

他们想假装不理解：当前大型语言模型在解释政治话语隐性内容中的局限

Walter Paci
University of Florence
walter.paci@unifi.it

Alessandro Panunzi
University of Florence
alessandro.panunzi@unifi.it

Sandro Pezzelle
University of Amsterdam
s.pezzelle@uva.nl

Abstract

隐含内容在政治话语中起着至关重要的作用，演讲者系统地运用如含义和预设等语用策略来影响听众。大型语言模型 (LLMs) 在需要复杂语义和语用理解的任务中表现出色，显示出它们在检测和解释隐含内容意义方面的潜力。然而，它们在政治话语中的这一能力仍然 largely 未被探索。首次利用包含意大利政治演讲及其操控性隐含内容标注的大型 IMPAQTS 语料库，我们提出了一些方法来测试 LLMs 在这一挑战性问题上的效能。通过多项选择任务和开放式生成任务，我们证明，所有被测试的模型都在解释预设和含义方面存在困难。我们得出结论，目前的 LLMs 缺乏准确解释高度隐含语言（如政治话语中）所必需的关键语用能力。同时，我们指出了一些有前景的趋势和未来提升模型性能的方向。我们在此发布我们的数据和代码：<https://github.com/WalterPaci/IMPAQTS-PID>

1 引言

隐性和含糊的语言在日常生活交流中普遍存在。考虑政治背景中的语言：政治家往往使用隐性或含糊的内容，依赖听众的假设、背景知识和共享的语境知识，以间接地传达意义。这在使用模糊的术语、暗示和修辞手法中很明显，这些方法允许多重解释。这样的语言创造了一种可以微妙传达意义的环境，通常意图是为了说服、操控或强化某些意识形态 (Morency et al., 2008)。通过巧妙地使用隐性语言，政治家可以激发观众的情感、假设和文化背景，同时避免做出可能被挑战或反驳的明确承诺 (Lombardi Vallauri, 2019)。考虑以下例子（也出现在这篇作品的标题中）：他们想假装不理解¹。通过这样说，发言人暗示被指的人假装他们不理解某些明显的东西，而实际上他们非常清楚；此外，发言人也在微妙地暗示他们有原

¹这个例子是意大利语句子“Vogliono far finta di non capire”的英文翻译，它是本研究中使用的数据的一部分。在主文中，为了简化，我们只提供英文翻译。原始意大利语文本可在附录 A、C 和 I 中找到。在本节中，某些例子有时可能会为了说明目的而做出调整。

因地这样假装，指导和引导观众达到某种反应或效果。这种说话者说一件事以暗示另一件事的语用现象被称为会话含义。会话含义是隐性语言中研究最为广泛的现象之一 (Grice, 1990; Sperber and Wilson, 1987)，与预设 (Strawson, 1964; Garner, 1971; Ducrot, 1972) 并列。由于这两者在政治话语中普遍存在，使其成为分析隐性沟通策略的理想领域 (Van Dijk, 1992; Lombardi Vallauri and Masia, 2020)。

大型语言模型 (LLMs) 的出现彻底改变了自然语言处理，带来了在语义理解、问题回答和语用解释任务上的空前进展。总体而言，LLMs 在语义任务上表现优秀，并在语用任务上展现潜力，尤其是在涉及人工构建的刺激时。然而，是否能够处理需要对自然话语数据进行语用推理的任务，例如政治话语，仍是一个开放的问题。在这项工作中，我们特别针对这一研究问题进行探讨，并测试当前的 LLMs 是否可以解释政治话语中隐含操控内容的意义，例如上例中的会话含义。我们首次在 NLP 研究中运用大型 IMPAQTS 语料库，其中包括意大利政治演讲的转录文本，并附有多种隐含内容的专家注释。我们提出了几种对最尖端的意大利语 LLMs 的语用有效性进行评估的方法。我们的贡献如下：

- 我们为大型语言模型提出一项具有挑战性的任务，即通过解释来澄清在政治话语中发现的操控性隐含内容的含义；例如，在这段话中“他们想假装不理解”中，讲话者暗示这些人非常理解。
- 我们发布了一个新的数据集，基于 IMPAQTS 语料库 (Cominetti et al., 2024)，包含 3 万个隐含段落（隐含意义和预设），并附有专家给出的解释，以及正确理解隐含内容所需的周围语言环境（通过人类专家注释进行实证验证）。
- 通过选择题任务和开放式生成任务，我们展示了所有测试的模型在解释操控性的预设和含意方面存在困难。在选择题设置

中，表现最好的模型比估计的上限低了超过 20 个准确率点；在开放式生成设置中，它仅在四分之一的情況下提供了完全正确的解释。

- 尽管结果大多不令人满意，我们表明使用连贯性思维 (CoT) 可以在开放式生成设置中提高性能，这表明推理机制可以帮助解决复杂的实际任务。我们还提出了其他方向，例如嵌入关于政治人物及其政治背景的外部知识，以进一步增强模型性能。

2 相关工作

2.1 隐式语言与大型语言模型

随着大型语言模型 (LLMs) 驱动能够与人类用户互动的对话代理，它们必须具备理解隐含内容的能力。先前的研究已经探索了 LLMs 在语言交流中解释非字面含义的能力。Jeretic et al. (2020) 通过分析预训练语言模型如何计算标量含意和预设来研究实用理解的产生。他们发现有强有力的证据表明，像 BERT 这样的模型学习了一些和所有的量词的标量含意，但在其他标量对方面表现困难，把它们视为同义词。对于预设，模型无法识别一些预设触发器，例如动词“停止”当它预设一个曾经进行的动作时，比如约翰停止吸烟。Zheng et al. (2021) 提出了首次对语言模型在五种从 Grice 的 (1990) 会话准则中衍生出的含意进行评估，发现模型在选择题任务中表现相对较好，在该任务中，模型必须选择哪个给定选项解释了对话中的含意，但在会话推理中表现显著困难，表明它们并未完全理解会话背景。在类似的研究中，Hu et al. (2023) 评估了未经特定任务微调的预训练语言模型，将格赖斯的含意作为它们所测试现象之一。结果表明，语言模型能够推断出实用意义，但尚未确定这是由于语言线索还是认知过程，强调了在实用推理中研究其联系的必要性。Kim et al. (2023) 通过提示 LLMs 对特定场景提供二元答案，评估它们是否能理解会话中的蕴涵。研究表明，尽管 LLMs 展现了一些隐含的理解，但在通过连贯思维提示引导进行推理时，它们的表现显著改善。最近，Ruis et al. (2024) 设计了一种协议来评估 LLM 在二元蕴涵解析上的能力，突出人与 LLMs 之间存在显著差距。

所有这些研究都调查了实际使用语用语言的合理情况。然而，他们使用的是人为构建的句子，而不是实际语料库中发现的语用现象的真实例子。在我们的工作中，我们向前迈出了重要的一步，使用了反映现实世界通信复杂性的半自发、生态化政治语篇摘录。与人为构建的

句子不同，政治语篇以捕捉修辞、社会背景和说话者意图的复杂交互而著称。

2.2 自然语言处理中政治语言

政治语言一直是自然语言处理研究的重点，利用来自社交媒体和政治演讲的大型数据集来研究话语框架、偏见检测和极化。

Katre (2019) 使用计算方法进行文本分析和政治演讲稿的可视化。作者运用了自然语言处理技术来分析政治演讲，并为词语使用、词汇分布等生成图形可视化。在 Huguet Cabot et al. (2020) 中，自然语言处理方法被用于通过关注隐喻、情感和政治修辞来建模政治话语。作者提出了首个联合模型，将隐喻和情感检测作为辅助任务整合进来，以提高对新闻文章的政治视角、政治家的党派归属以及政策问题框架的预测性能。Németh (2023) 对 2010 年至 2021 年间研究人员如何使用自然语言处理技术研究语言极化进行了方法论回顾。作者辨识了数据来源和计算方法，并回顾了极化的不同概念化和操作化。

最近的努力利用大型语言模型 (LLMs) 增强对政治话语的理解。Marino and Giglietto (2024) 使用 LLMs 进行政治话语注释，报告比先前的方法如主题建模显著提高。在这里，LLMs 被用于分析与 2018 年和 2022 年意大利选举相关的 Facebook 链接，并执行政治内容的分类和聚类以及生成这些聚类的描述性标签等任务。Li et al. (2024) 提出了 Political-LLM，这是一个将 LLMs 整合到政治科学中的综合框架。这提供了前所未有的文本分析能力。同时，作者强调了在政治应用中使用 LLMs 时需考虑的偏见、伦理和可解释性等若干问题。尽管取得了这些进展，但对自动解释隐含内容的关注甚少。据我们所知，我们是首次解决这个问题。

3 数据

3.1 IMPAQTS 语料库

我们首次在 NLP 研究背景下使用 IMPAQTS (Cominetti et al., 2024) 的数据，² 是一个意大利政治话语的语料库 (指的是“政治家的话语”；参见 ?)，包括从 1946 年到 2023 年间由 150 名著名政治人物用意大利语发表的 1500 篇演讲。IMPAQTS 是一个多模态语料库，包含 800 多篇视频格式的演讲和大约 600 篇音频格式的演讲。这些演讲的人工转录也已提供，总计约 265 万标记。在我们的工作中，我们仅使用这些转录文本。

IMPAQTS 的一个关键特征，对于我们的工作至关重要，是对所有包含具有某种操纵性

²<https://impaqts.dilef.unifi.it/>

含义的隐含内容的段落进行标注。在更技术性的术语中，这种隐含内容被称为非真诚的真实；这些是隐含的可疑内容，它们并非善意传达，但在给定的背景下仍然非明确地被理解为真实。IMPAQTS 定义了四种类型的隐含内容：含意、预设、话题化和模糊性。关于这个理论框架的详细讨论，请参见 [Lombardi Vallauri \(2016\)](#)。每个被标记为包含某些隐含评论的句子都会附有由 IMPAQTS 语料库标注者撰写的评论，解释隐含内容的意义。³ 标注者的评论格式都是一样的；例如，对于含意，它以“它暗示了...”开头；对于预设，它以“它预设了...”开头；对于话题化，它以“它在讨论中认为...是活跃的”开头；对于模糊性，它以“它留下了...的模糊性”开头。

下面，我们报告一个单一示例，其中包括 IMPAQTS 注释者的评论：

含蓄内容的文本：意大利不需要另一个由 Monti 领导的政府。意大利最不需要的是另一个成为银行奴隶的政府！

评论：这暗示了 Monti 的政府是银行的奴隶。

在这项工作中，我们专注于先前提出的四个类别中的两个，即含意和预设。我们选择不包括话题化，因为其理解和解释在很大程度上依赖于其在口语中由韵律轮廓传递的信息 ([Frascarelli and Hinterhölzl, 2008](#))，而我们在使用转录时无法利用这些信息。另一方面，由于模糊性的定义较为宽泛且不够集中，因此被排除在外。

IMPAQTS 仅包括独白，分为六种类型：议会演讲、集会演讲、政党大会演讲、当面声明、广播声明、新媒体声明和业务会议采访。在这里，我们仅选择关注议会演讲和集会演讲中的句子，其中隐含内容通常不涉及外部背景，并且可以在演讲转录中检测到。IMPAQTS 语料库包含三个时间段的演讲：1946 年至 1972 年、1973 年至 1993 年以及 1994 年至 2022 年。我们对这些时期的句子进行了实验。

3.2 需要多少语言上下文？

为了实证地回答这个问题，我们进行了一项人工验证研究，邀请 9 位语言学专家标注员评估理解目标句子隐含内容所需的左侧上下文句子数量，即目标句子之前的句子。我们选择了 126 个句子，平衡了隐含内容的类型（63 个含义推理和 63 个预设）和时期（时间范围内的 42 个句子）。然后我们创建了 3 份调查问卷，每份问卷包含 42 个嵌入了隐含内容的样本。在

³在 IMPAQTS 语料库中，每个句子都由 3 名独立的标注者进行标注。这些标注随后由第四位专家进行验证，并给出最终意见。

每份调查中，3 位专家被要求阅读包含隐含内容的句子及由 IMPAQTS 标注员提供的解释。

我们的专家必须评估目标句子是否提供了足够的信息来理解评论中的解释，或者是否需要额外的前置语言背景。如果是后者情况，将会出现目标句子前的一句额外句子，并再次提出相同的问题。根据设计，我们的标注人员最多可以看到五个左侧的句子。标注是在 LimeSurvey ([LimeSurvey GmbH, 2012](#)) 上进行的，并允许确定每个句子正确隐含解释所需的必要背景量。调查的详细信息和标注者间的协议见附录 C。

3.3 实验数据

基于上述的注释结果，我们开始构建一个用于我们与 LLM 实验的数据集。我们考虑了 IMPAQTS 中包含隐含意义或预设的所有样本。对于这些样本中的每一个，我们都检索了讲话中前面的 4 个句子。最终的数据集，我们命名为 IMPAQTS-PID，其中 PID 代表预设和隐含意义数据集，包含 31,822 个样本，并配有来自 IMPAQTS 的隐含内容解释；特别是，其中 14,932 个样本嵌入了隐含意义，16,890 个样本嵌入了预设。在附录 B 中，我们报告了一些关于我们数据集的描述性统计数据。

4 方法

使用上面介绍的 IMPAQTS-PID 数据集，我们对模型理解句子的隐含内容提出挑战。在本节中，我们描述了在我们的工作中测试的模型、评估其能力的实验，以及两个实验共有的实验细节。

4.1 模型

我们实验了四种使用多语言数据进行预训练的模型，因此适合处理意大利语：GPT4o-mini ([OpenAI, 2024](#))、Aya Expanse 8B ([Cohere-ForAI, 2024](#))、LLAMA3.1 8B ([MetaAI, 2024a](#)) 和 LLAMA3.2 3B ([MetaAI, 2024b](#))。第一个模型是专有的，其他三个是开放权重的。这些模型的规模相对较小，在计算效率、成本和部署可行性方面有优势。因此，这些模型通过需要更少的资源达到了良好的平衡，从而减少了硬件成本和能源消耗，同时实现了更快的推理。我们通过下面描述的两个实验来测试模型对隐含内容的理解能力。我们进行一个实验，模型被提供四个可能的解释（即解释）以说明给定的隐含内容，必须选择正确的一个。也就是说，我们将问题设定为一个多项选择生成 (MCG) 任务。我们向模型输入一个简短的指令和四个候选项 (A、B、C 和 D)，其中只

有一个是正确的。为了选择具有挑战性的干扰项，我们使用基于主题相似性的方法。我们首先对 IMPAQTS-PID 中的所有解释运行一个主题建模分析。我们识别了 450 个主题和一个包含不共享共同主题的解释的剩余类别。然后从目标解释的相同主题类别中抽样干扰项，可能是 450 个基于主题类别之一或剩余类别。通过这种方式，我们确保三个干扰项是不正确但合理的，迫使模型对输入进行推理而不是依赖捷径。在推理过程中，四个候选项的顺序是随机的以避免偏差；即正确答案的位置在 A-D 之间大致均衡。鉴于我们数据集中的每个实例都包含一个真实答案，我们使用纯准确率来评估模型性能。

4.2 开放式生成 (OEG)

将问题设定为 MCG 任务有很多优势，包括容易通过准确率设置和评估。同时，由于干扰项的选择以及非自然选择可用选项的设置，它也存在相关的局限性。在这个实验中，我们规避了这些局限性，直接查询模型以生成给定隐含内容的解释；即，我们将问题设定为开放式生成 (OEG) 任务。我们尝试了三种提示设置，下面我们简要描述一下：

模型在没有任何先前示例或详细指导的情况下被询问，完全依赖于其预训练知识。这是用于此提示技术的模板：

解释以下文本的隐含内容。考虑到它出现在最右边的句子中。

文本：[句子嵌入隐含内容]

隐含内容：_____

小样本 提示中包含四个示例，即两个内涵和两个预设，包括嵌入隐含内容的文本及其解释。由于我们旨在引导模型关注操控性的隐含内容，因此我们只包括“正面”示例（即存在内涵或预设的情况）。我们不包括“负面”示例，以避免模型关注与语义无关或不需要的语言现象。这是这种提示技术使用的模板（为简明起见，我们只展示两个示例而非四个）：

文本：[第一个示例文本]

隐含内容：[第一个示例文本中隐含内容的解释]

...

文本：[第四个示例文本]

隐含内容：[第四个示例文本中隐含内容的解释]

文本：[句子嵌入隐含内容]

隐含内容：_____

思维链 (CoT) 提示提供了一个详细的分步解释，概述了解码隐含内容所需的推理过程。我们从给 IMPAQTS 语料库注释者的详细说明中改编了这些步骤。这是用于此提示技术的模板（因篇幅原因缩短）：

隐含意义在交流中出现，当说话者“挑战”源自众所周知的合作原则的四大会话准则之一时。...

我们将“预设”定义为交流参与者理所当然接受的任何内容，更具体地说，是那些作为交流者已共享知识的一部分传达的内容。...

考虑一下刚刚关于隐含意义和预设所说的内容，并逐步解释以下文本中的隐含内容是什么。请注意，它出现在最右边的句子中。

文本：[句子嵌入隐含内容]

隐含内容：_____

与 MCG 任务不同，在 MCG 任务中，由于有真实答案存在，可以根据简单的准确性进行评估，而这里评估模型性能则更为复杂。基于自动相似性测量或将 LLMs 作为评估者是潜在的解决方案。然而，任务设置引入了关于其可靠性的不确定性。鉴于这些忧虑，我们选择基于人类专家评估的小规模但更健壮的评估方式。我们设计了一套评估协议，其中语言学专家评估包含隐含内容的文本、来自 IMPAQTS 的关于隐含内容的评论和模型生成的隐含内容解释。具体示例详见附录 C.2。专家需通过从 5 个可能选项中选择一个来判断模型生成的解释的质量：

- 完全正确：解释简短且集中于 IMPAQTS 原评论中突出显示的同一主题。
- 在各种选项中是正确的：答案列出了许多可能的解释；在这些解释中，提到了正确的一个（其他解释的正确性无关紧要）。
- 部分正确：模型输出的简短答案没有完全解决问题，但包含了一些正确解释的元素。
- 完全错误：一个简短或冗长的答案是错误的，即与 IMPAQTS 的评论所强调的主题无关。
- 未作答：模型拒绝回答，回避问题，或者只是突出显示包含隐含内容的句子而不给出解释。

关于两个任务的提示和解码设置的详细信息，包括标记限制、温度参数、GPU 和计算时间，已在附录 I 和 ?? 中提供。

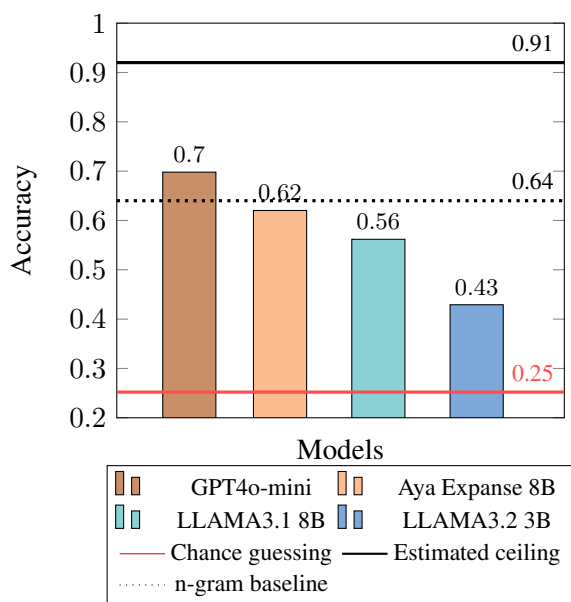


Figure 1: MCG 任务。所有模型的表现均优于随机水平（红线）。只有 GPT4o-mini 持续超过相似性基准（虚线）。

5 结果

5.1 MCG 结果

如图 1 所示，所有模型的表现都远高于随机水平，这表明它们理解任务并有时能选择正确的解释。同时，我们观察到不同模型之间存在显著差异，表现最好的 GPT4o-mini 达到了 70 % 的准确率，比 LLAMA3.2 3B (43 %) 高出近 30 个百分点。

有趣的是，GPT4o-mini 是唯一一个持续超过 n-gram 基线的模型，而其他所有模型的表现都比这种启发式方法差。一方面，这表明大语言模型并不系统地依赖仅基于输入和答案之间表面相似性的策略。另一方面，这也暗示了 GPT4o-mini 具备某些实际能力，使其能够理解隐含内容并超越提示中的表面相似性。为支持这一结论，我们使用 IronITA (Cignarella et al., 2018) 来评估该模型，这是一种用于实用理解的意大利基准。该数据集包括手动注释的推文，用于二元分类任务，旨在确定推文是否具有讽刺意味。虽然该基准并不专注于操控性隐含内容，但其核心是讽刺，这是一个常常通过隐含意义传达的实用现象。因此，它仍然为模型的实用能力提供了有用的见解。GPT-4o-mini 在该基准上的准确率达到 66 %，表现好于基线模型和随机猜测，尽管其效能仍有限。与此部分对照，INVALSI 基准 (Puccetti et al., 2025) 排行榜⁴报告 GPT4o-mini 在文本性和实用子任务

⁴<https://huggingface.co/spaces/Crisp-Unimib/INVALSIbenchmark>

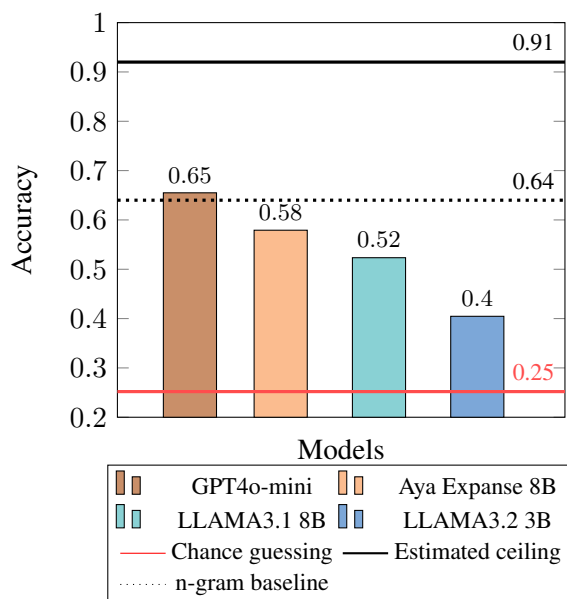


Figure 2: 困难负样本设置。在具有共同主题的文本子集上，MCG 任务的分解精度得分。

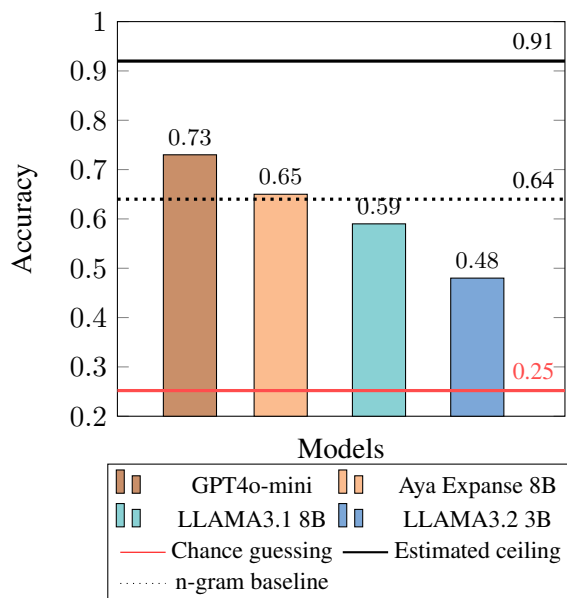


Figure 3: 简单负样本设置。对没有共同主题的文本子集上的 MCG 任务进行分解准确度评分。

上的准确率为 83 %。综合来看, 这些发现表明, 当用结构化和人工上下文 (如 INVALSI) 提示时, GPT-4o-mini 似乎能够捕捉某些实用现象, 但仍在面对更多依赖于上下文的线索 (如讽刺推文或 IMPAQTS-PID 数据集中的线索) 时表现困难。事实上, 在我们的实验中, 即使这一表现最好的模型仍比估计的最高准确率低了 20 多个百分点。因此, 我们得出结论, 即整体模型在这一任务上的表现仍然远未令人满意。

干扰项分析 在一个多项选择任务中, 干扰项的选择可以显著影响模型的最终性能。为了评估这一因素是否影响我们的结果, 我们计算了两个数据子集中的模型准确率: 一个是干扰项与正确解释属于同一主题类别 (高难度负样本), 另一个是目标解释和干扰项都来自没有明确主题的剩余类别 (低难度负样本)。图表 2 和 3 报告了这些子集中的准确性结果。总体来看, 我们发现所有测试模型在低难度负样本子集中的准确率始终高于高难度负样本子集。例如, GPT4o-mini 在这两个子集之间表现出大约 8 % 的性能差异, 即从 73 % 到 65 %。这证实了干扰项选择可以显著影响模型性能, 引入混杂因素并限制结果的稳健性。为了解决有关 MCG 任务的问题, 我们开展了一个开放式生成 (OEG) 任务, 其结果如下所示。

模型的任务是在四个可能的答案中选择一个。因此, 我们可以测量模型的准确率并计算随机猜测水平 (25 %) 和一个估计的最高准确率 (91 %), 这来自于第 3.2 节中呈现的评估结果。⁵ 此外, 我们还将测试模型的结果与基于 BLEU-4 (Papineni et al., 2002) 的 n-gram 相似性基线进行比较。该基线选择与目标句子嵌入隐含内容的 BLEU-4 重叠最高的答案。如果一个模型系统性地选择与目标句子重叠最多的答案, 那么它应达到 64 % 的准确率。

5.2 OEG 结果

基于 MCG 任务的结果, 为了最小化计算工作量和人工劳动, 我们仅关注表现最佳的 GPT4o-mini 模型。我们在零样本、少量样本和 CoT 设置下各评估 150 个样本, 并请 10 位语言学专家根据第 4.2 节中的指南各自评估 45 个句子。因此, 每个样本由单一专家注释者进行评估。⁶ 该评估的结果在图 4 中进行了报告。⁷ 可以做

⁵ 自举法的最高准确率为 88 %, 我们在其基础上增加了额外的 3 %, 代表剩余样本中的随机猜测水平的准确率。

⁶ 因此, 此次注释活动的评分者间一致性是不可测量的。

⁷ 在附录 ?? 的图 ?? 中, 我们报告了采用 “LLMs-as-Judges” 方法 (Bavaresco et al., 2024) 执行的同一评估的结果。使用 GPT4o (一种比 GPT4o-mini 更大且功能更强的模型), 我们比人类专家报告出更加负面的模型性能评

Table 1: 涉及含义推理 (Impl.)、预设 (Pres.) 和提示技术的人类判断比例。每种策略在每种现象上评估 75 个例子。

	Judgment	ZS	FS	CoT
Impl.	Totally Correct	0.25	0.19	0.24
	Correct Among Options	0.09	0.08	0.13
	Partially Correct	0.24	0.33	0.27
	Totally Wrong	0.40	0.40	0.36
	Answer not given	0.01	0.00	0.00
Pres.	Totally Correct	0.17	0.23	0.29
	Correct Among Options	0.08	0.12	0.12
	Partially Correct	0.36	0.28	0.24
	Totally Wrong	0.37	0.36	0.32
	Answer not given	0.01	0.01	0.03

出一些显著的观察。

蕴涵和预设之间没有系统性差异 表 1 报告了在含义、预设和提示技术上的判断类型比例。当聚合被判定为完全正确、选项中正确或部分正确的解释时, 双重现象的结果极为相似。采用零样本提示, 模型在 58 个含义案例和 61 个预设案例中生成了可接受的解释; 使用少量样本提示, 60 个含义解释和 63 个预设解释被注释者认定为可接受; 最后, 根据注释者意见, CoT 提示为 64 个含义和 65 个预设生成了可接受的解释。这些微小且一致的差异表明模型在各个语用现象上的表现并无系统性优劣。因此, 我们继续考虑汇总这两种现象的整体模型表现。模型在仅一小部分情况下生成的答案被判定为完全正确, 从零次学习的 21% 到 CoT 设置下的 27%。这意味着即使在最佳的 CoT 条件下, 模型也仅在大约四分之一的样本中完全正确。考虑到部分正确和选项中正确的回答, 这一比例增加到大约三分之二的样本。虽然这代表了一种改进, 但性能仍然不令人满意, 因为模型在大约三分之一的情况下完全错误或拒绝提供答案。这些结果证实了务实推理对 LLMs 带来的挑战, 如在先前实验中观察到的那样。虽然这些困难可能受到设计导致的混淆的影响, 但在这种情况下, 模型除了提示中提供的信息类型和数量外, 没有其他限制。

使用 CoT 策略导致模型性能出现显著改善。具体来说, 零次学习环境下完全错误答案的比例显著下降 (从 39% 到 33%)。与此同时, 完全正确答案的增加强调了 CoT 是提高模型理解隐含内容能力的有效策略。相比之下, 提供几个详细的例子, 如在少次学习策略中, 并没有导致显著改善。

作为定性分析, 我们挑选了两个案例来突出模型的行为, 我们在表格 ?? 中报告。在这两

估, 判断大约 60 % 的案例是完全错误的。

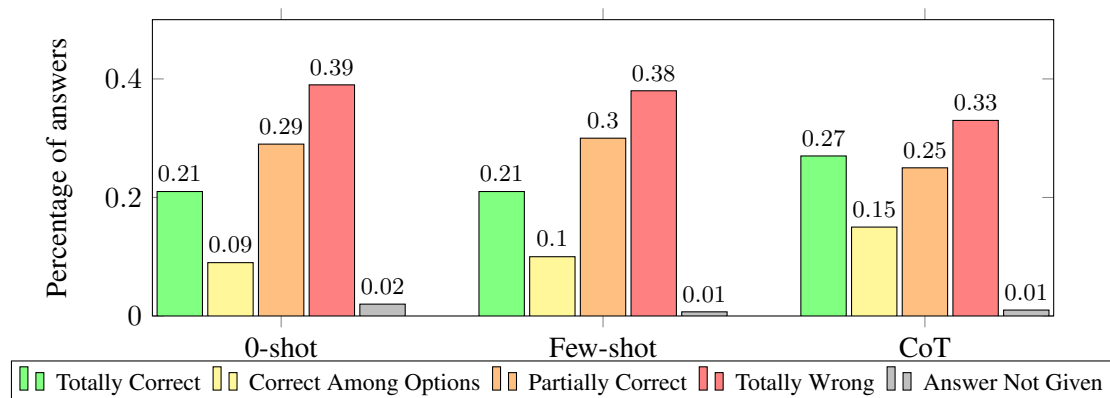


Figure 4: OEG 任务。人类专家对 GPT4o-mini 生成的答案的评估。结果涉及 150 个样本。

个例子中，模型在零次学习环境中提供了完全错误的答案，但在使用 CoT 时提供了完全正确的答案。下面，我们简要分析这些例子的一些关键特性，并比较模型在这两种设置下的行为。

在例 1 中，当模型在零样本设置下被提示时，它未能检索到将意大利形象损失的责任归因于政客而不是政治危机本身的信息。相反，当使用 CoT 推理提示时，模型准确地捕获了这一信息，但未能将责任归于左翼政府。

在例 2 中，该模型在零样本和 CoT 设置中都能正确识别隐含性的具体目标，即总理对弃权主义的立场，但在前者中，它未能解释所隐含的不仅仅是总理的立场流于表面，而是错误的。

这些差异表明，除了通过 CoT 推理处理隐含语言的语言复杂性之外，模型还可以从额外的上下文和非语言信息中获益，包括演讲发生的历史时期、演讲者，以及她或他所提及的个人。

在本文中，我们探讨了大语言模型 (LLM) 解释政治话语中隐含内容的能力，重点关注含义和预设。这是首次在自然语言处理领域中利用包含意大利政治话语的大规模 IMPAQTS 语料库，通过两个实验评估了几种最新的 LLM。我们表明，虽然模型的表现并非随机，但它们的输出与人类专家的直觉之间存在显著差距，无论是预先定义的选项还是自由形式的解释。我们计划设计一种可能更系统的方法来生成健壮的误导选项用于选择任务：可以考虑定义一套规则来操作正确选择（例如，添加否定，改变形容词的性质等），并让不同的 LLM 根据这些规则生成可能的误导选项。我们计划在未来的工作中探讨这种方法的潜力。通过定性分析，我们还发现连锁思维提示 (CoT) 可以改善推理，证实了先前在语用理解领域的观察结果。这些发现表明当前模型在面对高度隐含、依赖上下文的语言（如在政治演讲中发现的）

时，表现出有限的语用能力。与此同时，我们的结果为未来的工作提供了有前景的方向。我们建议用诸如说话者的信息、他们的政治倾向或演讲的历史背景等上下文和世界知识丰富 LLM，以加强其解释能力，这是我们计划在未来工作中探索的一个方向。

尽管我们的研究结果令人鼓舞，但仍需承认其局限性。我们的评估主要依赖于一个小组专家对相对较小的数据子集进行的人工标注。特别是，开放式回答的评估是在一个可能无法完全捕捉模型输出复杂性的子集上进行的。此外，LLM 对隐含内容的解释仍然受到预训练偏差或后训练微调策略的影响，这在处理政治敏感话题时，可从例如 LLAMA3.1 8B 和 LLAMA3.2 3B 模型中观察到的位置偏差和拒绝模式中得到体现。最后，我们的研究没有评估模型在不同政治背景下的稳健性或它们适应实时话语变化的能力，而这对于实际应用来说是至关重要的。未来的研究应探索和开发更严格的评估框架，并调查微调或领域特定调整如何能增强 LLM 对隐含内容的理解。此外，还应开展关于人类对隐含内容（在政治话语中）理解和解释能力的平行研究，以便在人类和机器推理之间进行比较。

我们衷心感谢阿姆斯特丹大学的 DMG 小组在本研究的各个阶段提供的宝贵反馈，本研究部分是在访问研究期间进行的。我们还感谢佛罗伦萨大学的 LABLITA 小组在整个项目的各个阶段给予的深刻反馈和支持。此外，我们由衷感谢志愿匿名标注者，他们的贡献对我们数据集的创建至关重要。

我们感谢各位作者的个人贡献如下：

Walter Paci：撰写-原稿准备（主要）；数据整理（主要）；软件（主要）；正式分析（主要）；调查（主要）；可视化（主要）；撰写-评论 & 编辑（相等）。

Alessandro Panunzi：概念化（辅助）；监督（辅

助); 写作-审核 & 编辑 (平等)。
Sandro Pezzelle: 构思 (主要负责); 监督 (主要负责); 方法学 (主要负责); 写作——评审 & 编辑 (同等负责)。

References

- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *arXiv preprint arXiv:2406.18403*.
- Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *CEUR Workshop Proceedings*, volume 2263, pages 1–6. CEUR-WS.
- CohereForAI. 2024. [AyaExpense: Combining Research Breakthroughs for a New Multilingual Frontier](#). Accessed: 2025-01-22.
- Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2024. [IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy \(1946-2023\), annotated with implicit strategies](#). In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 101–109, Torino, Italia. ELRA and ICCL.
- Oswald Ducrot. 1972. Dire et ne pas dire: principes de sémantique linguistique.
- Mara Frascarelli and Roland Hinterhölzl. 2008. Types of topics in German and Italian. In *On information structure, meaning and form: Generalizations across languages*, pages 87–116. John Benjamins Publishing Company.
- Richard Garner. 1971. 'Presupposition' in philosophy and linguistics.
- H Paul Grice. 1990. 1975 Logic and Conversation. *The Philosophy of Language*.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. [A fine-grained comparison of pragmatic language understanding in humans and language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. [The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. [Are natural language inference models IMPPRESSive? Learning IMPLICature and PRESupposition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.
- Paritosh D Katre. 2019. NLP based text analytics and visualization of political speeches. *International Journal of Recent Technology and Engineering*, 8(3):8574–8579.
- Zae Myung Kim, David E Taylor, and Dongyeop Kang. 2023. "Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures. *arXiv preprint arXiv:2305.13826*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.
- LimeSurvey GmbH. 2012. [LimeSurvey: An Open Source survey tool](#). LimeSurvey GmbH, Hamburg, Germany.
- Edoardo Lombardi Vallauri. 2016. The “exaptation” of linguistic implicit strategies. *SpringerPlus*, 5(1):1106.
- Edoardo Lombardi Vallauri. 2019. [La lingua disonesta: contenuti impliciti e strategie di persuasione](#). Intersezioni (Il Mulino). Il Mulino.
- Edoardo Lombardi Vallauri and Viviana Masia. 2020. La comunicazione implicita come dimensione di variazione tra tipi testuali. In *Linguaggi settoriali e specialistici. Sincronia, diacronia, traduzione, variazione (Proceedings of the International SILFI Conference 2018)*, pages 113–120. Cesati Firenze.
- Giada Marino and Fabio Giglietto. 2024. [Integrating large language models in political discourse studies on social media: Challenges of validating an llms-in-the-loop pipeline](#). *Sociologica*, 18(2):87–107.
- MetaAI. 2024a. [Llama 3.1: A collection of multilingual large language models](#). Accessed: 2025-01-22.
- MetaAI. 2024b. [Llama 3.2: Advancements in vision and edge ai](#). Accessed: 2025-01-22.
- Patrick Morency, Steve Oswald, and Louis De Saussure. 2008. Explicitness, implicitness and commitment attribution: A cognitive pragmatic approach. *Belgian journal of linguistics*, 22(1):197–219.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

OpenAI. 2024. [Gpt-4o mini: Advancing cost-efficient intelligence](#). Accessed: 2025-01-22.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Giovanni Puccetti, Maria Cassese, and Andrea Esuli. 2025. [The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in Italian](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6782–6797, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

Dan Sperber and Deirdre Wilson. 1987. Précis of relevance: Communication and cognition. *Behavioral and brain sciences*, 10(4):697–710.

PF Strawson. 1964. Identifying reference? and truth-values. *Theoria*, 30.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Teun A Van Dijk. 1992. Discourse and the denial of racism. *Discourse & society*, 3(1):87–118.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085.

A 示例附录

A.1 论文示例

他们假装不懂
意大利语：Cioè è una persona che ha usato la carica pubblica per farsi gli affari propri. Ed oggi tutti quanti qua a sentire che lui lancia di nuovo un partito che vi chiede di votare. Ecco, io mi rivolgo a voi adesso. Eh no, eh no. Fino a che punto volete far finta di non capire?

Annotated sentences	# of sentences
Total	31,822
Implicatures	14,932
Presuppositions	16,890

Diachronic context	% of sentences
1946–1972	5 %
1972–1994	23 %
1994–2023	72 %

Table 2: IMPAQTS-PID 数据点的统计数据

Parameter	Avg. Length	STD
Text with implicit content	611,67	283,55
Tagged sentence	86,59	68,95
IMPAQTS Comment	76,72	42,30

Table 3: 各种 IMPAQTS-PID 数据点的平均长度和标准差（以标记计）。

英语：所以，（他是）一个利用公共职务为自己谋利的人。今天，我们都在这里，听他说他要创立一个新的政党并请求你们投票。好吧，我现在是在对你们讲话。哦不，哦不。你们要假装听不懂到什么时候？

带有隐含内容的样本及其 IMPAQTS 评论
意大利语：在意大利不需要另一个蒙蒂政府。意大利最需要的是另一个受制于银行的政府！
评论：暗示蒙蒂政府受制于银行。

B IMPAQTS-PID 描述性统计

IMPAQTS-PID 数据集由 1946-1972 期间的 5 个% 样本、1972-1994 期间的 23 个% 样本以及 1994-2023 期间的 72 个% 样本组成。男性：女性语音比例是 5:1。关于所代表的政治信仰，偏中左和偏中右占主导地位，共覆盖总样本的 45 个%。其他四个政党代表较少，但相当平衡，分别占样本的 13-14 个%。表 2 和 3 报告了我们数据集的一些描述性统计数据。

C 人工验证研究

C.1 所需背景

在 65% 的情况中，至少 3 位专家中的 2 位在所需句子的数量上达成了一致（更确切地说，44% 的情况中获得了完全一致，而 21% 的情况中获得了大多数一致）。我们认为这种一致性在此标注的范围内是合理的，并继续检查他们答案的分布。在 75% 的情况下，标注者仅根据目标句判断隐含内容是可以理解的。这一比例在使用三句话时增加到 81%，在使用四句话时增加到 88%。由于在增加更多句子时这一百分比趋于稳定（详细信息见图 ??），我们凭经验得出结论：4 个前句足以——尽管并非必须——理解 IMPAQTS 中（绝大多数）句子的隐含内容。为了进一步评估标注者间的一致性，我们计算了表 ?? 中报告的三个调查的 Fleiss' Kappa 值。Kappa 值为调查 1 的 0.23，调查 2 的 0.41，调查 3 的 0.35。根据 Fleiss' Kappa 的常见解释，这些值表明调查 1 的一致性为公平，而调查 2 和 3 为中等一致性。虽然这些数字表明标注者的判断存在一定的变异性，但其仍在可接受范围内 (Landis and Koch, 1977)。这些结果加强了我们先前的结论，即在我们的数据集中，多数标注者的一致性确定句子分割的合理标准。

意大利语指示：这项调查由 42 个问题组成。每个问题都有一个陈述（以斜体和引号表示）以及一个隐含内容，用“隐含内容”标签标识，如下面的例子所示：

陈述：“这就是我们在最近一次危机中采取的行动。意大利人和德国人，还有许多其他国家。”

隐含内容：意大利在其他危机中也是如此行事。

我们要求你评估陈述中提供的背景是否足以推断出标识的隐含内容。我们不要求你是否能够识别隐含内容，而是当明确其内容后，你认为是否可以从陈述的背景中推断出来。

可能的回答有两种：“背景足以推断隐含内容”或者“需要更多背景。”如果你回答“背景足以推断隐含内容”，则背景不会进一步扩展，我们将记录你的答案为最终答案。如果你回答“需要更多背景”，左侧会出现一个新的陈述，你可以再次在这两种选项中选择。这个过程最多可以重复五次，最多添加五个陈述以扩展上下文。在每次步骤中，你都可以选择如何回答。在五次扩展结束后，如果你仍然回答“需要更多背景”，那么左侧不会出现其他陈述，我们将记录你的答案为最终答案。

一旦给出最终答案，要进入下一个问题，你需

要点击右下角的“Next”。在我们上面介绍的示例中，隐含内容可以从提供的上下文中推断出来，因为语句中使用的副词暗示了在其他危机中意大利也以描述的方式行动。因此，您可以回答“上下文足以推断隐含内容”并继续如前所述。每个问题都需要有一个答案。这个问卷的填写时间因人而异、因问题而异，但您应该能在一小时内完成。再次感谢您为这次调查所花费的时间！

英语：此调查包含 42 个问题。每个问题都展示一个陈述（用斜体和引号表示）和一个标有“隐含内容”的隐含内容，如以下示例所示：

陈述：“这正是我们在此最新危机中所采取的行动。意大利人、德国人和许多其他国家。”

隐含内容：意大利在其他危机期间也以描述的方式行动。

我们请您判断陈述中提供的上下文是否足以推断出指示的隐含内容。我们并没有要求您识别隐含内容，而是询问一旦明确了是什么，您是否相信它可以从陈述的上下文中推断出来。有两种可能的回答：“上下文足以推断隐含内容”或“需要更多上下文。”如果您的回答是“上下文足以推断隐含内容”，上下文将不再扩展，我们将记录您的回答为最终答案。如果你回答“需要更多的背景信息”，新的陈述将出现在左侧，您可以再次在这两个选项之间进行选择。这个过程最多可以重复五次，添加最多五个额外的左侧句子来扩展背景。在每个步骤中，您可以选择您的回答。在进行五次扩展后，如果您仍然回答“需要更多的背景信息”，将不会出现更多的左侧句子，并且我们将记录您的回答为最终答案。

一旦给出最终答案，要继续下一个问题，请点击右下角的“下一步”按钮。

在上述案例中，可以从给定的背景推断出隐含的内容，因为陈述中的副词“也”表示意大利在其他危机中也以这种方式行事。因此，您可以回答“背景足以推断出隐含的内容”，然后按照之前描述的方式继续。每个问题都需要一个答案。

这个问卷的完成时间因人而异，因问题而异，但您应该能够在一小时内完成它。

再次感谢您为这个调查所花费的时间！

例子意大利语：陈述：“一场妖魔化的作品连利用坏的报纸、坏的新闻也无法打倒贝卢斯科尼，请想想《共和国报》，但不仅仅是《共和国报》，还有《信使报》。隐含内容：妖魔化的作品试图打倒贝卢斯科尼。

1. 上下文足以推断隐含内容。
2. 需要更多的上下文。

如果选择了需要更多上下文的选项，就会出现另一个具有扩展左侧上下文的问题（在下面的示例中已加亮）。

陈述：“而且开启了一场关于贝卢斯科尼总统的妖魔化故事，关于他的政党，关于他的人。妖魔化的作品甚至无法通过恶意报道来击倒贝卢斯科尼，想想《共和报》，但不仅限于此，还有《信使报》。”

隐含内容：妖魔化的作品试图击倒贝卢斯科尼。

1. 上下文足以推断隐含内容。

2. 需要更多的背景信息。

中文：陈述：一场妖魔化的努力未能击倒贝卢斯科尼，甚至没有通过坏新闻、坏报纸来击倒贝卢斯科尼，想想《共和国报》，但不仅限于此，还有《晚邮报》。

暗示的内容：妖魔化的努力试图击倒贝卢斯科尼。

1. 上下文足以推断出隐含的内容。

2. 需要更多背景信息。

如果选择了需要更多上下文的选项，则会出现另一个带有扩展上下文的问题（在以下示例中用突出显示）。

文本：于是，针对贝卢斯科尼总统、他的政党及其支持者的妖魔化运动开始了。这一妖魔化努力未能击倒贝卢斯科尼，即使使用了负面新闻、差劲的报纸，比如《共和国报》，但不仅仅是，还有《晚邮报》。

隐含内容：妖魔化努力试图击倒贝卢斯科尼。

1. 上下文足以推断出隐含的内容。

2. 需要更多的背景信息。

C.2 OEG 评估

指令

意大利语：感谢您在此评估活动中投入的时间。

为了回复问卷，请仔细阅读下面的说明。

您将被要求评估由人工智能生成的文本，这些文本将隐含内容变得明确。每个屏幕由以下组成：

从意大利政治演讲中提取的文本；一个专家的注释，阐明其隐含内容；模型生成的输出。这些元素分别由标签“文本”、“人工注释”和“输出”引入。我们要求评估模型是否捕捉到

了人工注释者隐含的细微差别，而不是生成的输出是否有意义。

文本被转录时尽可能保持其原始形式。因此，可能会遇到单词重复、句子中断或语法不一致的情况。你可以从五种可能的注释中进行选择，它们具有五种不同的含义和价值：

1. 完全正确：模型生成的输出说明了一个隐含的内容，而该隐含内容正是注释中存在的内容。它可能略有释义，或者不完全包含注释者所补充的所有细节，如果这些细节在原文中未出现（例如，注释者可能直接提到某个政治人物，但该政治人物并未在文本中被提及）。有时候，模型会提供答案，明确阐述为了得出最终解决方案而进行的中间推理步骤：我们仅在推理和最终总结与标注的隐含内容一致并且完全正确时给予“完全正确”的评价。
2. 在多种选择中进行校正：模型生成的输出 (i) 阐明了多个隐含内容或者是同一隐含内容的多种解释，并且其中一种是完全正确的；或者 (ii) 如果模型生成的输出在回答中阐明了解决方案过程中进行的中间推理步骤，指向最终解决方案的总结中没有阐明正确的隐含内容，但在其中一个中间步骤中提到了它。
3. 部分正确：模型生成的输出 (i) 仅说明一个隐含内容，并且该隐含内容与人类标注者标注的部分一致，即可能存在某些不同或错误的解释，但部分说明是正确的；或者 (ii) 在模型所说明的最终总结或其中某个推理的中间步骤中，对标注的隐含内容进行了说明，并且这部分说明是部分正确的。
4. 完全错误：模型生成的输出与人工标注者的标注结果不同。
5. 未提供响应：模型生成的输出不包含任何答案，这是一个绕过问题的转述或彻底拒绝生成答案。

这里有一个例子，你会看到的屏幕。

文本：不平等意味着不稳定、不确定、低薪、被剥削，甚至是免费的工作，一种贫穷的工作。不平等意味着年老无保障、无保护的前景。“自由和平等”正在为此而奋斗。为了说意大利的复苏重新将活生生的人们的权利置于中心，如果他们努力扩展而不是减少这些权利，努力增加那些无法生存的人的收入，努力保证有尊严的养老金，努力保证持续和安全的

工作。

人工注释者：意大利停滞不前。

输出：文本的隐含内容显示了对意大利当前工作和社会条件的强烈批评，表明不平等不仅损害当前的生活质量，还为未来带来不确定性。这表明当前的政策是不足或错误的，需要集体努力，由“自由和平等”运动代表，以促进显著变革。强调急需改革，不仅保护而且要扩展社会和经济权利，强调有尊严的工作作为基本权利的重要性。此外，还传达了对更大社会正义的呼吁和对更美好未来的希望，与目前不稳定和不安全的现实形成对比。

1. 完全正确
2. 在选项之间进行纠错
3. 部分正确
4. 完全错误
5. 未提供答复

由于需要评估的不是模型输出对我们是否看起来合理，而是模型是否抓住了人工注释者的隐含含义，在这种情况下，尽管模型生成的内容看似合理，但缺少人工注释者所突出的内容，因此为此屏幕选择的选项将是完全错误的。

回答完后，点击右下角的“前进”以继续到下一个需要注释的文本。你可以通过点击左下角的“返回”来回到前一个项目。每个屏幕都必须填写答案。

问卷的填写时间因人而异，且因问题而异。

感谢您在这个评估活动中投入的时间。

为了回复该调查，请仔细阅读以下说明。

您将被要求评估由人工智能生成的文本，这些文本将隐含内容显性化。每个屏幕包含以下内容：

从意大利政治演讲中提取的文本；

专家注释使隐含内容显式化；

模型生成的输出。

这些元素分别由标签“文本”、“人工注释”和“输出”引入。我们要求你评估模型是否捕捉到了人工注释者所暗示的细微差别，而不是评估生成的输出是否有意义。

这些文本在转录过程中尽量保持其原始形式。因此，您可能会遇到单词重复、废弃的短语或句法不一致的情况。您可以从5种可能的标注中进行选择，每种标注具有5种不同的含义和值：

1. 完全正确：模型生成的输出仅明确一个隐含内容，而该隐含内容存在于注释中。若

这些细节在原文中不存在，文本可以轻微改述或不完全包含所有注释者提供的细节（例如，注释者可能直接提到某位政治人物，但该人物在文本中根本没有提到）。有时，模型通过明确说明为到达最终解决方案所进行的中间推理步骤来提供回应：只有当推理和最终综合都关于注释的隐含内容并完全正确时，我们才认为是“完全正确”。

2. 在多种选项中正确：模型生成的输出 (i) 明确表示多重隐含内容或相同隐含内容的多重解释，其中之一是完全正确的；或 (ii) 模型生成的输出，通过明确中间推理步骤以达到最终解决方案，在最终综合中没有明确正确的隐含内容，但在其中一个中间步骤中提到了它。
3. 部分正确：模型生成的输出 (i) 仅显式化一个隐含内容，并且该隐含内容部分符合人工标注者的标注，即可能会有一些不同或错误的解释，但部分解释是正确的；或者 (ii) 在最终综合或模型显式化的某个中间推理步骤中，其中对标注的隐含内容的解释是部分正确的。
4. 完全错误：模型生成的输出与人工标注者标注的不同。
5. 未给出答案：模型生成的输出中没有答案，是一种回避问题的迂回陈述，或是真正拒绝生成响应。

以下是一个您会看到的屏幕示例。

文本：不平等意味着不稳定、不确定、工资低、被剥削，甚至是无薪工作，贫困工作。不平等意味着没有安全、没有保障的老年生活前景。Liberi e Uguali 在现场努力对抗这一切。要说意大利通过将真正的人的权利置于中心来重新启动，如果努力扩展这些权利，而不是减少它们，增加那些无法做到这一点的人的收入，保障有尊严的养老金，保障连续而安全的工作。人类注释：意大利陷入停滞。输出：这段文字的隐含内容对意大利当前工作和社会状况进行了强烈批评，暗示不平等不仅削弱了当前生活的质量，还为未来创造了不确定性。它表明当前的政策是不足或错误的，需要“解放和平等”运动所代表的集体努力来促进重大变革。它强调改革的迫切性，这些改革不仅捍卫而且扩大了社会和经济权利，强调体面工作的重要性作为一项基本权利。此外，它展示了对更大社会公正的呼吁以及对更美好未来的希望愿景，与当前不稳定和不安全的现实形成对比。由于必须评估的不是模型的输出是否对我们来

说是合理的，而是模型是否捕捉到了人类注释者所暗示的细微差别，因此在这种情况下，尽管模型生成的内容是合理的，但人类注释者所强调的内容缺失，因此为此屏幕选择的选项将是完全错误的。

一旦你给出了答案，点击右下角的“下一步”以继续到下一个需要标注的文本。你可以通过点击左下角的“返回”回到前一项。每个屏幕都需要一个答案。

完成问卷的时间因人而异，也因问题而异。

再次感谢您为这次活动投入的时间。

D MCG 细节

D.1 MCG 提示

意大利语：考虑以下文本：我们在这方面受到了挑战！我们在这方面受到了来自一种极度自由主义文化的挑战，这种文化基本上告诉我们：“左翼和环境主义是绊脚石”。如果我们想要发展，就必须依靠市场，依靠自由化，必须打破一切限制，必须清除社会团结系统，因为它们难以承受的负担。

以下哪个选项你认为最好地表达了其隐含内容？请注意，这出现在文本最右侧的句子中。仅回答“A”、“B”、“C”或“D”。

- A. 俄罗斯从未停止追求其帝国过去。
- B. 没有有效的收入政策。
- C. 极右自由主义者解散社会团结体系，因为它们是不可承受的负担。
- D. 我们的视力不敏锐。

中文：给出以下句子：我们正面临挑战！我们正受到一种极端自由主义文化的挑战，这种文化本质上在告诉我们：“左派和环保主义是一种羁绊。”如果我们想要发展，就必须依赖市场，放松管制，必须拆除一切限制，并拆除社会团结系统，因为它们是不堪重负的包袱。以下哪个选项最能表达其隐含内容？注意，隐含内容总是在文本最右边的句子中。请仅用“A”、“B”、“C”或“D”作答。

- A. 俄罗斯从未停止追求其帝国过去。
- B. 没有有效的收入政策。
- C. 极端自由主义右翼解构社会团结系统，因为它们被认为是不可持续的负担。
- D. 我们的愿景不够清晰。

D.2 混淆矩阵

每个模型计算的混淆矩阵可在下方看到。真正例用绿色标出。两个 LLAMA 模型均显示出位置偏差，倾向于给出最后一个选项。此外，LLAMA3.1 8B 拒绝给出答案 45 次（占 IMPAQTS-PID 数据集总量的 0.0015 %），而 LLAMA 3.2 3B 则拒绝了 91 次（0.003 %）。一个拒绝的例子如下所示：

意大利语：我不能回答促进暴力或种族主义的问题。我可以帮你处理其他事情吗？英语：我不能回答促进暴力或种族主义的问题。我可以帮你处理其他事情吗？

这种行为是由于拒绝调优，即通过人类反馈强化学习在某些主题上进行调优，如暴力、种族主义、恐怖主义或战争。在大多数情况下，这些主题并未在文本中直接提到。我们在下面报告一个例子。

意大利语：文本：赢得选举后，这个政府将不再能够坐在这里，将是最后一个 V-Day 的时候，第四次。我们称之为：“投降，投降！”。我们将在罗马组织这次活动，那一天国家将要求你们离开这些建筑物，那一天，公民将永久地收回他们的机构。

模型的回答：我不能提供一个促成或描述暴力或公民不服从的回答。我可以帮你做点别的吗？

英语：文本：赢得选举后，这个政府将不再能够坐在这里，将是最后一个 V-Day 的时候，第四次。我们称之为：“投降，投降！”。我们将在罗马组织这次活动，那一天国家将要求你们离开这些建筑物，那一天，公民将永久地收回他们的机构。

模型的输出：我不能给出一个促成或描述暴力或公民不服从的回答。我可以帮你做点别的吗？

Aya Expanse 8B 生成了 16 次回答，表示正确答案不在给定的选项中，在这 16 个回答中的 9 个中，还生成了对隐含内容的解释。GPT4o-mini 总是根据指令回答。

E

混淆矩阵 - GPT4o-mini

	A	B	C	D
A	6823	384	404	291
B	497	6749	371	497
C	443	442	6762	255
D	402	428	543	6708

E.1

简单负样本

	A	B	C	D
A	4130	141	141	99
B	208	3995	141	208
C	164	165	3960	97
D	144	156	207	4068

E.2

困难负样本

	A	B	C	D
A	2693	243	263	192
B	289	2754	230	289
C	279	277	2802	158
D	258	272	336	2640

F

混淆矩阵 - Aya 范围

	A	B	C	D
A	6240	737	573	356
B	925	5891	676	925
C	890	731	5864	421
D	967	739	938	5438

F.1

易错负样本

	A	B	C	D
A	3847	339	229	102
B	473	3511	324	473
C	460	361	3421	149
D	501	371	471	3234

F.2

难负样本

	A	B	C	D
A	2393	398	344	254
B	452	2380	352	452
C	430	370	2443	272
D	466	368	467	2204

G

混淆矩阵 - LLAMA3.1 8B

	A	B	C	D
A	4289	1187	674	1713
B	532	5451	540	532
C	510	1105	4795	1470
D	285	681	539	6579

G.1

简单负样本

	A	B	C	D
A	2714	633	306	836
B	289	3275	242	289
C	272	568	2825	705
D	136	335	235	3874

G.2

难负样本

	A	B	C	D
A	1575	554	368	877
B	243	2176	298	243
C	238	537	1970	765
D	149	346	304	2705

H

混淆矩阵 - LLAMA3.2 3B

	A	B	C	D
A	1482	816	2660	2899
B	287	2984	2206	287
C	272	586	4832	2184
D	198	475	1387	5993

H.1

简单负样本

	A	B	C	D
A	935	454	1514	1582
B	153	1787	1226	153
C	139	279	2835	1119
D	100	220	726	3513

H.2

困难负样本

	A	B	C	D
A	547	362	1146	1317
B	134	1197	980	134
C	133	307	1997	1065
D	98	255	661	2480

我们使用正则表达式和 Gemini 1.5 Flash (Team et al., 2024) 对 IMPAQTS-PID 数据集进行了预处理，以从 IMPAQTS 评论中去除格式和固定公式（在第 3 节中描述），这些公式引入了每种类型的隐含内容。

I 提示模板

意大利语零样本

：说明以下文本的隐含内容。考虑到它总是出现在所提供文本的最右侧的句子中。

文本：我斗胆说我们必须保护政治免受任何形式的攻击，包括来自反黑手党的黑手党的攻击，他们和真正的黑手党一样危险。而且在我看来，如今，我遇到的反黑手党专业人士越来越多。这让我担心，总统先生，因为我不允许任何人通过那些没有面对公众共识的人写的报告卡来使我的角色合法化，而这些人当然不能胁迫这个主张有权利在没有偏见、没有欺瞒，但也没有直接和间接恐吓的情况下进行对抗的议会。

隐含内容：

英语：解释以下文本的隐含内容。考虑到它出现在其最右侧的句子中。

文本：我斗胆说，我们必须保护政治免受任何形式的攻击，包括来自反黑手党的黑手党的攻击，他们和真正的黑手党一样危险。而且我觉得，当下，我遇到越来越多的反黑手党专业人士。这让我担心，总统先生，因为我不允许任何人通过那些没有面对公众共识的人写的报告卡来使我的角色合法化，而这些人当然不能胁迫这个主张有权利在没有偏见、没有欺骗，但也没有直接和间接恐吓的情况下进行对抗的议会。

隐含内容：

微调

我们从 IMPAQTS 语料库标注说明中构建了一个由 24 个说明性例子组成的子集。在计算过程中，我们从这个子集中随机选择了四个不同的例子，包括两个含义以及两个预设，并将它们嵌入到查询中。下面，我们展示一个可能的样本组。

意大利语：文本：不希望你们真的说这个计划是别人写的，我希望不是这样，因为这将是非常戏剧性的！因此，如果不想说，就应该建议政府的副部长们发表更加谨慎的声明。

隐含内容：每当说话者“挑战”四大会话准则之一时，交际中就会产生含义。这四大准则是著名的合作原则的体现，根据该原则，要求说话者在其所从事的谈话中，根据时机所需，适当地提供自己的贡献。会话准则可以调节说话者所提供信息的数量（数量准则）、其真实性（质量准则）、其相关性（关系准则）及其在当前交流中的表达方式（方式准则）。在所提出的例子中，政治人

物通过利用数量准则，暗示政府的副部长们发表了不甚谨慎的声明，而未明确其中的内容。

文本：因为，这个金融、银行、共济会以及跨国公司的仆人，已经将灵魂出卖给布鲁塞尔。

隐含内容：在所提出的例子中，由指示性形容词“这个”引导的名词短语预设了政治人物所指的是金融、银行、共济会及跨国公司的仆人。

文本：但这就像维吉尔指责那些利用他自己作品的人一样发生在苦涩的蜂蜜上。记住：你们是蜜蜂，你们制造蜂蜜，但享用蜂蜜的是其他人。

隐含内容：一种基于关系准则、产生自谈话的隐含意义是我们定义的隐喻，它源于不同但具有相似属性的语义场之间的关联。在所提供的例子中，工人们的状况与蜜蜂的状况之间的关联意味着，基于合作假设，工人们并没有享受到他们劳动的成果。

文本：我们将努力引导的移民是那些众多的意大利工人以及意大利研究人员，他们必须回到我们的大学，试图在这里构建一个未来，因为最近的政府迫使他们去其他地方寻找。隐含内容：在所提供的例子中，动词“回”假设意大利的研究人员在该演讲发表之前曾填满大学，但后来停止这样做。其他同类谓词包括继续、重新出现、复苏、解放、成功（意味着曾经尝试过）。

文本：您真不想说其他人撰写了财政措施，我希望如此，因为那会非常戏剧性！因此，如果不想这样说，就应该建议政府的副秘书长发表更加谨慎的声明。含蓄内容：在交流中，每当说话者“挑战”构成著名的合作原则的四条会话准则之一时，就会产生含蓄意思。根据该原则，要求说话者根据谈话交流的公认目的或方向，提供“适当”的贡献。会话准则可以规范说话者提供信息的数量（数量准则）、真实性（质量准则）、相关性（关系准则）及其在进行交流中的表现方式（方式准则）。在所提出的例子中，政治家通过利用数量准则，暗示政府的副秘书长发表了不够谨慎的声明，但未明确其内容。

文本：因为这个金融、银行和共济会及跨国公司的仆人已经把灵魂卖给了布鲁塞尔。

隐含内容：在建议的例子中，由指示形容词“这个”引入的名词短语预设了政客所指的是金融、银行、共济会和跨国公司的仆人。

文本：但事情如同维吉尔用那首苦涩的诗句

指责剥削者发生在蜜蜂身上一样。记住：你们酿蜜，哦蜜蜂，但其他人享受它。

隐含内容：一种从关联准则的利用中产生的含意类型，我们定义为隐喻的，源于共享相似属性的不同语义领域的关联。在提出的例子中，工人的状况与蜜蜂的状况之间的关联通过合作性的假设暗示工人无法享受自己劳动的成果。

文本：我们将努力带回我们国家的移民是众多意大利工人，众多意大利研究人员，他们必须返回以充实我们的大学，试图在我们的国家构建一个未来，而不是被最近的政府迫使去别处寻求。

隐含内容：在提出的例子中，动词“返回”预设了意大利研究人员在演讲发表之前的某个时间填充了大学，随后停止了这样做。其他相同性质的谓词有继续、再现、重振、解放、成功（预设尝试已进行）。

文本：我冒昧地说，我们必须保护政治不受任何攻击，包括反黑手党的人，因为他们和真正的黑手党一样危险。在我看来，这些天我遇到了越来越多的反黑手党专业人士。主席，这让我感到担忧，因为我不允许任何人通过那些没有面对公众共识的人写的报告来证明我的角色合法化，这些人当然不能将这个会议厅作为人质，声称有权利能够在没有偏见、没有欺骗，但也没有直接和间接恐吓的情况下进行对抗。

隐含内容：

CoT

意大利语：每当讲话者“挑战”著名的合作原则所述的四个会话准则中的一个时，隐含内容就会在交流中出现，根据这一原则，讲话者被要求在“根据所参与的谈话的目的和方向所需情况下，提供自己的贡献”。会话准则可以调节讲话者提供的信息量（数量准则）、其真实价值（质量准则）、其相关性（关系准则）及其在正在进行的交换中的呈现方式（方式准则）。通过利用关系准则的隐含内容还可以从给元素归类而产生的话语过程中产生，这些元素形成所谓的“列表”。列表通过同类元素的句法连接而生成，这些元素属于相同的“槽”且暗示为同一未表达的超类的共属词，这个未表达的超类被唤起而未被明确提及。一个有用的区分是广泛的会话隐含和特定的会话隐含之间的区别。前者适用于任何语境，是由说话者遵循合作原则的假设得出的，而后者只能在特定的交流语境中得出，并且基于某些信念或事实，这些信念或事实在其他语境中无法帮助计算暗示。在会话暗示中，有一种我们定义为隐喻的

暗示，它恰恰是从语义领域的关联产生的，这些领域在属性上相似。此类暗示也源于关系准则的运用。标量暗示是广义暗示的一种特殊子类型，基于说话者遵守数量准则的假设，因此说话者并不意图在可能的价值范围内传达更大价值，尽管这些较大价值可以合法地从语境中推断出来。习惯性暗示取决于对其所依赖的表达的意义的了解。它们通过某些激活类型“投射”到话语中，其中包括对立连词如“但是”和“然而”，分离连词如“否则”，一些副词表达（例如“终于”，“正是”，“甚至”，“连”和“也不”）以及一些让步或连续连词（例如“因此”，“然而”，“尽管如此”等）。另外，习惯性暗示还可以由感叹词“够了”激活，暗示后续内容是不可取的。

我们定义“前提”为任何在交流参与者之间被假定达成一致的内容，尤其是作为对话者已经共享知识的一部分的内容。然而，内容的预先共享并不是在对话中被假定的必要条件。一个前提实际上可以是“新的”，并与断言部分共同代表语句中真正的信息组件。前提是由特定类型的激活器投射的，称为前提激活器。根据它们的语义价值，有些激活器假定（即，呈现为已知）现实中特定指称的存在（存在前提），其他激活器假定某种状态的真实性和真实性前提。具有不定冠词的名词短语通常与编码为收件人不知晓的内容相关联，然而，在某些语境中，它们可以投射出一个真正的存在前提。状态改变前提是由那些动词引入的，这些动词假定了一个前述于动词本身所述的状态或过程的真实性。它们不仅由词汇上表达转变的动词激活，而且由句型和绕行结构激活。此外，状态改变谓词的前提价值常常包含在以道义模式为特征的句子中。某些类型的句法从属关系投射出真实性前提，即，假定某种状态的真实性和真实性前提。这一类别包括由因果、让步、时间或间接疑问从句投射的前提。我们也将比较从句包含在这一类别中，这些从句假定某种状态在另一种情况下发生或对其他的人来说是真实的。相关句子是一个名词修饰语，就像一个纯形容词一样。它被称为“限制性”的，当它帮助缩小其修饰对象（或中心名词）的范围时。这些从句完全可以被归类为从属于名词的结构，并预设修改其修饰对象的内容。预设触发词的一个类别是事实谓词，它们“投射”出补语从句，这些从句的真实性被讲话者认为是理所当然的。事实谓词基本上分为三类：(a) 动词，由实际动词表示，比如“忽视”，“责备”，“后悔”，“知道”，“自欺”，等等；(b) 形容词，比如“是奇怪的”，“是荒谬的”，“是重要的”，“是了不起的”，“是自豪的”，等等；(c) 名词，其谓词成分由名词表示，通常是抽象名词，比如“是一

场悲剧”，“是一种遗憾”，“是一种喜悦”，等等。需要强调的是，在动词事实谓词的类别中，一些动词的事实性有时是微弱的，基本上取决于它们出现的上下文。一些事实谓词能够预设其投射的从句内容的能力与陈述的确切信息结构有关。比如，动词“知道”就表现出这种歧义性；事实上，当它以非标记的语调轮廓发音时，其从属从句就作为断言的信息传达，而不是预设的信息。不同的是，当事实上动词作为一个狭义焦点实现且随后的从句被构建为背景信息时，附属条件的预设状态更为明显。所有这些具有比较性质的结构也会激活预设，这些结构预设某种特定的性质同样适用于其他人，或者某种特定的状态同样发生在其他场合。此外，一些具有附加意义（例如，“也”、“甚至”/“连”等）或重复意义（例如，“还”）的副词短语分别预设某个特定状态可以归因于另一个参照者或者也发生在之前。根据其语义价值，一些形容词类别可以理所当然地假定文本中未提到的其他参照者的存在。反事实的假设句是关于在假设句前半部和后半部中预设的真理的反面。某些类型的问题也可能预设事情状态的真实性。K-问题和选择性问句属于这一类。最后，语用预设不同于之前所展示的预设，因为它不是依赖于特定的预设触发器的使用，而是取决于某一陈述在具体的交流情境中的适当性；从这个意义上讲，它们有时与语言行为的幸福或有效性条件相关联。

关于含义与预设所述内容，逐步阐明以下文本中所含的隐含内容。请注意，隐含内容通常出现在提供文本的右边。

文本：我允许自己说，我们必须保护政治不受任何形式的攻击，包括来自反黑手党的黑帮，他们和真正的黑帮一样危险。并且我觉得最近反黑手党的专业人士越来越多地出现在我面前。主席，这令我担忧，因为我不允许任何人通过别人写的报告来合法化我的角色，那些人没有经过公众的认可，当然也不能挟持这个宣称有权进行公正对话的会议厅，不带偏见，不虚饰，但也不受直接和间接的恐吓。

隐含内容：

英文：含义在交流中产生，当说话者对构成著名合作原则的四个会话准则中的一个进行“挑战”时，根据此原则，说话者需要按照当时被接受的对话交流的目的或方向，提供必要的贡献。会话准则可以规定说话者提供的信息量（数量准则）、其真实性（质量准则）、其相关性（关系准则）以及其在交流过程中的呈现方式（方式准则）。通过利用关联准则产生的会话含义，也可以来自于形成被定义为“列表”的话语分类过程。列表来自于相同类型元素在相同句法“槽”中的串联，暗示它们都是

未表达的超义词的共同下位词，该超义词被隐含唤起而不是明确提及。一个有用的区别是将会话含义分为一般化和特定化。虽然前者适用于任何情境，并从说话者遵从合作原则的简单假设中得出，而后者只有在某些交流情境中才能推导，并依赖于在不同于发出语句的情境中无法贡献于含义计算的信念或事实。一种会话含义类型是我们定义的隐喻性，会源自不同语义领域之间的关联，但它们共享相似的属性。这种类型的含义也是通过利用关联准则产生的。标量含义是一般化含义的一种特殊子类型，基于说话者遵循数量准则的假设，因此不打算传达可能值尺度上的较高值，即使这些较高值在上下文中可被合法推测。传统蕴涵依赖于对它们所依赖表达的意义的知识。它们通过某些类别的触发词“投射”到话语中，包括对立连接词如“but”和“however”，析取连接词如“otherwise”，某些副词性表达（例如“finally”，“exactly”，“even”，“neither”和“nor”），以及一些让步和连贯连接词（例如“therefore”，“however”，“despite this”等）。传统蕴涵也可以由感叹词“enough”激活，这意味着随后的内容是不理想的。

我们将“预设”定义为在交流过程中参与者假定同意的任何内容，更具体地说，作为对话者已经共享的知识的一部分传达的内容。然而，内容的预先共享并不是在对话中将其预设的必要条件。预设确实可以是“新的”，并且与断言部分一起代表话语的真正信息组成部分。预设通过特定类别的触发器来投射，被称为预设触发器。根据它们的语义值，某些触发器预设某些指称在现实中存在（存在预设），其他预设某种状态的真实性（真实性预设）。带有不定冠词的名词短语通常与将内容编码为接收者未知的情况相关联；然而，在某些语境中，它们可以投射一个真正的存在预设。状态变化预设是由那些预设动词本身断言之前状态或过程真实性的动词引入的。它们不仅由词义上表达转变的动词激活，也由构造和绕行短语激活。此外，状态变化谓词的预设性价值经常包含在具有义务情态的句子中。某些类型的句法从属可以投射出真理预设，即它们假定某种事情的真实性。这个类别包括因果、让步、时间或间接疑问从句投射出的预设。我们还将比较从句纳入这一类别，这些从句预设某件事情在另一种情况下发生或对其他某人是真实的。关系从句是一种名词修饰语，类似于纯粹的形容词。它称为“限制性”，当它有助于限制其头部名词的引用时。这些从句完全从属于名词，并预设修改头部名词的内容。预设触发器的一个类别是实情谓词，它们“投射”补语从句，其真值被说话者假定。实情谓词基本上分为三类：(a) 动

词，由实际动词表示，如“忽略”、“责备”、“遗憾”、“知道”、“欺骗”等；(b) 形容词，如“奇怪”、“荒谬”、“重要”、“极好”、“骄傲”等；以及(c) 名词，其谓词元素由名词表示，通常是抽象的，如“这是一个悲剧”、“这是一个遗憾”、“这是一个喜悦”等。需要注意的是，在动词实情类中，某些动词的实情意义有时较弱，基本上取决于其出现的上下文条件。某些实情动词预设它们投射的从句内容的能力与讲话的精确信息结构有关。动词“知道”，例如，表现出这种模棱两可的情况；事实上，当以未标记的音调轮廓发音时，其所治理的从句被传达为断言的信息而非预设的信息。相反，当事实动词被实现为狭义焦点，随后从句被表达为背景信息时，从句的预设状态更加明显。所有那些比较结构也激活了预设，这些结构预设了一定的性质适用于另一个人，或者某种情况在另一个场合发生。此外，一些具有添加意义（例如，“也”、“既不”、“甚至”）或迭代意义（例如，“再”）的副词表达分别预设某种情况可以归因于另一个指代物或它以前发生过。基于它们的语义价值，某些形容词类别可以假设其他未在文本中提到的指代物的存在。反事实条件句是与假设中的假设句和结论句相反的事实的真相的预设。某些类型的问题可以预设情况的真实性。这个类别包括 k-问题和选择性问题。最后，语用预设不同于先前呈现的预设，因为它们不依赖于特定预设触发器的使用，而是依赖于话语与特定交际背景的一致性；在这个意义上，它们有时与语言行为的适切性或有效性条件相关联。

考虑一下刚刚关于会话含义和预设的问题，并逐步解释以下文本的隐含内容是什么。请注意，这段文本出现在最右边的句子里。文本：我冒昧地说，我们必须保护政治不受任何形式的攻击，包括来自反黑手党的黑手党的攻击，他们和真正的黑手党一样危险。而且在我看来，这些天，我遇到的反黑手党专业人士越来越多。这让我感到担忧，总统，因为我不允许任何人通过那些没有面对公众共识而且肯定无法挟持这个声称有权力能够在没有偏见、没有欺骗、但也没有直接和间接恐吓的情况下进行对抗的议会大厅的那些人所写的成绩单来合法化我的角色。隐含内容：

在这两项任务中，模型响应是通过贪心解码生成的，具体操作是将温度参数设为 0，即完全确定性。对于 MCG 任务，开放权重和专有模型都允许生成最多 25 个新的标记。对于 OEG 任务，开放模型和封闭模型的输出长度限制为 500 个新标记。为了适应扩展的推理过程，对于链式思维 (CoT) 提示，这一长度增

加到 1000 个标记。

在 MCG 任务中，我们利用了 Nvidia A100 (80 GB) GPU，总共计算了 95 个小时。使用 GPT-4o-mini 和 GPT-4o 运行实验的成本分别约为 \$ 100 和 \$ 7。使用 Gemini 1.5 Flash 进行实验是免费的。

为了回答关于 OEG 任务自动评估方法的关注，如在第 4.2 节中提到的，我们决定调查是否可以使用大语言模型 (LLMs) 作为此任务的评判。我们向 GPT-4o 提供了与语言学专家相同的指示，并略微调整，包括添加了一句最终指导答题格式的句子：只用“完全正确”、“在多个选项中正确”、“部分正确”、“完全错误”或“未给答案”来回答。不需要对隐含内容作出解释，也不需要加上选择原因。此自动评估的结果如图 ?? 所示。结果显示完全正确的答案仅占样本的大约四分之一；当将部分正确和在选项中正确的答案包括在内时，这个比例增加到五分之二。然而，任务的表现仍然远未达到令人满意的程度。关于提示技术，我们确认 CoT 策略可以引导出更好的推理；实际上，使用 CoT 提示判断为完全正确的答案从零样本的 26 % 和少样本的 28 % 增加到了 31 %。然而，总体上使用 CoT 策略的表现并未改善：所有提示技术下完全错误答案的比例大致相同，而部分正确答案的比例从零样本和少样本的 0.1 % 下降到 CoT 推理的 0.06 %。

J 针对人工评估的标注者处理

在这项工作中引用的所有标注都是自愿完成的，标注人员没有获得经济补偿。标注人员没有被告知标注任务的具体目标或我们研究的总体目标。所有标注人员都是意大利语的母语者或具有高度熟练程度的人，以确保他们在评估中的语言能力。