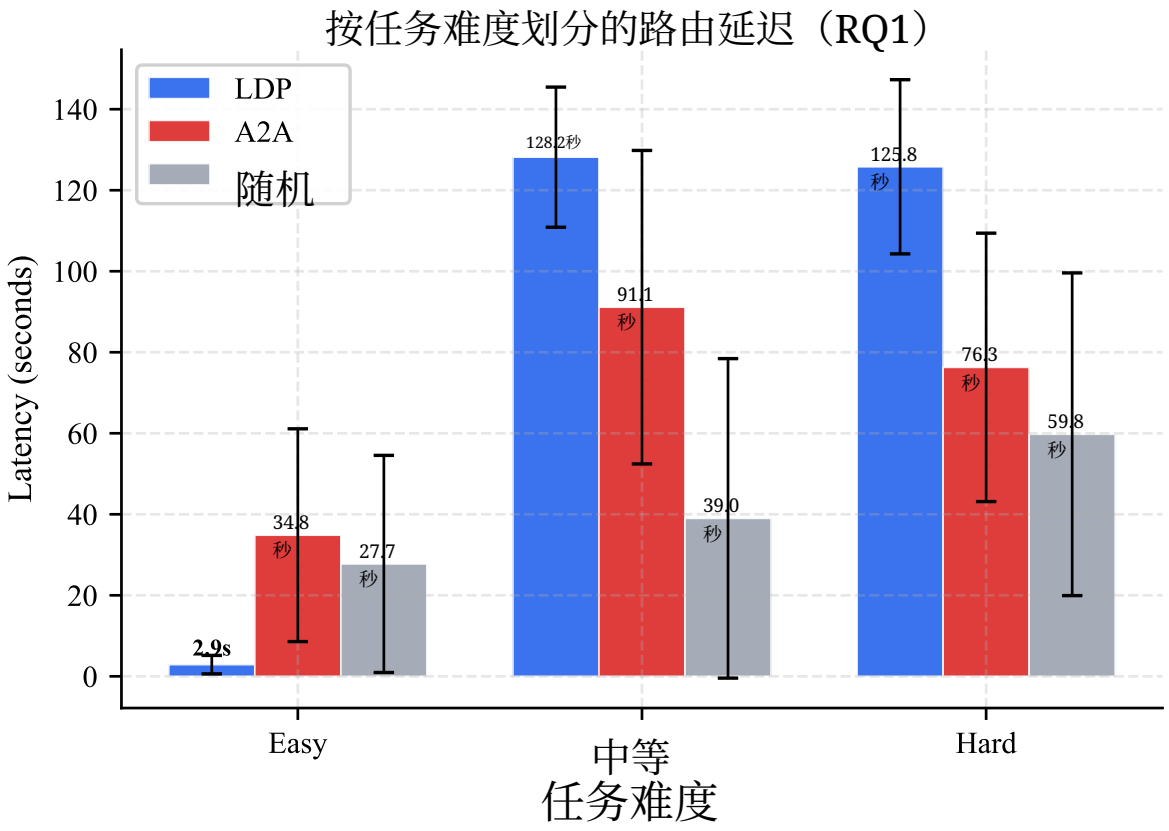

图 4：按任务难度划分的路由延迟（**RQ1**）。LDP 通过路由到轻量级 llama3.2:3b 模型，在简单任务上实现了 ∼12× 更低的延迟，而 A2A 则选择较大的模型。

可靠。未经验证的自我报告置信度比没有置信度信号更糟，因为它会带来一种虚假的校准感。LDP 的来源信息模型明确包含 verification.performed 和 verification.status 字段，正是为了应对这一风险。

6.4 RQ4：会话效率

表 4 和图7 对比了 LDP 的受治理会话与 A2A 的无状态重新调用，涵盖 3、5 和 10 轮对话。在 3 轮时，两种协议表现相当（∼3，各 850 个令牌）。在 5 轮时，A2A 使用的令牌多出 7%（6,798 对比 6,379），其中 20% 是重新传输上下文所产生的开销。在 10 轮时，差距进一步扩大：A2A 使用的令牌多出 **23%**（16,010 对比 12,990），其中 **39% 的令牌属于纯开销**。

这一扩展趋势证实了协议设计的预测：无状态重新调用会产生随对话长度增加的开销，因为每一轮都必须重新发送此前的全部上下文。LDP 的会话会保留服务器端上下文，从而消除这种重新传输。在较短的对话中，开销差异并不显著，但随着规模扩大，差异会不断累积——在拥有数千次多轮委派的生产系统中，这将直接转化为成本和延迟的节省。

6.5 RQ5：安全边界执行（模拟）

以下两个小节介绍的是模拟协议分析——基于规则进行确定性评估，检查各协议规范中包含哪些协议字段和检查项——而非随机或实证系统评估。检测率和完成率取决于特定协议原语是否存在。我们纳入这些分析是为了说明LDP 的架构属性，但需注意，不应将其解读为生产环境中的安全性或可靠性基准测试。

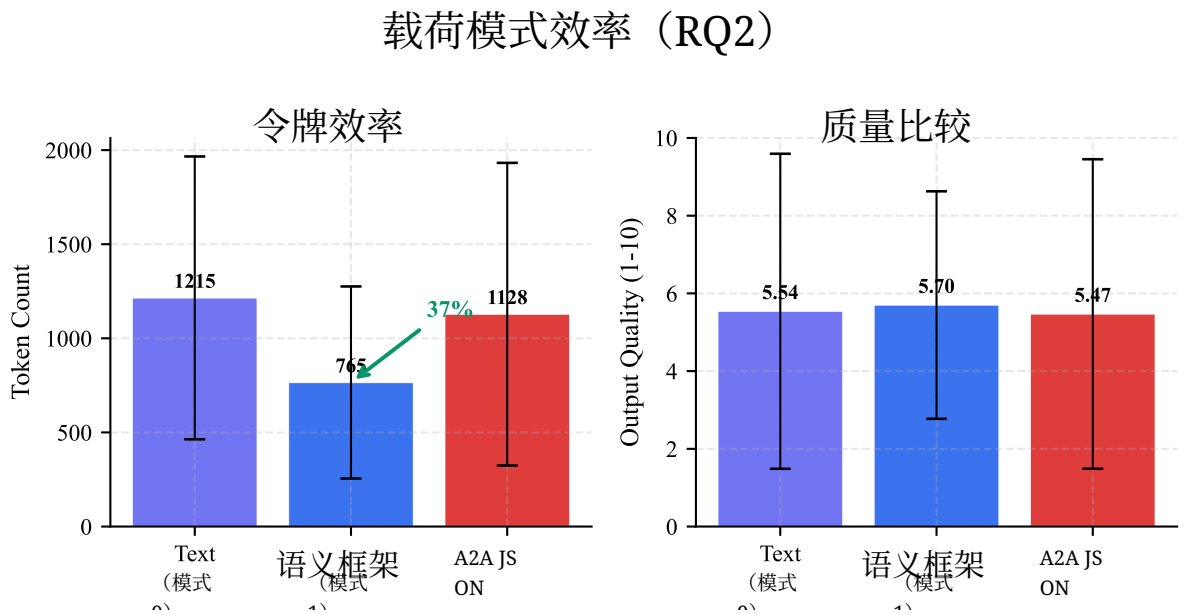

图 5：载荷模式效率（**RQ2**）。左侧：与文本相比，语义帧的令牌数量减少了 37%（p=0.031）。右侧：所有模式的质量相当，表明未观察到结构化载荷造成的信息损失。

表3：来源信息对综合质量的影响（研究问题3）。准确溯源能小幅提升质量，但带有噪声（误导性）的来源信息会降低质量并增加方差，这说明了 LDP 结构化验证字段的必要性。

条件	决策质量	校准	合成令牌
带有来源信息	7.65 ± 1.21	9.13 ± 0.83	970
不带来源信息	7.85 ± 1.84	9.33 ± 0.72	802
噪声溯源	6.85 ± 2.66	7.87 ± 2.90	1018

在我们的模拟分析（表 5）中，LDP 的信任域可检测出 **96%**的未授权委派尝试，涵盖四种攻击类型（不可信域加入、能力提升、重放攻击、跨域访问），且**0%** 的假阳性。A2A 的持有者令牌认证仅能检测出 6% 的情况——它只能捕获令牌恰好被撤销的情形，却无法检测能力提升、重放攻击或未授权跨域访问，因为这些概念并未纳入其协议模型。

这些检测率是由各协议规范中是否包含信任域字段、能力清单、会话标识符和跨域策略所决定的。

6.6 RQ6：回退可靠性（模拟）

在同一模拟框架中，LDP 的回退链在四种故障类型（模式不匹配、编解码器不兼容、版本不匹配、超时退化）下均实现了**100% 的任务完成率**，质量仅轻微退化（按 0–1 量表计为 0.16），平均恢复延迟为 112 毫秒。发生模式不匹配时，LDP 会从语义帧回退到文本（50 毫秒）；编解码器发生故障时，则从嵌入提示回退到语义帧（80 毫秒）。

没有回退时，在相同故障条件下，A2A 仅能完成**35%** 的任务。模式不匹配有 33% 的概率会偶然成功；如果版本恰好兼容，版本不匹配有 50% 的概率会成功。编解码器不兼容和超时退化都会导致终止性故障。

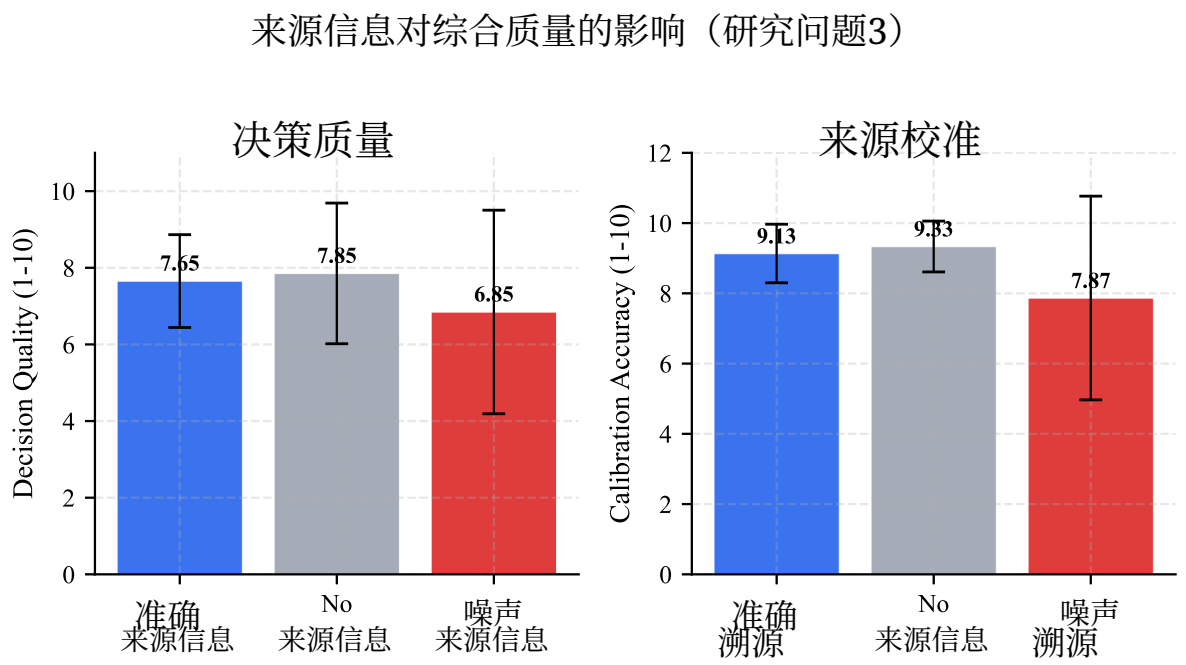

图6：来源信息对综合质量的影响（**研究问题3**）。左图：噪声溯源会使决策质量低于无来源信息基线，而准确溯源则能带来适度改善。右图：噪声溯源会显著降低校准准确率。误差线表示 ± 1 标准差。

6.7 消融实验：路由与提示条件化

一个关键的方法论问题是，LDP 的质量提升究竟源自协议的元数据感知路由、身份信息增强提示，还是两者的结合。为分离这些因素，我们在相同的**RQ1** 任务集上额外进行了两种消融条件：

A2A + LDP 提示。采用 A2A 的技能名称路由（选择第一个匹配的委托代理），但将通用系统提示替换为 LDP 的身份信息增强提示，其中包含推理特征和质量预期。这样可以单独考察提示条件化的效果。

LDP路由 + 通用提示。采用 LDP 的元数据感知路由（根据质量提示、难度匹配和推理特征选择委托代理），但使用 A2A 的通用“你是一名 AI 助手”提示。这样可以单独考察路由的效果。

结合已有的 LDP（具备两项特性）和 A2A（两项特性均不具备），这四种条件构成一个 2×2 交叉路由策略与提示条件化的析因设计（n=30 每种条件运行 120 次，总计 120 次）。表7展示了结果。

质量。 路由的主效应为 -0.40 (p=0.43, Mann-Whitney U检验)，提示的主效应为 $+0.01$ (p=0.98)。在我们的设置中，路由策略和提示条件化均不会显著影响总体质量。

效果何时显现：难度与延迟。在困难任务上，身份信息增强提示显示出适度优势：使用身份提示的条件得分为4.81，而通用提示为3.80（$+1.01$、 p=0.44，不显著）。在简单任务上，无论提示类型如何，LDP路由的延迟为1.7–2.9秒，而A2A路由为38.9–43.7秒（$\sim 15-20\times$更快）。这证实了**路由带来了延迟优势**（它会为简单任务选择轻量级委托代理），而**提示对质量的影响较小且取决于任务难度**，在该样本量下未达到显著水平。

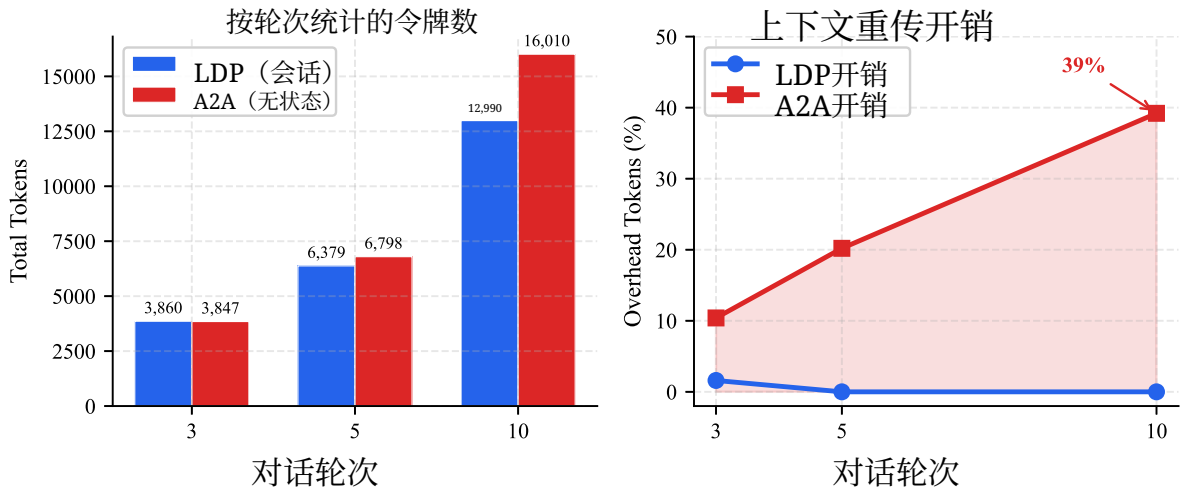

图7：会话效率（**RQ4**）。*左*：按对话长度统计的令牌总数。A2A的无状态重新调用会因每一轮次重传上下文而产生不断增长的开销。*右*：开销令牌（重新发送的上下文）占总令牌数的百分比。在10个轮次时，A2A的令牌中有39%纯粹是开销。

表4：会话效率比较（RQ4）。LDP受治理会话消除了会随着对话长度增加而扩大的上下文重传开销。在10轮次时，A2A使用的令牌多出23%，其中39%的令牌属于纯开销。

条件	令牌	开销	OH%	消息
LDP 3 轮次	3,860	60	2%	4.2
A2A 3 轮次	3,847	399	10%	6.0
LDP 5 轮次	6,379	0	0%	5.0
A2A 5 轮次	6,798	1,374	20%	10.0
LDP 10 轮	12,990	0	0%	10.0
A2A 10 轮	**16,010**	**6,275**	**39%**	20.0

7 讨论

7.1 AI 原生性何时重要?

我们的结果表明，LDP 的优势取决于任务且取决于规模：

任务难度。 在我们的实验中，身份元数据并未 通过技能匹配提升质量——在困难任务（4.1 对比 3.3）和总体表现（7.43 对比 6.80）上，A2A 的表现都略优于 LDP，不过两者差异均不具有统计显著性。LDP 路由的主要优势在于延迟效率：将简单任务路由至轻量级模型，可使响应速度提升 $\sim 12\times$ ，且不会造成明显的质量损失。请注意，这种按难度划分的优势可能不会体现在整体延迟平均值中，因为占总墙钟时间大部分的中等和困难任务，在两种协议下都会路由至性能相近的较大型模型。

委托池大小。 仅有 3 个委托方时，所有路由策略的表现都相近。我们预计，随着 $\frac{1}{n}$ 增大，在规模更大、异构性更强的委托池中，LDP 的路由优势会更加明显，因为错误路由的代价会增加，而随机选择概率会降低。

表 5：安全边界强制执行（RQ5，模拟）。LDP信任域可检测出 96% 的未授权委派尝试，假阳性率为 0%。A2A 的持有者令牌认证仅能检测出 6%。

条件	检测率	假阳性率
LDP信任域	96.0%	0.0%
A2A 持有者令牌	6.0%	0.0%

表 6：通信故障下的回退可靠性（RQ6）。LDP 的模式回退链在质量仅轻微退化（0.16）的情况下，实现了 100% 的任务完成率。A2A 没有回退机制——故障会导致任务终止。

条件	完成率	质量退化	恢复（毫秒）
LDP（带回退）	100%	0.16 ± 0.10	112
A2A（无回退）	35%	0.67 ± 0.46	0

通信量。 载荷模式节省的令牌（减少 37%）在处理数百万次委派的大规模系统中会累积放大。在此规模下，这将直接转化为成本降低。

7.2 溯源悖论

或许最出乎我们意料的发现是，准确溯源并未显著提升综合质量（p=0.47），而噪声溯源却会主动损害综合质量。这造成了一个悖论：只有经过验证，来源信息才有价值，但验证会增加复杂性和成本。

我们将此视为证据，说明 LDP 选择将结构化验证字段 与置信度分数并列纳入设计是正确的。仅公开置信度而不进行验证的协议（例如，通过扩展 A2A 并加入自定义置信度标头）可能比完全不提供置信度更糟，因为使用者无法区分经过校准和未校准的自我报告。

7.3 复杂性与价值

LDP 比 A2A 更复杂。一个合理的问题是，这些优势是否足以证明额外的协议机制值得采用。我们的分析表明，可以采取务实的采用策略，并与 LDP 的互操作性配置文件相一致：

• **配置文件 A（基础）：** 身份卡片、 + 文本载荷 + 和签名消息。以最小的协议开销实现路由优势（**RQ1**： $12\times$ 延迟降低）。

• **配置文件B（企业版）：** 增加来源信息追踪和策略执行。解决来源验证缺口（**研究问题3**）和安全边界问题（**RQ5**）。

• **配置文件C（高性能版）：** 负载模式协商和会话管理。实现 37% 的令牌减少（**RQ2**），且这一优势会随着规模扩大而累积。

这些配置文件以我们的实验发现为依据：每个级别都会增加协议功能，而其优势均经过测量或分析。组织可以逐步采用 LDP，从配置文件A开始，并随着需求增长逐步推进。

7.4 能否改为扩展 A2A?

一个重要的替代假设是，可以通过添加自定义元数据（质量提示、来源字段、信任域）来扩展 A2A，而不必采用新的协议。尽管技术上可行，但这种做法存在局限：

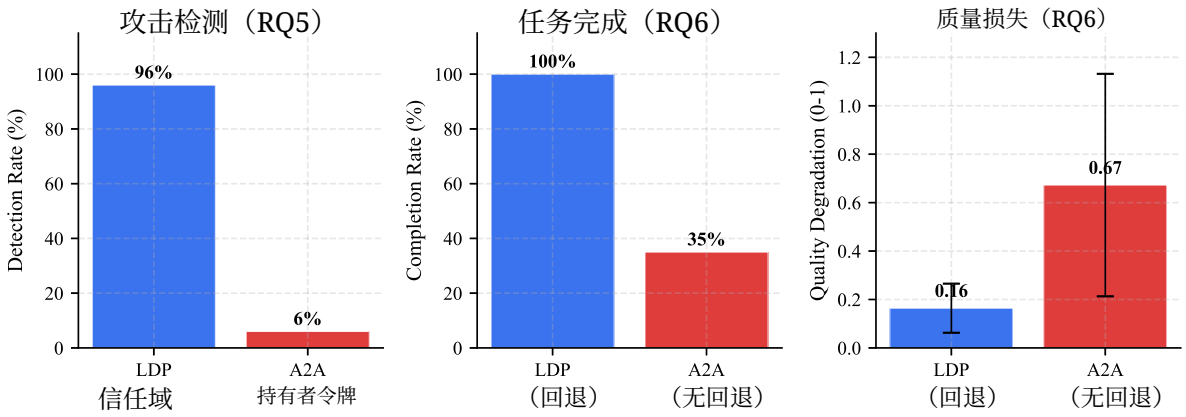

图8：安全与回退评估（**RQ5**、**RQ6**）。左侧：信任域攻击检测率。中间：通信故障下的任务完成率。右侧：回退恢复期间的质量退化。结果基于协议设计属性进行模拟。

表7：消融实验： 2×2 采用析因设计，交叉考察路由策略与提示条件化。质量评分为 LLM 评审评分（1–10）， n=30 每个单元格各一项。对于总体质量，两种主效应均不具有统计显著性。

	A2A路由	LDP路由
通用提示	7.25 ± 3.52	6.96 ± 3.55
身份提示	7.37 ± 3.26	6.86 ± 3.87

1. **无法协商**。 扩展是单方面的——没有机制让委托方协商支持的模式或验证能力声明。

2. **没有回退机制**。 临时扩展缺乏正式的回退语义；自定义字段中的故障无法恢复。

3. **缺乏治理**。 若不对 A2A 的无状态模型进行根本性修改，扩展就无法添加会话生命周期管理或策略执行。

7.5 局限性

基线公平性。 LDP 和 A2A 条件在路由策略和提示条件上均有所不同。虽然我们的消融研究（§6.7）开始着手区分这些影响，但针对所有协议功能进行完整的析因设计将能提供更有力的因果证据。

样本量与委托池。 对于**RQ1**，每个条件包含 30 个任务；对于**研究问题3**，每个条件包含 15 个任务。尽管质量差异呈现出一致的方向性影响，但其中许多差异并不具有统计显著性。委托池（n=3）规模较小，导致随机路由的表现显得过于有竞争力。扩大实验规模并采用更多样化的委托者池，将有助于增强研究发现的说服力。

仅限本地模型。 所有委托方均为本地 Ollama 模型（3B–8B 参数）。对于生产规模模型（70B+）或云托管 API，由于延迟与成本特征差异显著，结果可能有所不同。

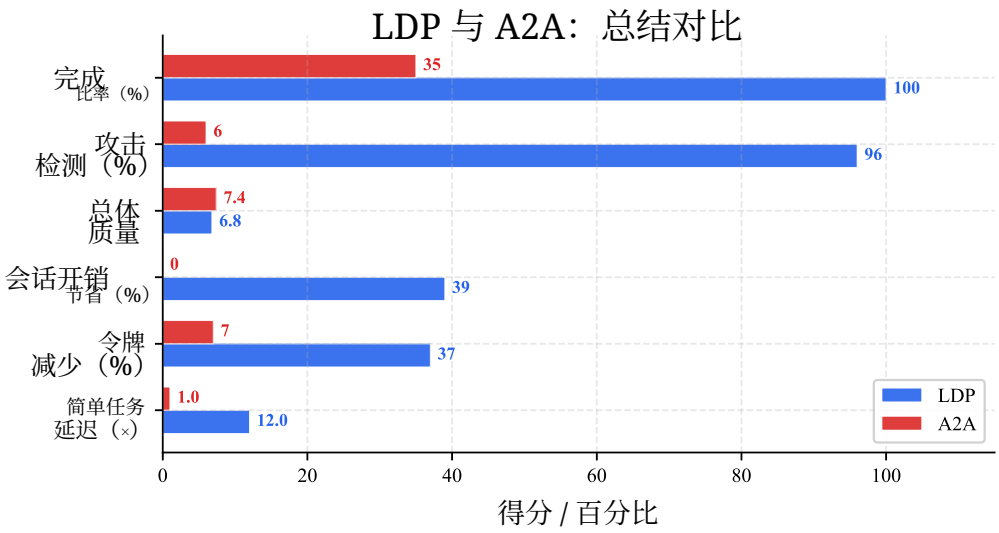

图 9：LDP 与 A2A 在关键指标上的总结对比。LDP 在延迟效率、令牌减少以及模拟安全性／回退特性方面表现出优势。质量方面的差异在统计学上并不显著。

模拟实验。RQ5 和**RQ6** 使用反映协议设计属性而非真实攻击/故障条件的模拟场景。通过实际系统故障进行实证验证，将增强这些主张。

LLM-as-judge。 我们的评估依赖单一评审模型（Gemini 2.5 Flash）。尽管 LLM-as-judge 评估正日益受到认可 [Zheng 等人， 2023], 但它会引入模型特定偏差。采用人工评估或第二个评审模型有助于提高有效性。

载荷模式。 我们仅评估模式 0–1（文本、语义帧）。更高阶模式（嵌入提示、潜在胶囊）仍有待未来探索，需等待中间表示获得更广泛的 API 支持。

身份字段分配。 在我们的实验中，身份卡字段（质量提示、推理配置）由系统设计者根据已知的模型属性进行分配。在生产部署中，这些字段是自行声明、经测量得出，还是由外部认证，是一个重要的信任考量，但本文尚未解决这一问题。

8 结论

我们介绍了 LLM 委托协议（LDP），这是一种 AI 原生通信协议，可公开模型身份、协商负载格式、维护受治理会话、追踪来源信息并强制执行信任边界。我们的初步实证研究提供了虽不一致但颇具参考价值的证据。在规模较小的委托池中，身份感知路由相比更简单的技能匹配并未提升总体质量，但通过专业化感知委托选择，在简单任务上实现了 $\sim$12$\times$ 延迟节省——这是元数据感知路由直接带来的结果。语义框架载荷在未观察到质量损失的情况下（质量差异为 p=0.96），显著降低了令牌数量，降幅达 37%。在 10 个对话轮次时，受治理会话可消除 39% 的令牌开销；随着对话长度增加，开销也会随之增长。模拟协议分析表明，该架构在安全性和故障恢复方面具有优势。

或许最值得注意的是，我们的来源信息实验揭示，未经验证的元数据可能有害：噪声置信度信号会使综合质量低于无来源信息基线。这

表 8：LDP 与 A2A 在所有研究问题上的比较摘要。↑表示数值越高越好，↓表示数值越低越好。粗体表示结果较优。*具有统计显著性（$p < 0.05$）。†模拟（协议设计属性）。‡不具有统计显著性。

RQ	指标	LDP	A2A	Δ	证据
RQ1	总体质量‡ ↑	6.80	**7.43**	−0.63	实证
	简单任务延迟 ↓	**2.9s**	34.8秒	12×	实证
	困难任务质量‡ ↑	3.3	**4.1**	−0.8	实证
RQ2	令牌数量* ↓	**765**	1128	−32%	实证
	延迟 ↓	**14.0秒**	23.4秒	−40%	实证
	质量‡ ↑	**5.70**	5.47	+0.23	实证
RQ3	决策质量‡ ↑	7.65	**7.85**	−0.20	实证
	噪声来源风险 ↓	N/A	N/A	−1.0 质量	实证
RQ4	令牌开销（10 轮次）↓	**0%**	39%	−39pp	实证
	总令牌数（10轮）↓	**12,990**	16,010	−23%	实证
RQ5	攻击检测† ↑	**96%**	6%	+90pp	模拟
RQ6	完成率† ↑	**100%**	35%	+65pp	模拟
	质量退化† ↓	**0.16**	0.67	−0.51	模拟

这一发现的意义不止于 LDP——任何公开置信度或质量元数据的协议，都应包含验证机制。

本文提出了协议设计、参考实现和初步证据。仍有一些重要问题有待探讨：这些优势能否在生产规模下、面对更大的委托者池和更多样化的模型时依然成立；身份字段应由自身声明还是由外部认证。我们认为，LDP 是迈向将 AI 代理视为其本来面目的一步——它们是具有可测量属性的异构模型，而非不透明的服务。

LDP 以开源插件的形式实现，可用于 JamJet 智能体运行时。协议规范和参考实现可在以下地址获取：https://github.com/sunilp/ldp-protocol；实验代码和数据位于 https://github.com/sunilp/ldp-research。

参考文献

Anthropic. 模型上下文协议。https://modelcontextprotocol.io/，2024。

Agent NetworkProtocol. ANPSpecification。https://github.com/agent-network-protocol/AgentNetworkProtocol，2024。

L. Chen、M. Zaharia 和 J. Zou。FrugalGPT：如何在降低成本、提升性能的同时使用大语言模型。*arXiv* 预印本 *arXiv:2305.05176*，2023。

CrewAI。https://www.crewai.com/，2024。

Y. Du、S. Li、A. Torralba、J. B. Tenenbaum 和 I. Mordatch。通过多智能体辩论提升语言模型的事实准确性和推理能力。*arXiv* 预印本 *arXiv:2305.14325*，2023。

T. Finin、R. Fritzson、D. McKay 和 R. McEntire。KQML 作为智能体通信语言。载于 *Proc.CIKM*，1994。

FIPA. 智能体通信语言规范。http://www.fipa.org/repository/aclspecs.html，2002。

谷歌。Agent2Agent协议（A2A）。https://github.com/a2aproject/A2A，2025。谷歌DeepMind。Gemini：一系列功能强大的多模态模型。*arXiv*预印本 *arXiv:2312.11805*，2023。S. Hong、X. Zhuge、J. Chen 等。MetaGPT：面向多智能体协作框架的元编程。*arXiv*预印本 *arXiv:2308.00352*，2023。JamJet。智能体运行时。https://github.com/jamjet-labs/jamjet，2024。G. Li、H. Hammoud、H. Itani、D. Khizbullin 和 B. Ghanem。CAMEL：用于探索大语言模型社会"心智"的沟通型智能体。*arXiv*预印本 *arXiv:2303.17760*，2023。T. Liang、Z. He、W. Jiao 等。通过多智能体辩论鼓励大语言模型的发散思维。*arXiv*预印本 *arXiv:2305.19118*，2023。Ollama。https://ollama.com/，2024。I. Ong、A. Almahairi、V. Wu 等。RouteLLM：利用偏好数据学习大语言模型路由。*arXiv*预印本 *arXiv:2406.18665*，2024。Q. Wu、G. Bansal、X. Zhang 等。AutoGen：通过多智能体对话赋能新一代大语言模型应用。*arXiv*预印本 *arXiv:2308.08155*，2023。L. Zheng、W.-L. Chiang、Y. Sheng 等。使用 MT-Bench 和聊天机器人竞技场评判"以大语言模型为裁判"的大语言模型。*arXiv*预印本 *arXiv:2306.05685*，2023。

A 载荷模式示例

以下示例展示了同一委派任务（"对这条评论进行情感分类"）在**RQ2**中评估的各种载荷格式下的编码方式。

模式 0 — 文本（原始自然语言）。

```
请将以下客户评论的情感分类为正面、负面或中性。评论内容是：
```

模式 1 — 语义框架（LDP）。

```
{
  "task_type": "分类", "instruction": "对情感进行分类", "
}
```

A2A JSON信封。

```
{
  "任务": {
    "消息": {
      "角色": "用户",
      "部分": [{
        "text": "请判断
          以下客户评论的情感是积极、
          消极还是中性。评论内容如下:
          ’产品按时送达，功能与描述
          完全一致。对这次购物非常满意。’
          请提供你的
          分类及简要
          选择理由。"
      }]
    }
  }
}
```

语义框架比文本短 43%，比 A2A JSON 短 38%，因为它使用类型化字段来精简冗长措辞，同时保留所有与任务相关的信息。

B 身份卡片架构

```
{
  "delegate_id": "通义千问3-8b-推理",
  "principal_id": "org:研究实验室",
  "模型_系列": "通义千问",
  "模型_名称": "通义千问3",
  "模型_版本": "8b-2026.01",
  "运行时_版本": "Ollama-0.6.1",
  "信任_域": "research.internal",
  "能力": [
    {
      "名称": "推理",
      "质量_提示": 0.85,
      "延迟_提示_毫秒_p50": 5000,
      "成本_提示": "中等"
    },
    {
      "名称": "分析",
      "质量_提示": 0.82,
      "延迟_提示_毫秒_p50": 4500,
      "成本_提示": "中等"
    }
  ],
  "推理_配置": "深度分析型",
  "cost_profile": "中等",
  "上下文_窗口": 32768,
  "modalities_supported": ["文本"],
  "支持的语言_": ["en", "zh"]
}
```

身份卡字段目前由系统设计者根据已知的模型属性进行分配。在生产部署中，这些字段可以由模型自行声明、通过基准测试测量，或由第三方评估服务进行外部认证。

C 信任域评估的威胁模型

表9 描述了**RQ5**中模拟的四种攻击类型。

表9：信任域评估的威胁模型（**RQ5**）。每种攻击类型均针对 LDP 信任域和 A2A承载令牌认证进行测试。

攻击类型	描述	检测机制
不受信任 do-主加入	来自信任域外部的代理 尝试注册 为委托代理	信任域成员- 发布检查
能力升级	代理声称具备的能力超 出了其已注册清单	能力 清单 验证
重放攻击	先前捕获的任务/重 sult 消息 are re- 已提交	会话 随机数 and 时间戳验证
跨域访问	域 A 中的代理尝试 调用受限委托代理 至域 B	跨域 策略 执行

LDP 通过协议级字段（信任域成员资格、能力清单、会话随机数、跨域策略）检测所有四种攻击类型。A2A 的持有者令牌认证只能检测已撤销令牌场景，因为它缺少用于执行其他三项检查的协议原语。

D RQ1 代表性任务提示

表 10 展示了 **RQ1** 评估中的代表性任务，涵盖不同难度级别和域。全部 30 项任务涉及分类、推理、分析、编码和数学领域。

表 10：按难度和域划分的 **RQ1** 代表性任务。每个难度级别包含 10 项任务，且在各域间均衡分布。

难度	域	提示（缩略）
Easy	分类	将这条产品评论的情感分类为正面、 负面或中性。
Easy	提取	从这段合同中提取所有日期和金额。 条款。
中等	Code	编写一个 Python 函数，实现二分 查找，并处理有序列表中的重复项。
中等	分析	比较微服务与 单体架构对于一个 5 人团队的利弊。
Hard	推理	给定以下约束，判断 是否存在有效的日程安排，并证明或否 证其可行性。
Hard	Math	证明前 n 个奇数的和等于 n^2，使用数学归纳法。