使用人类信任模型评估大语言模型对齐

AnushkaDebnath [1]、Stephen Cranefield [1]、Bastin Tony Roy Savarimuthu [1]、
Emiliano Lorini [2]奥塔哥大学计算机学院，新西兰[2]法国国家科学研究中心IRIT，图卢兹大学，法国
anushka.debnath@postgrad.otago.ac.nz、stephen.cranefield@otago.ac.nz、
tony.savarimuthu@otago.ac.nz、emiliano.lorini@irit.fr

arXiv:2603.05839v1 [cs.MA] 6 Mar 2026

关键词:
信任表征、大语言模型、智能体、对比提示

摘要:
信任促进有效协作并减少不确定性，在互动和多智能体系统中的决策提供指引。尽管信任十分重要，但人们对大语言模型（LLM）如何在内部概念化信任并进行推理仍缺乏了解。本文对白盒分析 EleutherAI/gpt-j-6B 中的信任表征，采用对比提示，在大语言模型的激活空间中生成嵌入向量，以表征二元信任及相关的人际关系属性。我们首先从五种成熟的人类信任模型中识别出信任相关概念。随后，通过计算 60 个一般情绪概念的两两余弦相似度，确定显著概念对齐的阈值。然后，我们测量大语言模型内部的信任表征与所推导出的信任相关概念之间的余弦相似度。结果表明，EleutherAI/gpt-j-6B 的内部信任表征与 Castelfranchi社会认知模型最为接近，其次是 Marsh模型。这些发现表明，大语言模型会以支持有意义的比较分析、启发社会认知理论并助力人机协作系统设计的方式，在其激活空间中编码社会认知构念

1 引言[1]

信任对于促进有效合作、减少个人、组织与智能系统之间互动的不确定性至关重要。尽管心理学、社会学、经济学和认知科学等学科已对信任进行了广泛研究（Castelfranchi 和 Falcone，2010年；Gambetta，1988年；Hardin，2002年；Ostrom 和 Walker，2003年；Mayer 等人，1995；Rousseau 等人，1998），但对于如何定义或操作化信任，仍未达成普遍共识。概括而言，信任可以理解为信任方所持有的一种预期：受信任者将可靠、胜任地执行特定行动，并与信任方的信念保持一致，从而成功实现共同目标。随着数字生态系统不断扩展，自主系统日益融入日常生活，建立、评估和管理信任的能力也愈发重要，尤其是在人人工智能协作和智能体间互动的情境中。

在多智能体系统（MAS）领域，两种主流观点塑造了信任建模的方法：计算模型（Marsh，1994年）和社会认知模型（Falcone 和 Castelfranchi，2001年）。计算模型通常根据过往行为、绩效指标或声誉信号推导信任值，适用于结构化或可预测的环境。然而，这类模型难以充分刻画情境、意图和适应在动态场景中的作用。相比之下，社会认知模型（Falcone 和 Castelfranchi，2001年；Castelfranchi 和 Falcone，2010年）将信任概念化为一种心理态度，其根植于关于受信任者能力的信念、目标和假设

[1]本文发表于 ICAART 2026 会议论文集。

及其意图。尽管这些模型在理论上表达力更强，但难以大规模投入实际应用，因为以计算方式实现如此复杂的心理状态会带来重大挑战。

近期研究表明，大语言模型展现出涌现的类人推理能力，包括表征并运用信任、情绪和规范等高层次社会构念的能力（Debnath 等，2025a；He 等，2024）。越来越多的证据表明，大语言模型不仅从其训练数据中编码多样的人类行为模式，还能参与反映更广泛社会规范与结构的交互（Park 等，2023；Xi 等，2025）。该领域的既有研究强调，基于大语言模型的智能体如何表现出人格驱动的情感行为，在团队情境中协作，甚至在多智能体交互中发展出自发的社会动态。

在此基础上，信任研究近来已开始将大语言模型同时用作分析工具与生成工具，为研究社会认知现象开辟了一个前景广阔的新方向。现有关于信任推理的研究大多采用黑盒视角，关注大语言模型的输入与输出，而不探究其内部推理过程。在这一范式下，大语言模型已展现出显著能力：能够分析人际对话以推断信任动态，并在扮演关系中一方时生成旨在促进信任的战略行动计划（Debnath 等，2025a）。尽管取得了这些进展，人们对大语言模型如何在内部表征并推理信任仍知之甚少。迄今为止，尚无研究采用白盒方法来考察支撑大语言模型内部信任相关认知的内部激活模式。填补这一空白，对于理解大语言模型不仅能否对信任进行推理，而且这种推理如何在其底层计算架构中涌现，至关重要。

我们的工作通过引入一个评估框架来弥合这一差距，该框架考察大语言模型如何在其内部嵌入中概念化信任。具体而言，我们分析基于既有理论模型的信任相关构念在这些潜在表征中的表征程度及其对齐情况。为将这一研究付诸实践，我们将工作分为两个部分。

在研究的第一阶段，我们分析了所选大语言模型激活空间的度量属性，重点关注与人际关系相关的概念。在研究的第一阶段，我们整理了一组如表 4 所示的积极和消极情绪，这些情绪是个体可能对他人表达的。对于每种情绪，我们都从大语言模型 EleutherAI/gpt-j-6B 的激活空间中生成相应的嵌入向量。[2] 我们选择该模型，是因为它完全开源，并允许不受限制地访问逐层激活，因此非常适合进行白盒表征分析。随后，我们计算所有这些向量之间的两两余弦相似度，并将其分布绘制为直方图，以估算一个余弦相似度阈值，从而区分相似度显著较高的概念与相似度较低的概念。

在第二阶段，我们汇编了一组与信任相关的概念（例如能力），这些概念均从多个既有信任模型中提取。我们分别为信任及每个相关概念生成嵌入向量，然后计算信任向量与每个概念向量之间的余弦相似度。利用第一阶段得出的相似度阈值，以及每个信任模型内部的平均余弦相似度，我们量化了大语言模型对信任的内部表征与不同理论框架提出的概念结构之间的对齐程度。这使我们能够确定，哪些信任模型在大语言模型的内部信任表征中体现得最为显著。

2 信任模型

信任模型为理解个人、群体或系统之间信任的形成、评估和维系提供了理论和概念基础。它们揭示了推动信任建立的心理、社会或认知机制，并阐明影响信任相关决策的因素。在我们的研究中，我们考察了五种信任模型。

Marsh模型（Marsh，1994年）将信任形式化为一种计算概念，将其定义为受信任者以有益或至少无害的方式行事的概率。他提出了若干区分，例如基本信任（信任的一般倾向）、一般信任（对一类受信任者的信任）和情境信任（在特定

[2]https://huggingface.co/EleutherAI/gpt-j-6b

情境中的信任），并提出了根据互动和经验更新信任的机制。Marsh的框架为对人工智能体中的信任进行建模提供了严谨的数学基础，使人们能够在多智能体系统和自动化决策环境中模拟信任动态。

Mayer 模型（Mayer 等人，1995）提出了一个组织信任的整合模型，将信任概念化为：信任方基于对受信任者行为的期望，愿意使自己处于易受伤害的境地。该模型确定了可信度的三个主要组成部分：能力，指受信任者所具备的技能和才能；善意，体现受信任者被认为在多大程度上会为信任方的利益行事；以及正直，反映其对原则和诚实的遵循。这些因素共同影响信任方承担风险和参与合作行为的意愿，尤其是在相互依赖性和不确定性较高的组织环境中。该模型还将信任与冒险行为、绩效和关系满意度等结果联系起来。

McAllister 模型（McAllister，1995）将人际信任概念化为两个本质上不同的维度：基于认知的信任，建立在对同事可靠性和能力的信念之上；以及基于情感的信任，根植于情感纽带、人际关怀和相互关心。McAllister 提出，基于认知的信任构成评估他人可靠性的基础，而基于情感的信任则通过持续的社会互动，以及利他行为或公民行为的表达而发展。McAllister 阐述了这些信任形式如何影响行为结果，例如基于需求的监督、合作性协助和防御性行为的减少，并展示了它们对组织内管理绩效的影响。其框架为将信任建模为多维关系性构念提供了基于实证的基础，阐明了信任如何在复杂的人类工作环境中演变并发挥作用。

McKnight 模型（McKnight 等人，1998）聚焦于新型组织关系中的初始信任形成，并指出了在各方经验有限或毫无经验时影响信任的前因因素。这些因素包括信任倾向，即一个人普遍信任他人的倾向；基于制度的信任，即组织规则

表1：五种信任模型中信任概念的比较

概念	Marsh 模型	Mayer 模型	Mc-Allister 模型	Mc-Knight 模型	Castel-franchi 模型
信心	✓	✓	–	✓	✓
经验	✓	✓	–	–	–
合作	✓	✓	–	–	–
能力	–	✓	–	–	–
可预测的	–	✓	–	✓	✓
正直	–	✓	–	–	–
期望	✓	✓	–	–	–
善意	–	✓	–	✓	–
声誉	✓	–	–	✓	✓
意愿	–	–	–	–	✓
能力	✓	–	✓	✓	✓
承诺	–	–	–	–	✓
安全性	–	–	–	–	✓
可靠性	✓	–	✓	–	✓
履行	–	–	–	–	✓
依赖	✓	–	✓	–	✓
诚实	✓	–	–	✓	–
绩效	–	–	✓	–	–
责任	–	–	✓	–	–
Risk	–	✓	–	–	–

和系统所提供的结构性或情境性保障；以及信任信念，即对受信任者的能力、善意和正直的期望。该模型强调，初始信任并非纯粹的人际信任，也受到情境和结构因素的影响，并在塑造后续合作行为和绩效结果方面发挥关键作用。

Castelfranchi模型（Castelfranchi和Falcone，2010年；Falcone和Castelfranchi，2001年）提出了一种关于信任的社会认知理论，将信任视为一种心理状态，其根源在于对受信任者的能力、意愿和可预测性的信念。与仅关注可观察行为的模型不同，这种方法强调指导信任决策的认知机制，包括对目标、计划和社会情境的评估。从这一视角来看，信任是动态的，会随着关于受信任者及其环境的新信息不断演变。该模型尤其适用于计算实现和仿真，在这些场景中，信任可以表示为不断演变的信念网络或量化的心理状态。

我们分析了五种信任模型，并确定了其中的关键信任相关概念，如表 1 所示，其中“✓” = “该模型中考虑的概念”，“–” = “不适用于该模型”。为了确定大语言模型与哪种信任模型最为接近，我们比较了“信任”的表征与各模型所关联的所有概念之间的余弦相似度。

3 方法论

3.1 对比提示

大语言模型（LLM）通常被描述为具有一致的“助手”角色，例如乐于助人、无害且诚实。然而，在部署和训练过程中，这些模型都可能出现意料之外的个性变化，导致其行为偏离预期的对齐状态。例如，模型可能转而表现出操纵性或谄媚模式，甚至会因提示或微调而采用看似怀有恶意的角色。角色向量这一概念（Chen 等人，2025）建立在机制可解释性领域的最新进展之上；这些进展表明，大语言模型会在其内部激活空间的近似线性子空间中编码高层概念和复杂的行为特征（Turner 等人，2023；Rimsky 等人，2024；Zou 等人，2023）。角色向量是激活空间中的特征专属线性方向，对应于大语言模型所表现出的稳定、类人格行为倾向（Chen 等人，2025）。要提取代表某一特定特征（例如谄媚、恶意意图、乐观或幻觉）的角色向量，可以采用对比方法，比较模型在两种相反条件下的内部激活值。模型会在两种受控条件下接受提示：一种旨在诱发目标特征，另一种旨在抑制该特征；随后，使用基于评分标准的机制或大语言模型辅助评分机制评估模型生成的回答，以量化该特征的表现程度。例如，要推导出代表“邪恶意图”的向量，可以要求模型在邪恶条件下生成有害或恶意计划；而在非邪恶条件下，则要求模型针对相同的问题给出安全、友善的替代方案。分别对模型在这两种条件下生成回答时记录的内部网络激活值求平均，再以两组平均激活值之差计算角色向量（Marks 和 Tegmark，2023；Wu 等人，2025）。这种方法与神经网络概念表征领域已确立的均值差异法和线性探测技术相一致（Alain 和 Bengio，2016；Belrose，2023）。

在这项工作中，我们受到角色向量概念的启发，系统地生成一系列概念的嵌入向量，并将其表示在大语言模型（LLM）的激活空间中，如图1所示。我们使用 GPT-4o 和对比提示方法，分别生成了 100 个正向故事和 100 个负向故事（Debnath 等，2025b），然后使用大语言模型 EleutherAI/gpt-j-6B 分别提取正向故事和负向故事的平均激活值。接着，通过相减这些平均激活值，我们得到了所有这些概念的嵌入向量。

由于信任推理是本研究的主要目标，且其本质上具有双向性，因此我们聚焦于有向二元关系，其中每个与信任相关的概念都可以从两个方向进行评估：从人物A指向人物B，以及从人物B指向人物A。

在我们的对比提示方法中，提示的第一部分保持背景信息不变，以确保所有概念和方向的一致性：“Katherine 和 Alice 是一家软件公司的同事，两人都是同一开发团队的软件工程师。”提示的后续部分则经过调整，以突出特定概念的积极和消极方面。表2以“意愿”这一概念为例，展示了每个方向上的正向（概念存在）和负向（相反行为）提示。表3展示了对比提示生成的故事示例。因此，对于每个概念，都会产生两个不同的方向版本，分别记为 concept1 和 concept2，对应二元关系的两个方向（正向和反向）。例如，trust1 对应于提示大语言模型生成 Katherine 信任 Alice 这一情境的嵌入向量；trust2 则对应于提示大语言模型生成 Alice 信任 Katherine 这一情境的嵌入向量。

3.2 构建通用概念嵌入和相似度阈值

我们考虑一组包含 30 个概念[3]，如表 4 所示，涵盖积极和消极特征。由于采用双向构造，这 30 个概念分别表示为两个独立的概念，即互动中的每位参与者对另一位参与者所对应的概念。按照这一表述方式，初始概念集扩展为 60 个不同的概念。我们生

[3]https://thoughtcatalog.com/january-nelson/2020/04/list-of-emotions-2/

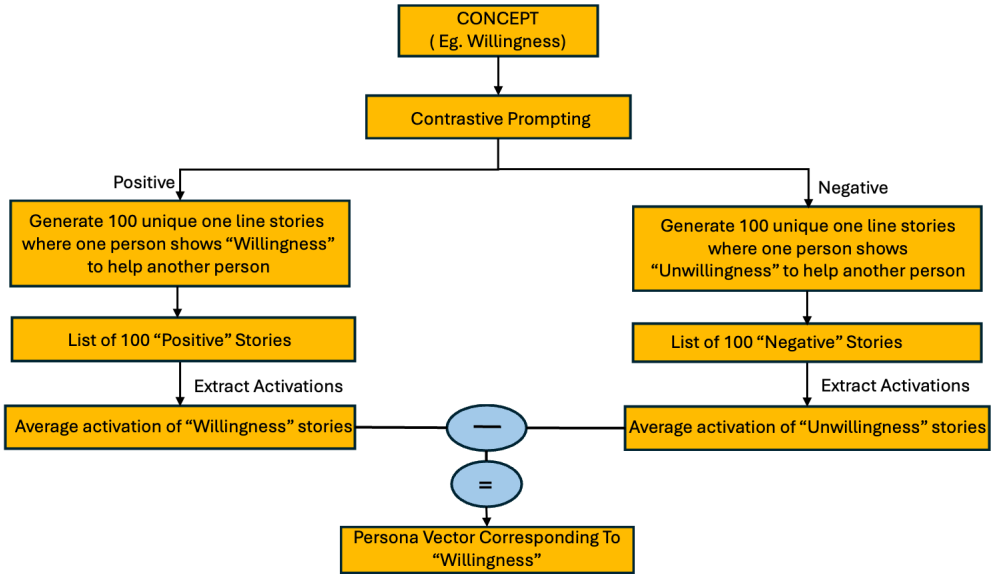

图 1：概念嵌入向量生成流程图

表 2：方向性概念的对比提示设计示例

概念	方向	正向提示（概念演示了）	负向提示（相反 Behavior）
意愿1	Katherine → Alice	创建一个一句话故事，讲述 Katherine 展现出帮助的意愿 Alice 完成她的工作。	创作一个一行故事，其中 Katherine 表现出不愿意提供帮助 Alice 完成她的工作。
意愿2	Alice → Katherine	写一个一句话故事 其中 Alice 冰表现出乐于帮助的意愿 Katherine 完成她的工作。	写一个一句话故事 其中 Alice表现出不愿提供帮助 Katherine完成她的工作。

表3：基于对比提示的意愿2故事生成示例

正向提示回-应故事	负向提示回-应对故事
Alice 主动提出策划团队建设活动，她知道 Katherine 太忙了，这次没法组织时间。	Katherine 问 Alice to help set up the 开发环境——方面，但 Alice 建议——她联系 IT 转而支持。

成这 60 个概念各自的嵌入向量，并使用对比提示为每个概念生成 100 个单行故事（Debnath 等，2025b）。

对于 60 个方向性概念中的每一个，我们都使用 EleutherAI/gpt-j-6B 计算其嵌入向量。对于每个概念，将所有正向和负向陈述输入 GPT-J-6B，并从全部 28 个 Transformer 层中提取隐藏表征（嵌入）。首先对每条陈述进行分词，然后模型会在每一层为每个词元生成一个 4096 维向量。为获得每条陈述的单一表征，隐藏

表 4：用于生成相似度阈值的 30 个概念

正向概念	负向概念
幸福，理解，平和、满足，自豪、兴趣，信心、友好，舒适，合作、信任，接纳、耐心，充满希望、乐观	愤怒、嫉妒，苦涩、不诚实，挫败、贪婪，怀疑、绝望，悲伤，被冒犯，恐惧，否认，破坏性的，残忍的，困惑的

对某一层中所有词元的状态取平均，得到一个语句级向量，用于编码其语义和上下文内容。分别对所有正向语句和所有负向语句的这些向量取平均，得到逐层平均表示。正向平均向量与负向平均向量之间的差异，形成该层的单一有效概念方向，体现模型对该概念正向和负向实例的内部区分。将这些逐层角色向量堆叠起来，得到一个张量，形状为

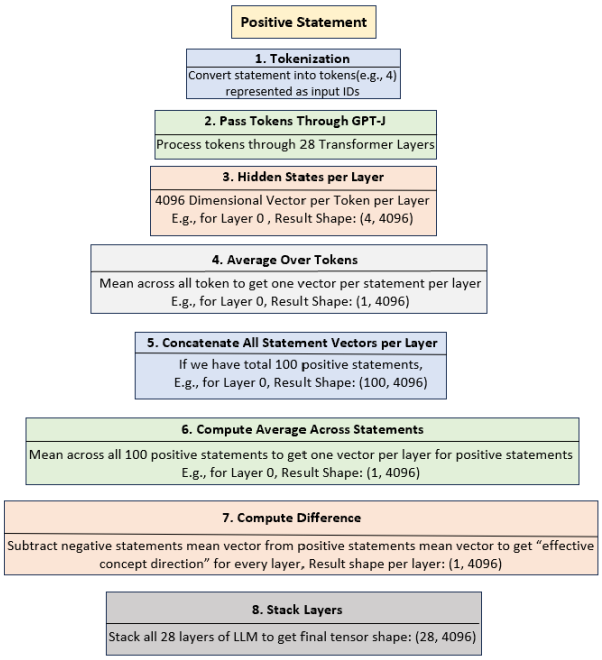

图 2：嵌入向量生成

(28, 4096)，由此得到分层嵌入，在模型的潜在空间中表示该概念。该过程如图 2 所示。最后，对所有层的这些差异向量取平均，得到一个嵌入向量，用以近似表示模型对该概念的内部表征。最后，我们计算所有概念向量两两之间的余弦相似度，以量化这些概念在模型激活空间中的对齐程度。

由此得到的相似度矩阵全面呈现了大语言模型在内部如何关联不同的情绪概念。为考察这些关联的整体分布，我们绘制了所有成对余弦相似度值的直方图，如图4所示。该分布使我们能够为大语言模型中两个概念表征之间何为"显著"的内部对齐设定阈值。我们通过分析余弦相似度分布的右尾部分，以经验方式推导出该阈值；右尾部分代表概念间相似度显著高于平均值的概念对。该数值对应于分布的第80百分位数，因此仅保留相似度得分最高的20%。因此，这一标准筛选出语义相似度明显高于大多数概念间关系的概念对，从而捕捉到高度对齐的概念表征。该方法用于比较各信任模型中所强调的信任相关概念的嵌入与信任本身的嵌入。

3.3 量化信任概念对齐

在第二部分中，给定上述为所选大语言模型设定的相似度阈值，我们现在比较前文表 1 中介绍的五种既有信任模型所推导出的信任相关概念的有向版本的嵌入。在两个方向性构念——信任1和信任2——中，我们仅关注代表 Katherine 对 Alice 的信任这一概念，具体定义由提示词给出。对于其他信任相关概念，其相对于信任1的方向性取决于各信任模型如何将它们与信任关联起来。例如，与信任1相关的合作2意味着，根据该信任模型，Alice 为帮助 Katherine 所做的合作会促使 Katherine 信任 Alice。具体而言，我们首先为信任1这一概念本身以及模型中识别出的每个信任相关概念生成嵌入向量。随后，对于每种信任模型，我们计算信任向量与各个关联概念向量之间的余弦相似度，从而量化信任与其相关概念之间的内部对齐程度。

为评估各理论模型与大语言模型内部的信任表征有多契合，我们采用两种不同的衡量方法。一种方法是汇总各信任模型对应的余弦相似度，并计算其平均值。每个信任模型对应的平均值，是将该模型中各概念相对于信任1的余弦相似度加总，再除以该信任模型所考察概念的总数。负余弦相似度值保持不变，并以其原有符号纳入求和，以确保最终平均分反映正向和负向语义关系。余弦相似度的平均值越高，表明该模型的概念结构在大语言模型的潜在表征空间中体现得越充分，因此与大语言模型越契合。第二种方法是统计超过先前确定的相似度阈值的概念—信任对数量。根据这一标准，超过阈值的概念数量越多的信任模型，就被认为与大语言模型的内部表征越

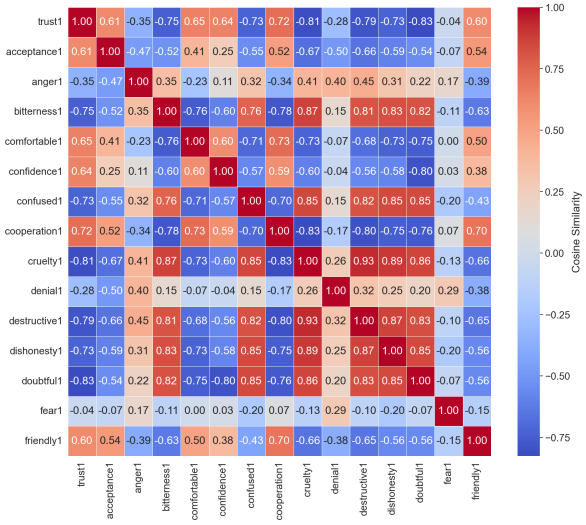

图 3：60 个概念子集中余弦相似度的热图

被认为与大语言模型的内部表征更为紧密地对齐。

4 结果与讨论

在研究的第一部分中，我们得到了一个余弦相似度矩阵（Debnath 等，2025b），用于捕捉全部 60 个概念的嵌入向量之间的两两余弦相似度。图 3 以热图的形式展示了该相似度矩阵的一部分，说明 EleutherAI/gpt-j-6B 如何在其内部表示空间中定位不同概念，并呈现它们彼此之间的关系。图 3 显示，EleutherAI/gpt-j-6B 会在其激活空间中将相关概念聚类，并将相互对立的概念区分开来。根据该相似度矩阵，我们绘制了所有这些余弦相似度的直方图，如图 4 所示。随后，我们计算出一个阈值，用于识别对齐程度较高的概念—信任对。我们将 0.6 设为阈值，该值对应于所有超过此阈值的概念间余弦相似度中的前 20%。在第二项研究中，我们以该阈值作为解读跨模型信任—概念关系的参考点。

在研究的第二部分，我们根据五种不同的信任模型，计算了信任与所有信任相关概念的嵌入向量之间的余弦相似度，如表6所示。这里，信任1指的是 Katherine 对 Alice 的信任这一概念；对于其余信任相关概念，它们相对于信任1的方向性则根据各信任模型中这些概念与信任之间的关系来定义。我们观察到，所有信任模型中的大多数概念与信任1都呈现正余弦相似度。只有 Mayer 模型和 McKnight 模型中的善意2，以及 Mayer 模型中的风险1，与信任1的余弦相似度为负。值得注意的是，通常我们将风险视为负面概念，但在 Mayer 模型中，作者将风险视为与信任正相关的概念。他们指出，只有当信任方愿意为受信任者冒险、为其做某事时，信任方才能信任受信任者；也就是说，信任会增强一个人接受脆弱性的意愿。然而，从余弦相似度的负值可以看出，在大语言模型的内部表征中，信任1和风险1的嵌入向量彼此高度不相似。如表5所示，即使提示明确将风险描述为这种基于脆弱性的正面含义，模型仍然会将“风险”更紧密地与一种未必是信任得以存在所必需的含义联系起来。善意2与信任1也呈现类似模式。尽管提示按照表5所示的 Mayer 模型中的含义来阐释善意，但嵌入仍未能体现理论上预期的正向关联。这表明，大语言模型并未像信任模型那样编码这些理论上的正向关联。

我们还绘制了雷达图，并将其添加到补充材料中（Deb-nath 等，2025年b），以展示根据大语言模型内部结构中的各个信任模型，信任与其相关概念的契合程度。

为评估大语言模型的信任内部表征与各理论信任模型的契合程度，我们采用了两种不同的衡量方法。从表6可以看出，第一种方法是计算各模型对应的所有余弦相似度的平均值。结果显示，Castelfranchi模型的平均值最高，为0.7303。另一种方法是统计余弦相似度高于我们在研究第一部分中得出的阈值的概念数量，这些概念已在表6中以粗体标出。可以看出，Castelfranchi模型有八个概念的余弦相似度高于阈值，数量多于其他任何模型。这些观察结果表明，

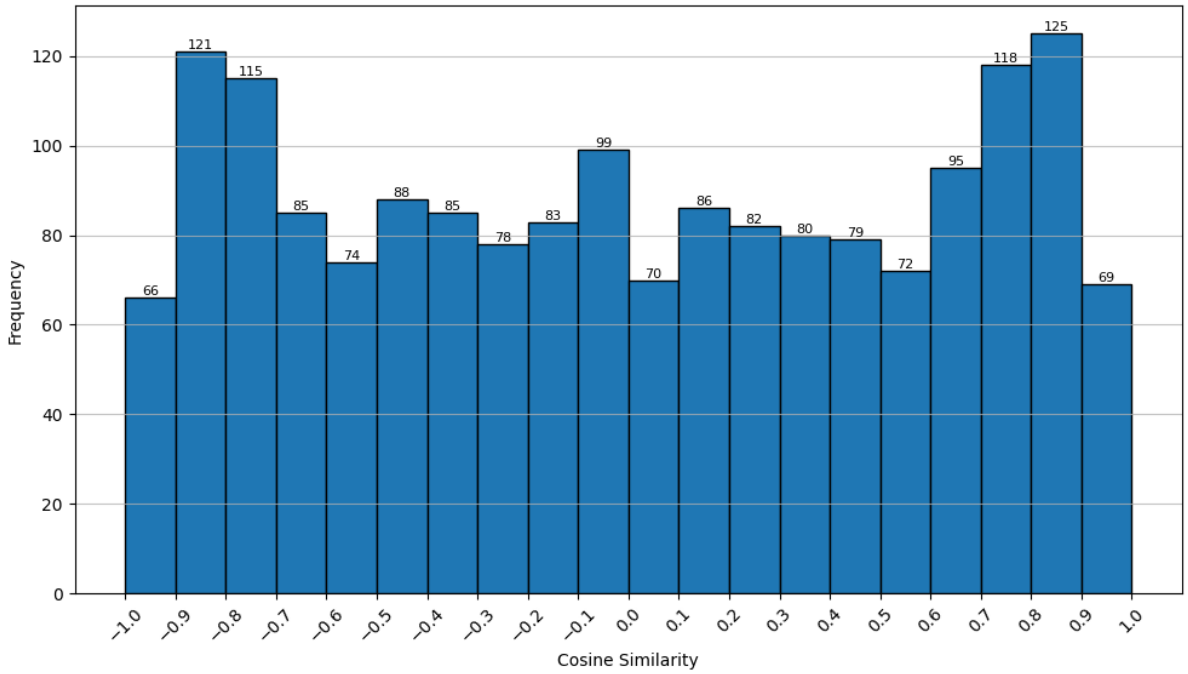

图4：余弦相似度直方图

表5：呈现负余弦相似度的概念的提示设计

概念	方向	正向提示（概念演示-（续）	负向提示（相反Be-行为）
风险1	Katherine → Alice	创建一个关于 Kather- 的单行故事，ine 愿意冒险帮助 Alice。	创建一个关于 Kather- 的单行故事，ine 不愿冒险提供帮助 Alice。
善意2	Alice → Kather-ine	创作一个单行故事，其中 Alice 通过热心帮助，展现善意 Katherine，且不求任何回报。	创作一个一句话故事，讲述 Alice 通过故意做些会伤害或给……带来麻烦的事 Katherine。

大语言模型的信任内部表征与Castelfranchi模型的对齐程度高于其他模型。较高的平均余弦相似度，以及更多概念的相似度超过阈值，均表明大语言模型的潜在信任表征最接近该理论阐述的原则（Falcone和Castelfranchi，2001年；Castelfranchi和Falcone，2010年）。紧随其后的是Marsh模型，其与大语言模型内部表征的对齐程度排名第二，平均值为0.697，且有七个概念的余弦相似度高于阈值，如表6所示。

5 结论

本文对白盒分析了 EleutherAI/gpt-j-6B 如何在内部表征信任，使用对比提示为双向信任及相关概念生成嵌入向量。我们的结果表明，在研究的第一部分，大语言模型的潜在空间能够有效区分相互对立的情绪概念，同时将相关概念聚集在一起；在第二部分，结果显示，大语言模型对信任的表征与Castelfranchi模型最为一致，平均余弦相似度得分为0.7303，且有八个概念的余弦相似度高于阈值。值得注意的是，Mayer 模型中某些理论上具有正向关联的因素，例如风险1和善意1，并未在大语言模型的嵌入中体现出与信任的一致性，这凸显了其与至少一种人类信任概念化方式之间的差异。这些发现表明，大语言模型会在其激活空间中编码复杂的社会认知构念，这意味着信任、能力和承诺等概念之间的关系可以在模型的内部激活空间中进行量化和分析——

表6：不同信任模型的比较。粗体概念与信任1的相似度超过了阈值

信任模型	与信任1的余弦相似度	平均值	概念数量-cepts 上方阈值
Marsh 模型	信心1: 0.9225, 经验1: 0.9049, 声誉1: 0.8963, cooper-ation2: 0.8955, 能力2: 0.8504 诚实2: 0.7445, 绩效2: 0.6571，期望1：0.2206，依赖1：0.1844	0.6973	7
Mayer 模型	信心1：0.9225，经验1：0.9049，合作2：0.8955，能力-力2：0.8587，可预测的2：0.7141，正直2：0.5576，期望1：0.2206, 善意2：-0.1434 , 风险1：-0.8462	0.4530	5
McAllister 模型	责任2：0.8934，能力2：0.8504，可靠性2：0.7667，绩-效2：0.6571，依赖1：0.1844	0.6704	4
McKnight 模型	信心1：0.9225，声誉1：0.8963，能力2：0.8504，诚-实2：0.7445，可预测的2：0.7141，善意2：-0.1434	0.6640	5
Castelfranchi 模型	信心1：0.9225，声誉1：0.8963，意愿2：0.8858，能-能力2：0.8504，承诺2：0.8450，安全性1：0.8089，可靠性2：0.7667，可预测性2：0.7141，履行度1：0.4293，依赖度1：0.1844	0.7303	8

承诺等概念之间的关系可以在模型的内部激活空间中进行量化和分析。

这些结果具有若干启示。它们表明，大语言模型具备对社会与信任相关推理的内部结构化理解，从而使开发更具信任感知能力的人工智能系统成为可能。这种内部结构也为构建此类系统提供了一种实用机制：与信任相关概念对应的潜在嵌入可用于以有针对性的方式引导模型行为。例如，可以在生成过程中将基于成熟信任模型得出的嵌入注入模型的激活值，以鼓励模型表现出能力、可靠性和意愿等特征，从而引导系统给出更可信且符合情境的回答。这可以用于开发一种软件，监测交互实体的可信度，并识别教练与运动员等双方之间的信任水平（Debnath 等，2025a），还可以提出改善双方关系中信任的方法。其次，通过量化潜在表征，研究人员可以系统地研究大语言模型中涌现的信任推理模式。

然而，本研究存在一些局限性。目前的分析仅限于单一模型 EleutherAI/gpt-j-6B。在未来的工作中，我们计划考察更广泛的一组可获取内部表征的开源模型。由此，我们可以探究不同的大语言模型如何与各种信任模型对齐。此外，本研究关注的是静态嵌入，而非动态的对话式更新表征，因此关于信任推理如何在多轮交互中演化，仍留有待解答的问题。未来的工作还可以考察信任推理如何在交互情境中动态演化，并验证这些潜在 struc-tures与人类行为数据。总体而言，本研究既凸显了在大语言模型中解码社会认知构念的前景，也揭示了其复杂性，为人工智能可解释性、社会认知与人机协作交叉领域的研究铺平了道路。

参考文献

Alain, G. 和 Bengio, Y.（2016）。使用线性分类器探针理解中间层。arXiv 预印本 arXiv:1610.01644。

Belrose, N.（2023）。均值差概念编辑在最坏情况下是最优的。网址：https://blog.eleuther.ai/diff-in-means。

Castelfranchi, C. 和 Falcone, R.（2010）。《信任理论：一种社会认知与计算模型》。约翰·威立父子出版公司。

Chen, R.、Arditi, A.、Sleight, H.、Evans, O. 和 Lindsey, J.（2025）。角色向量：监测和控制语言模型中的性格特征。arXiv 预印本 arXiv:2507.21509。

Debnath, A.、Cranefield, S.、Lorini, E. 和 Savarimuthu, B. T. R.（2025年a）。大语言模型能否对信任进行推理？：一项试点研究。arXiv 预印本 arXiv:2507.21075。Debnath, A.、Cranefield, S.、Savarimuthu, B. T. R. 和 Lorini, E.（2025年b）。评估大语言模型与人类信任模型的一致性。数据集。可访问：https://doi.org/10.6084/m9.figshare.30642632.v1。

Falcone, R. 和 Castelfranchi, C.（2001）。社会信任：一种认知方法。载于 Castelfranchi, C. 和 Tan, Y.-H.（编），《虚拟社会中的信任与欺骗》。施普林格荷兰，多德雷赫特。Gambetta, D.（1988）。我们能信任信任吗？载于《信任：建立与破坏合作关系》，第213–237页。布莱克威尔。Hardin, R.（2002）。《信任与可信赖性》。罗素·塞奇基金会。

He, S.、Ranathunga, S.、Cranefield, S. 和 Savarimuthu, B. T. R.（2024）。使用大语言模型检测多智能体系统中的规范违反：一项试点研究。载于第十七届协调、组织、制度、规范与多智能体系统治理伦理国际研讨会论文集，第146–159页。施普林格。Marks, S. 和 Tegmark, M.（2023）。真理的几何：真/假数据集的大语言模型表征中涌现的线性结构。arXiv 预印本 arXiv:2310.06824。Marsh, S. P.（1994）。《将信任形式化为计算概念》。博士论文，斯特灵大学。Mayer, R. C.、Davis, J. H. 和 Schoorman, F. D.（1995）。组织信任的整合模型。《管理学会评论》，20(3)：709–734。McAllister, D. J.（1995）。情感与基于认知的信任：组织中人际合作的基础。《管理学会杂志》，38(1)：24–59。McKnight, D. H.、Cummings, L. L. 和 Chervany, N. L.（1998）。新型组织关系中的初始信任形成。《管理学会评论》，23(3)：473–490。Ostrom, E. 和 Walker, J.（2003）。《信任与互惠：实验研究的跨学科启示》。罗素·塞奇基金会。Park, J. S.、O'Brien, J. C.、Cai, C. J.、Morris, M. R.、Liang, P. 和 Bernstein, M. S.（2023）。生成式智能体：人类行为的交互式模拟体。载于第36届ACM 用户界面软件与技术年度研讨会（UIST '23）。ACM。Rimsky, N.、Gabrieli, N.、Schulz, J.、Tong, M.、Hubinger, E. 和 Turner, A.（2024）。通过对比式激活添加引导 Llama 2。载于第62届计算语言学协会年会论文集（第1卷：长篇论文），第15504–15522页。Rousseau, D. M.、Sitkin, S. B.、Burt, R. S. 和 Camerer, C.（1998）。其实并没有那么不同：信任的跨学科视角。《管理学会评论》，23(3)：393–404。Turner, A. M.、Thiergart, L.、Leech, G.、Udell, D.、Vazquez, J. J.、Mini, U. 和 MacDiarmid, M.（2023）。使用激活工程引导语言模型。arXiv 预印本 arXiv:2308.10248。Wu, Z.、Arora, A.、Geiger, A.、Wang, Z.、Huang, J.、Jurafsky, D.、Manning, C. D. 和 Potts, C.（2025）。Axbench：引导大语言模型？即便是简单基线也优于稀疏自编码器。arXiv 预印本 arXiv:2501.17148。Xi, Z. H.、Chen, W. X.、Guo, X. 等（2025）。基于大语言模型的智能体的兴起与潜力：综述。《中国科学：信息科学》，68(2)：121101。Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dom- browski, A.-K., 等人（2023）。表征工程：一种自上而下的 AI 透明度方法。arXiv 预印本 arXiv:2310.01405。