scientific reports



OPEN

# 使用大型语言模型进行口头谎言检测

Riccardo Loconte , Roberto Russo , Pasquale Capuozzo , Pietro Pietrini & Giuseppe Sartori

事实证明，人类通过直觉判断检测欺骗的准确性不会超过偶然水平。因此，已经开发了几种采用机器学习和 Transformer 模型的自动口头谎言检测技术，以达到更高的准确度。这项研究首次探讨了大型语言模型 FLAN-T5（小尺寸和基本尺寸）在三个英语数据集（包括个人数据）中的测谎分类任务中的性能。

观点、自传体记忆和未来的意图。在执行文体分析来描述三个数据集中的语言差异后，我们使用 10 倍交叉验证在三个场景中测试了小型和基本尺寸的 FLAN-T5：一个使用来自同一数据集的训练集和测试集，一个使用来自同一数据集的训练集和测试集，训练集来自两个数据集，测试集来自第三个剩余数据集，其中一个训练集和测试集来自所有三个数据集。我们在场景 1 和 3 中达到了最先进的结果，优于之前的基准。结果还表明，模型性能取决于模型大小，模型越大，性能越高。此外，还进行了文体计量分析以进行可解释性分析，发现与认知负荷框架相关的语言特征可能会影响模型的预测。

测谎涉及确定给定通信的真实性的过程。在制作欺骗性叙述时，说谎者会采用言语策略在互动伙伴中制造错误的信念，从而陷入特定的暂时的心理和情绪状态。因此，温德奇假说表明，欺骗性叙述在形式和内容上与真实叙述不同。鉴于其在法医和法律环境中的重要和有前景的应用，该主题在认知心理学领域一直在不断地研究和发展。其潜在的关键作用是在调查和法律诉讼过程中确定证人和潜在嫌疑人的诚实，影响调查信息收集过程和最终决策水平。

数十年的研究重点是识别欺骗的言语线索，并开发有效的方法来区分真实和欺骗性的叙述，这种言语线索充其量是微妙的，通常会导致天真的人和专家的表现略高于偶然水平。社会心理学对这种令人不满意的人类表现的一个潜在解释是人类内在的真相偏见倾向，即诚实推定的认知启发法，这使得人们假设互动伙伴是诚实的，除非他们有理由相信相反的观点。然而，值得一提的是，最近的一项研究对这一可靠的结果提出了挑战，发现指导参与者仅依赖最好的可用线索，例如故事的细节，使他们能够始终如一地区分谎言和事实，准确度范围为59%至79%。这一发现引发了关于（1）法官在做出真实性判断之前应结合适当数量的线索的争论——并建议使用最佳启发式方法是最直接和准确的——因此关于（2）该提示的诊断水平。

最近，言语测谎问题也通过采用计算技术（例如文体测量法）得到了解决。文体计量学是指来自计算语言学和人工智能的一组方法和工具，可以对书面文本中的语言特征进行定量分析，以发现可以推断和表征作者身份或其他文体属性的独特模式。尽管存在一些局限性，但文体测量法已被证明在测谎方面是有效的。主要优点是可以独立于人类判断来编码和提取言语线索，从而减少编码器间一致性问题，因为研究人员对相同数据使用相同技术将提取相同索引。

1 卢卡 IMT 高级研究学院分子思维实验室，Piazza San Francesco 19, 55100 Lucca, LU, Italy。帕多瓦大学数学系“Tullio Levi-Civita”，意大利帕多瓦。帕多瓦大学普通心理学系，意大利帕多瓦。电子邮件：riccardo.loconte@imtlucca.it

沿着这一趋势，最近的几项研究探索了不同领域的语言计算分析，例如假新闻、法院案件的转录、欺骗性产品评论的评估、网络犯罪的调查、自传信息的分析以及有关欺骗性意图的评估。未来的事件。总的来说，大多数研究都集中在使用机器学习和深度学习算法与自然语言处理 (NLP) 技术相结合来自动检测言语线索中的欺骗（参见 Constâncio 等人对计算机化技术的系统回顾）测谎研究）。

最近，随着大型语言模型 (LLMs) 的出现，人工智能和自然语言处理领域向前迈出了一大步。 LLMs 是基于 Transformer 的语言模型，具有在大量语料库上训练的数亿个参数（即预训练阶段）。由于这个预训练阶段，LLMs 已被证明能够捕获语言的复杂模式和结构，并对语法、语义和语用学产生强大的理解，能够生成类似于人类自然语言的连贯文本。此外，一旦经过预训练，这些模型就可以使用较小的特定任务数据集针对特定任务进行微调。微调是指在新数据集上继续训练预训练模型的过程，使其能够将之前学到的知识适应新数据的细微差别和特殊性，从而获得最先进的结果。 LLMs 微调的常见任务包括 NLP 任务，例如语言翻译、文本分类（例如情感分析）、问答、文本摘要和代码生成。因此，LLMs 擅长处理各种 NLP 任务，而不是专门针对一项特定任务进行训练的模型。然而，据我们所知，尽管 LLMs 具有极大的灵活性，但在小型语料库上微调 LLM 以执行测谎任务的过程仍未被探索。

## 心理学领域的相关著作

在以往旨在识别言语欺骗的可靠线索的心理框架中，距离框架、认知负荷（CL）理论、现实监控（RM）框架和可验证性方法（VA）已被广泛研究，并获得了实证支持。功效不仅来自初步研究，还来自荟萃分析研究。

欺骗的疏远框架指出，说谎者倾向于将自己与自己的叙述保持距离，作为一种处理说谎时所经历的负面情绪的机制，通过使用更少的自我参照（例如“我”，“我”）并采用更多的其他参照（例如，“他”、“他们”）。

CL 框架指出，说谎者在编造虚假回答、检查其与其他编造信息的一致性以及在审查员面前保持可信度和一致性时会消耗更多的认知资源，从而导致更短、更不复杂、更不复杂的陈述。一项荟萃分析发现，基于 CL 理论的方法在检测欺骗方面比标准方法具有更高的准确率。

RM 框架的假设基于记忆特征文献，假设真实的回忆基于经历过的事件，而欺骗性的回忆基于想象的事件。因此，RM 从感觉、空间和时间信息以及事件期间经历的情绪和感受中得出对真实叙述的预测。相反，关于欺骗的预测是根据认知操作（例如思想和推理）的数量得出的。 RM 总分在真实性检测准确性方面似乎具有诊断意义（d = 0.55）（另请参阅对口头测谎方法的广泛回顾）。最近，RM 框架通过语言的具体性进行了研究。在这项研究中，一个基本且得到部分支持的假设是真实具体性假设，该假设表明真实的陈述通常由具体、具体和上下文相关的细节组成。相反，欺骗性或虚假陈述通常包含更抽象和不太具体的信息，与认知操作的 RM 标准更相关。

口头测谎中的 VA 表明，真实的陈述比虚假或欺骗性的陈述更容易被验证，因为说谎者会避免提及可以用独立证据验证的细节来掩盖他们的欺骗行为。可验证的细节可以通过涉及或由已识别个人见证的活动来表示，通过视频或照片证据记录，或留下数字或物理痕迹（例如电话或收据）。

值得注意的是，这些框架提供了可检测的语言线索，可以使用 NLP 技术轻松识别这些线索，并且在这个意义上已经得到了广泛的研究。

胡什等人。对基于计算机的测谎研究进行了荟萃分析，其中大多数纳入的研究依赖于语言查询和字数统计软件（LIWC）。 LIWC 是研究词汇多样性和文本语义内容的黄金标准工具。给定一段文本，LIWC 会计算字典中与不同心理社会维度相关的 100 多个类别对应的总单词的百分比，这些百分比已由人类评估人员使用严格的程序进行验证。在 Houch 的荟萃分析结果中，反映欺骗的距离、CL 和 RM 框架的 LIWC 指标得到了结果的支持，并且可以通过计算机技术检测言语欺骗。

通常，对于距离指标，研究人员通过将第一人称代词与第二人称代词和第三人称代词的频率相加来计算自我和其他参考的数量。当在文本中应用 CL 理论时，研究人员通常会使用和分析有关文本的单词和句子数量、可读性和复杂性的统计数据。 RM 通常与 LIWC 一起研究。舒特等人。提供的证据表明，人类对感知和上下文细节的编码在区分谎言和事实方面并不具有决定性的优越性，从而凸显了自动化技术的潜在优势。此外，最近的研究通过使用命名实体识别（NER）提取了可验证的细节，事实证明这是一种有效的自动化程序，可以检测酒店评论中的欺骗行为以及参与者周末计划的意图。

将 NLP 技术应用于心理学研究的有希望的结果表明，可以在一种新的基于理论的文体分析中结合不同心理框架的指标，从而为一次性从多个角度研究言语测谎提供了可能性。

## AI领域相关工作

人工智能领域之前的工作已将机器学习和深度学习模型应用于数据驱动的言语欺骗检测的二元分类任务中。

Kleinberg 和 Verschuere 开发了一个未来意图数据库，以研究机器和人类判断的结合是否可以提高预测欺骗的准确性。虽然发现人类判断会损害自动欺骗检测的准确性，但作者实现了两种机器学习模型（即普通随机森林），分别根据 LIWC 和词性特征（例如名称、形容词、副词、动词的频率）进行训练，达到准确度分别为 69%（95% CI：63-74%）和 64（95% CI：58%、69%）。在同一数据集上，Ilias 等人评估了六种深度学习模型，包括 BERT（和 RoBERTa）、MultiHead Attention、co-attention 和 Transformers 模型的组合。使用具有共同注意力模型的 BERT 达到的最佳准确率为 70.61% (± 2.58%)。作者还提供了可解释性分析，以了解模型如何结合使用 LIME（一种工具，通过显示文本中的哪些特定单词影响结果，以更直接和更容易理解的方式解释深度学习预测）和 LIWC 来做出决策。

Capuozzo 等人开发了一个新的跨领域和跨语言的意见数据集，要求英语和意大利语参与者就五个不同主题提供真实或欺骗性的意见。使用 FastText 字嵌入对文本进行编码后，他们使用 10 倍交叉验证在多个场景中训练 Transformers 模型，平均准确度范围从“主题内”场景的 63% (± 8.7%) 到高达 90.1 “基于作者”场景中的 % (± 0.16%)。

相比之下，Sap 等人开发了一个由记忆和想象生成的新叙事数据集，并使用 LLM (GPT-3) 来计算称为“顺序性”的新指标。顺序性是一种叙事流程的度量，它比较一个句子在有或没有前面的故事背景的情况下的概率。虽然通过创新的计算方法提供了对讲故事的认知过程的见解，但作者没有采用 LLM 的微调程序来对不同的叙述进行分类。

人工智能领域的研究结果表明，随着模型复杂性的增加，预测文本欺骗的准确性也会提高。然而，准确性的提高通常是以牺牲这些预测的可解释性为代价的。 LLMs 是目前能够处理大量和复杂的语言数据的最前沿模型之一，并且缺乏关于对测谎任务进行微调 LLMs 的文献，这提供了有价值的研究调查该领域的理由。

## 研究目的和假设

本研究的主要目标和假设概述如下：

- 假设 1a)：微调 LLM 可以有效地对原始文本中的简短叙述的真实性进行分类，1b) 在言语测谎方面优于经典机器学习和深度学习方法。
- 假设 2)：微调 LLM 可以有效地对原始文本中的简短叙述的真实性进行分类，1b) 在言语测谎方面优于经典机器学习和深度学习方法。
- 假设 3)：微调 LLM 可以有效地对原始文本中的简短叙述的真实性进行分类，1b) 在言语测谎方面优于经典机器学习和深度学习方法。
- 假设 4)：微调 LLM 可以有效地对原始文本中的简短叙述的真实性进行分类，1b) 在言语谎言检测方面优于经典机器学习和深度学习方法。
- 假设 5a)：区分真实陈述和欺骗性陈述的语言风格因上下文而异，5b)，并且可能是模型预测的一个重要特征。

为了测试假设 1a，我们使用三个数据集对开源 LLM FLAN-T5 进行了微调：个人观点（欺骗性意见数据集）、自传经历（Hippocorpus 数据集）和未来意图（意图数据集）。鉴于 LLMs 的极大灵活性，假设该方法可以检测高于机会水平的原始文本中的欺骗行为。为了测试我们的方法相对于经典机器和深度学习模型（假设 1b）的优势，我们决定将结果与两个基准进行比较，这在方法和材料部分中进行了进一步描述。

关于假设2和假设3，根据经验证据，经典机器学习模型在上述场景下进行训练和测试时往往会出现性能下降的情况。相比之下，LLMs在预训练阶段就获得了对语言模式的全面理解。我们假设经过微调的 LLM 能够在不同的上下文中推广其学习。与假设 4 相关，我们相信这种泛化能力在较大的模型中会进一步增强，因为它们的大小与更复杂的语言表示相关。

最后，为了检验假设 5，我们引入了一种新的基于理论的文体测量方法，名为 DeCLaRatiVE 文体测量，以提取与距离、认知负荷、现实监控和可验证性方法等心理框架相关的语言特征，为从话语中提取特征。我们将应用 DeCLaRatiVE 风格测量法来比较上述三个数据集中的真实和欺骗性陈述，以探索语言风格方面的潜在差异。我们的假设表明，区分真实和欺骗性陈述的语言风格可能在三个数据集中有所不同，因为这些类型的陈述源自不同的上下文。我们还应用了 DeCLaRatiVE 风格测量技术来提供最佳表现模型的可解释性分析。

## 方法和材料数据集

本研究使用了三个数据集：欺骗性意见数据集、从现在起意见数据集、Hippocorpus 数据集、从现在起记忆数据集和意图数据集。对于每个数据集，参与者需要在三个不同领域提供真实或捏造的陈述：对五个不同主题的个人观点（意见数据集）、自传经历（记忆数据集）和未来意图（意图数据集）。值得注意的是，每个领域内的具体主题在说谎者和说真话者之间是平衡的。每个数据集的更详细描述可在补充信息以及每篇原始文章的方法部分中找到。

表 1 显示了关于观点、记忆和意图的真实和欺骗性陈述的示例。
表 2 报告了每个数据集的描述性统计数据，包括总体数据和按真实和欺骗性陈述组分组的情况。这些统计数据包括字数的最小值、最大值、平均值和标准差。使用 spaCy（一个用于文本处理的 Python 库）在文本标记化后计算字数。此外，表 2 提供了真实词汇集和欺骗词汇集之间的 Jaccard 相似度指数值。 Jaccard指数是通过计算这些词的交集（常用词）和并集（总词数）得出的。

| | 真实 | 欺骗性的 |
|---|---|---|
| 观点<br>（流产） | 虽然我在这个问题上感到道德上的撕裂，但我相信归根结底这是一个女人的身体，她应该能够随心所欲地处理它。我相信人们不应该为了让自己感觉更好而对胎儿进行非人性对待。有关此问题的法律决定应由各州决定。为了解决这个问题，避孕措施应该很容易实现 | 堕胎是生命的终结，不应该被允许。如果胎儿已经能够在母体之外“独立”生存，我们有什么权利缩短它的生命呢？如果母亲有生命危险，她在同意生育的时候就已经选择了愿意牺牲自己的生命来生孩子。 |
| 回忆（我和男朋友一起去听了一场音乐会，玩得很开心。我们在那里遇到了一些朋友，看日落真的很开心。） | 这一天的开始很完美，我们开车前往丹佛观看演出。我和男朋友在去红岩的路上没有遇到任何交通堵塞，而且天气很好。我们在演出时与我的朋友们在剧院顶部附近会面，并铺上了毯子。揭幕战开始了，我们随着舞台上演奏的班卓琴和曼陀林起舞。我们很高兴能在那里。那是日落开始的时候。真是太美了。天空是柔和的粉红色，看起来很美丽。就在菲尔·莱什上场的时候，我差点就死了。这是我一生中最幸福的时刻，在近十年没有见到他之后再次见到他。我很高兴能和我的朋友和我的爱人在一起。那天晚上没有什么可以超越的。我们开车回家，看到满天繁星，在一个观景台停下来仰望它们。我喜欢我居住的这个地方。我喜欢现场音乐。我很高兴 | 音乐会是我最喜欢的事情，我男朋友也知道这一点。这就是为什么在我们的周年纪念日，他给我买了去看我最喜欢的艺术家的门票。不仅如此，门票还是户外演出的门票，我更喜欢户外演出，而不是在拥挤的体育场里。因为他知道我是音乐的忠实粉丝，所以他为自己买了票，甚至还为我的几个朋友买了票。他对我和我喜欢做的事情非常友善和体贴。我将永远记住这件事，我将永远珍惜他。演唱会那天，我做好了准备，他来接我，我们提前去了一家餐厅。他是如此的浪漫。他不用问就知道要带我去哪里。我们一边吃一边笑，在大型活动之前进行了一次愉快的晚餐约会。我们到达了音乐会现场，音乐非常美妙。我喜欢它的每一分钟。我的朋友、男朋友和我都坐在一起。当音乐慢慢消失时，我发现我们都迷失了方向，只是盯着星星。这是一个令人难以置信且美丽的夜晚 |
| 意向（和女儿去游泳） | 我们每周都会去水宝宝班，我 16 个月大的孩子正在那里学习游泳。我们在水中做了很多活动，例如学习吹泡泡、使用浮标帮助游泳、戏水以及学习如何在落水时自救。我发现这项活动很重要，因为我喜欢与女儿共度时光和游泳是一项重要的生活技能 | 这个星期六我将带我 8 岁的女儿去游泳。我们会一大早就去，因为那个时候通常会安静得多，而且我女儿总是很早起床看动画片（凌晨 5 点！）。在她九月份开始新学校之前，我正在努力教她如何在深水区游泳，因为他们每周有两次游泳课 |

表 1. 关于观点、记忆和意图的真实和欺骗性示例陈述。括号中是分配给参与者在欺骗条件下编造叙述的主题。

| 数据集（总数） | 最小–最大字数 | 平均字数 (SD) | 杰卡德相似指数（定性解释） |
|---|---|---|---|
| 所有意见 (2500) | 6–338 | 59.05 (30.66) | 0.35<br>（相似度低） |
| 真实意见 (1250) | 7–338 | 66.74 (31.95) | |
| 欺骗性意见 (1250) | 6–232 | 51.36 (27.24) | |
| 所有的意图 (1640) | 15–251 | 50.44 (30.11) | 0.34<br>（相似度低） |
| 真诚的意图 (783) | 15–206 | 47.04 (28.36) | |
| 欺骗意图 (857) | 15–251 | 53.55 (31.31) | |
| 所有的回忆 (5506) | 22–625 | 255.24 (92.36) | 0.34<br>（相似度低） |
| 真实的回忆 (2770) | 22–625 | 269.78 (94.14) | |
| 欺骗性的记忆 (2736) | 22–609 | 240.51 (88.12) | |

表 2. 每个数据集以及真实和欺骗性陈述集的单词数的汇总统计。杰卡德相似度指数及其括号中的定性解释是指每个数据集的真实词汇集和欺骗词汇集之间的相似度。

两套。得到的索引范围为0到1，0表示两个集合之间的词汇完全不同，1表示两个集合之间的词汇完全相同。我们报告了杰卡德相似度指数，以提供相应数据集中真实和欺骗性陈述的单词选择之间的相似性或重叠度。补充信息提供了计算杰卡德相似指数的详细方法。

## 法兰-T5

我们采用了 FLAN-T5，这是由 Google 研究人员开发的 LLM，可通过 HuggingFace Python 的 Transformers 库免费获取 (https://huggingface.co/docs/transformers/model_doc/flan-t5)。 HugginFace 是一家通过 Python API 提供对最先进的 LLMs 的免费访问的公司。在可用的 LLMs 中，我们选择了 FLAN-T5，因为它在计算负载和学习表示的优良性之间进行了有价值的权衡。 FLAN-T5 是 MT-5 的改进版本，MT-5 是一种文本到文本的通用模型，能够解决许多 NLP 任务（例如情感分析、问答和机器翻译），并通过预训练进行了改进。该模型的独特之处在于，他们接受的每项训练任务都被转换为文本到文本的任务。例如，在执行情感分析时，输出预测是训练集中用于标记每个短语的正面或负面情感的字符串，而不是二进制整数输出（例如，0 = 正面；1 = 负面）。因此，它们的力量在于预训练阶段学习的自然语言的广义表示，以及在不调整其架构的情况下轻松调整模型以适应下游任务的可能性。

## 陈述式风格分析

本研究采用风格测量分析来实现两个主要目标。首先，我们的目的是在初始化微调过程之前描述区分三个数据集的语言特征。其次，我们进行了可解释性分析，以深入了解语言风格在模型分类过程中区分真实和欺骗性陈述的作用。为此，采用了我们称为 DeCLaRatiVE 风格测量的新框架，该框架涉及提取 26 个语言特征，并结合距离、认知负荷、现实监控和可验证性方法等心理框架。表 3 显示了 26 种语言特征的完整列表及其简短描述。这种综合方法能够从多维角度分析欺骗的言语线索。

与 CL 框架相关的特征包括有关文本长度、可读性和复杂性的统计信息，并使用 Python 库 TEXTSTAT 进行提取。与距离相关的功能

| 标签 | 描述 |
|---|---|
| 句子数 | 句子总数 |
| 字数 | 总字数 |
| 音节数 | 总音节数 |
| 每个单词的平均音节数 | 每个单词的平均音节数 |
| fk_等级 | 理解课文所需年级的索引 |
| fk_read | 文本可读性指数 |
| 分析型 | LIWC 摘要统计根据分析思维分析文本风格（0-100） |
| 真正的 | LIWC 摘要统计分析文本风格的真实性 (0–100) |
| Tone | ‘tone_pos’ — ‘tone_neg’ 的标准化差（0-100） |
| 音调位置 | 与积极情绪相关的单词百分比（LIWC 词典） |
| 音调否定 | 与负面情绪相关的单词百分比（LIWC 词典） |
| 认识 | 与认知过程语义域相关的单词百分比（LIWC 词典） |
| 记忆 | 与记忆/遗忘语义领域相关的单词百分比（LIWC 词典） |
| 焦点过去 | 与过去相关的动词和副词的百分比（LIWC 词典） |
| 焦点呈现 | 与现在时相关的动词和副词的百分比（LIWC 词典） |
| 焦点未来 | 与将来相关的动词和副词的百分比（LIWC 词典） |
| 自参考 | LIWC 类别“i”+“we”的总和 |
| 其他参考 | LIWC 类别“她和”+“他们”+“你”的总和 |
| 感性细节 | LIWC 类别“注意力”+“视觉”+“听觉”+“感觉”的总和 |
| 上下文嵌入 | LIWC 类别“空间”+“运动”+“时间”的总和 |
| 现实监控 | 感知细节+情境嵌入+情感之和——认知 |
| 具体性得分 | 单词具体性得分平均值 |
| 人们 | 与人相关的独特命名实体：例如“玛丽”、“保罗”、“亚当” |
| 时间细节 | 与时间相关的唯一命名实体：例如“星期一”、“下午 2:30”、“圣诞节” |
| 空间细节 | 与空间相关的独特命名实体：例如“机场”、“东京”、“中央公园” |
| 数量详情 | 与数量相关的唯一命名实体：例如“20%”、“5 $”、“first”、“ten”、“100 m” |

表 3. 与 DeCLaRatiVE Stylometry 技术相关的 26 个语言特征的列表和简短描述。

和 RM 框架是使用 LIWC（用于分析单词用法的黄金标准软件）计算的。我们使用英语词典对每个文本以及 LIWC-22 中存在的所有类别进行评分。 LIWC 评分是使用英语词典对标记化文本进行计算的。与距离和 RM 框架相关的 LIWC 类别的选择是以先前关于计算机化言语测谎的研究和最近的荟萃分析为指导的。 RM 还通过单词的语言具体性进行了研究。为了确定每个陈述的平均具体性水平，我们利用了 Brysbaert 等人开发的具体性注释数据集。为了计算具体性分数，使用 Python 库 SpaCy 对文本数据应用了预处理管道：文本被转换为小写并标记化；然后删除停用词，并对剩余的内容词进行词形还原。然后将这些内容词与带注释的具体性数据集交叉引用，以在找到匹配时分配相应的具体性值。然后，将每个陈述的具体性得分计算为该陈述中所有内容词的具体性得分的平均值。对于涉及可验证细节的内容，它们是通过唯一命名实体的频率来估计的。使用Python的SpaCy库，通过英语语言的Transformer算法，使用NER技术提取命名实体（en_core_web_trf，https://spacy.io/models/en#en_ core_web_trf）。

补充信息中提供了有关如何计算 26 种语言特征的更多详细信息。

## 实验设置

在本节中，我们将描述我们在这项工作中应用的方法。第一步，我们想要对数据集进行描述性语言分析，试图对假设 5a) 做出回应，即区分真实和欺骗性陈述的语言风格是否在不同的上下文中有所不同。为了实现这一结果，我们采用了 DeCLaRatiVE 风格测量分析。第二步，我们继续测试 FLAN-T5 模型在谎言检测任务上进行微调的能力。为此，我们提供了三种场景来验证以下假设：

- 假设 1a)：微调 LLM 可以有效地对原始文本中的简短叙述的真实性进行分类，1b) 在言语测谎方面优于经典机器学习和深度学习方法。
- 假设 2)：对欺骗性叙述进行微调 LLM 使模型也能够检测新类型的欺骗；
- 假设 3)：对欺骗性叙述进行微调 LLM 使模型也能够检测新类型的欺骗；
- 假设4）：模型性能取决于模型大小，模型越大，精度越高；

我们期望假设 1a、1b、3 和 4 得到验证，而我们对第二个假设没有任何先验期望。场景描述如下：

1. 场景 1：模型在单个数据集上进行了微调和测试。每次使用同一模型的不同副本（即微调过程之前的相同参数）对每个数据集重复此过程（图 1）。该场景评估模型学习如何检测与相同背景相关的谎言的能力，并对假设 1a 做出回应；
2. 场景 2：模型在三个数据集中的两个上进行了微调，并在剩余的未见过的数据集上进行了测试。对于之前的场景，该过程迭代了 3 次，使用同一模型的单独实例，每次都使用不同的数据集配对组合（图 2）。该场景评估模型在训练阶段从未接触过的新环境中的样本上的表现，并为假设 2 提供响应；
3. 场景 3：我们首先聚合场景 1 中的三个训练集和测试集。然后，我们在聚合数据集上微调模型，并在聚合测试集上测试模型（图 1）。该场景评估了模型从多种背景下的真实和欺骗性叙述样本中学习和概括的能力，并为假设 3 提供了答案。

在场景 1 和 3 中，每个实验都经过 10 倍交叉验证。 N 折交叉验证是一种统计方法，用于通过将数据集划分为 n 个分区（本研究中 n = 10）来估计模型的性能。对于每个分区 i，我们创建一个由剩余 n−1 个分区组成的训练集，使用第 i 个分区作为测试集（即 90% 的数据属于训练集，剩余的 10% 数据属于训练集）测试集）。对于每次迭代，都会在测试集上计算、存储性能指标，然后求平均值。此过程可确保公正的性能估计，并允许在不同模型之间进行公平比较。在我们的研究中，我们在场景 1 和 3 中以及两种模型大小中采用了相同的训练测试分割，以保证公平的性能比较。每次折叠的平均测试精度及其相应的标准偏差以性能指标的形式呈现。相反，在场景 2 中，使用整个两个配对数据集作为训练集对每个配对组合进行微调，同时使用完整的未见过的数据集作为测试集来评估模型的性能。

值得注意的是，意见数据集的开发包含了每个参与者的真实和欺骗性陈述，总共有五种意见。因此，我们将每个意见视为一个单独的样本。为了避免模型因学习参与者的语言风格而在测试集上表现出夸大的性能，我们采取了以下预防措施。具体来说，我们确保了参与者之间的独家划分

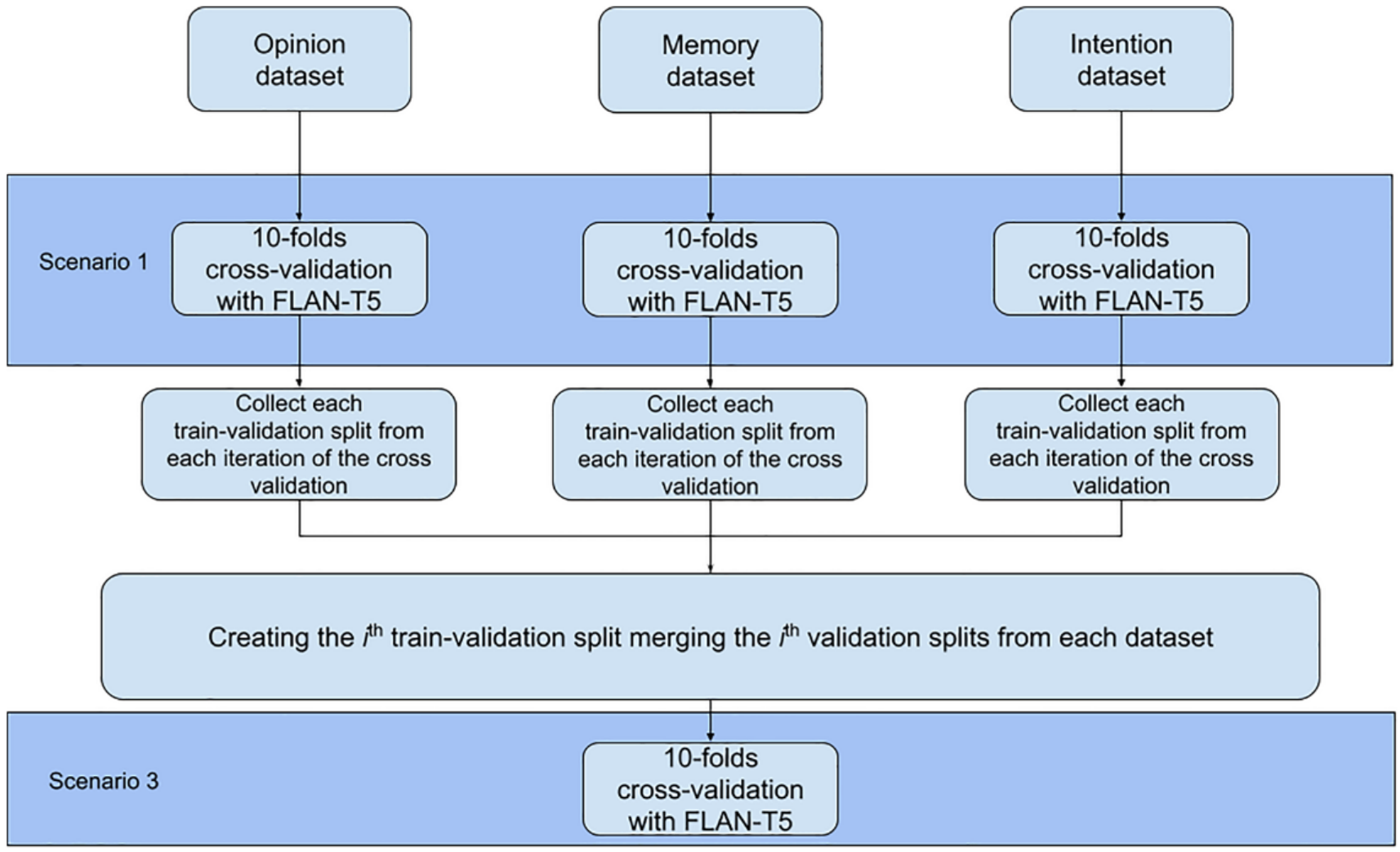


图 1. 场景 1 和 3 的直观图示。

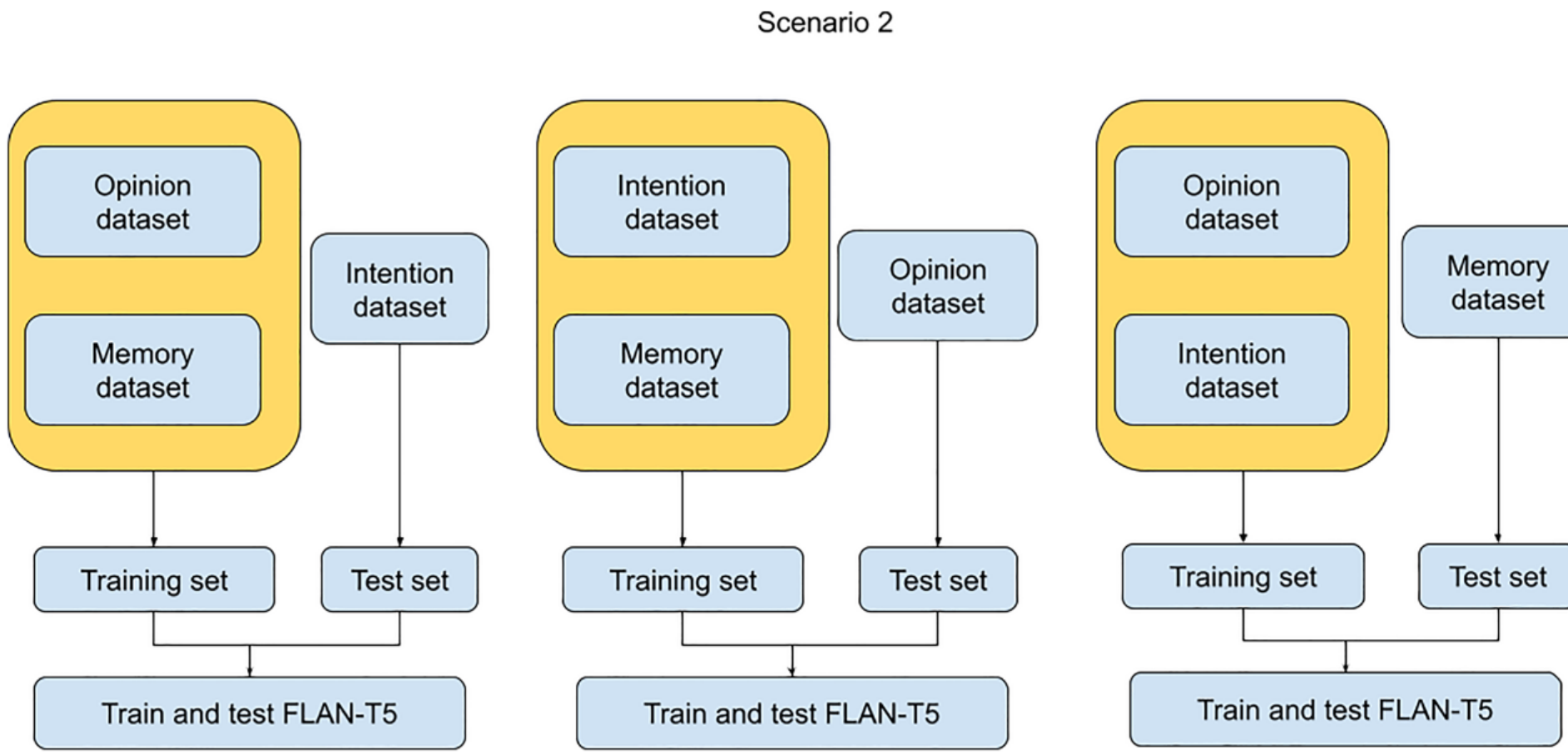


图 2. 场景 2 的直观图示。

训练集和测试集，这样任何将其意见分配给训练集的个人都不会将其意见分配给测试集，反之亦然。
场景 2 和场景 3 共同提供了关于在未见过的数据和多域数据集上进行测试时，经过微调的 FLAN-T5 模型在测谎任务中的通用能力的证据。此外，我们测试了模型性能是否取决于模型大小。因此，我们首先在每个场景中对小尺寸版本的 FLAN-T5 进行微调，然后在每个场景中使用基本尺寸版本重复相同的实验，为假设 4 提供了答案。
为了测试假设 1b，即测试我们的方法与经典机器学习模型相比的优势，我们决定将结果与两个基准进行比较：

1. 由词袋（BoW）编码器加逻辑回归分类器组成的基本方法（遵循场景1的实验过程）；

2. 基于先前研究的文献基线，使用机器学习或深度学习方法提供相同数据集的准确性指标。对于意见数据集（以每个主题对五个不同主题的意见为特征），我们将我们的结果与他们的“主题内”实验获得的性能进行比较，因为我们的方法与他们的方法相同，唯一的区别是我们解决了所有问题一个模型中的主题。

最后一步，我们进行了可解释性分析，以调查模型正确分类和错误分类的真实陈述和欺骗性陈述之间的语言风格差异。该过程旨在回答假设 5b，即模型是否考虑其最终预测的陈述的语言风格。为了实现这一结果，我们采用了 DeCLaRatiVE 风格测量分析。

在图 3 中，我们提供了整个实验设置的流程图。

## 微调策略

LLMs 的微调包括通过在特定于任务的数据上进一步训练模型，使预先训练的语言模型适应特定任务，从而增强其生成符合上下文的上下文相关且连贯的文本的能力。期望的任务目标。我们使用三个数据集并遵循上述实验设置，对 FLAN-T5 的小尺寸和基本尺寸进行了微调。考虑到三个数据集包含与二进制标签相关的原始文本，特别是分类为真实或欺骗性的实例，我们将测谎任务视为二元分类问题。

据我们所知，文献中没有针对这种新颖的下游 NLP 任务的微调策略。因此，我们的策略遵循 Hugginface 关于微调 LLM 翻译的指南。具体来说，我们选择了与预训练原始模型相同的优化策略和相同的损失函数。

值得注意的是，在 FLAN-T5 预训练阶段从未执行过欺骗性陈述和真实陈述之间的分类任务，也没有包含在模型预训练的任何任务中。因此，我们使用不同的学习率值（即 1e−3、1e−4、1e−5）多次执行实验设置部分中描述的相同实验，最终选择表 4 中所示的配置，这在准确性方面产生了最佳性能。三个场景的所有实验和运行都是在 Google Colaboratory Pro + 上使用 NVIDIA A100 Tensor Core GPU 进行的。

## 描述性语言分析的统计程序

应用 DeCLaRatiVE 文体测量技术后，我们为三个数据集的每个文本获得了包含 26 个语言特征的文体向量。

为了评估观察到的组间差异的显着性，采用了排列 t 检验。这种非参数方法涉及汇集所有观察结果，然后将它们随机重新分配为两组，同时保留原始组大小。然后计算这些排列组的感兴趣的检验统计量（即平均值的差异）。通过重复此过程数千次（即 n = 10,000），我们在组间没有差异的零假设下生成了检验统计分布。然后，将实际数据中观察到的检验统计量与该分布进行比较，以计算 p 值，表明如果原假设为真，则观察到这种差异的可能性。使用排列 t 检验的优点是不需要对数据分布进行假设。该分析是使用 SciPy 和 Pinouin 库在 Python 中进行的。

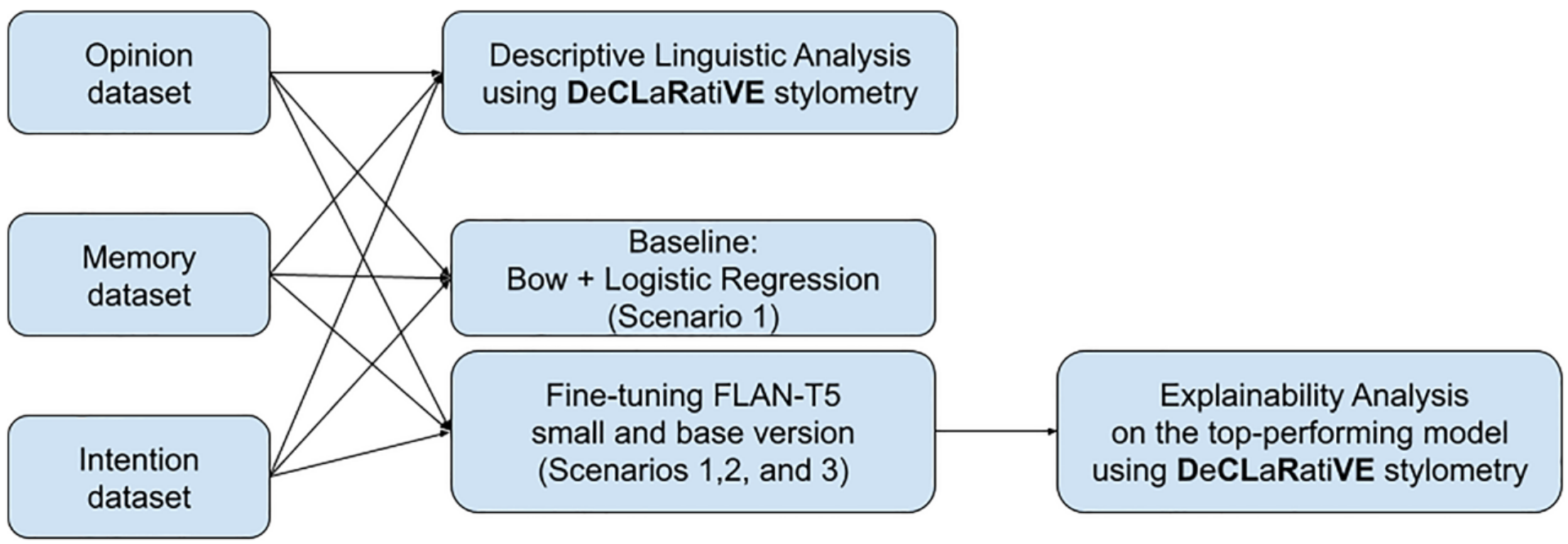


图 3.整个实验装置的直观图示。使用 DeCLaRatiVE 文体测量法对意见、记忆和意图数据集进行了描述性语言分析。还为这三个数据集建立了由词袋（BoW）和逻辑回归（场景1）组成的基线模型。然后，小版本和基本版本的 FLAN-T5 模型在场景 1、2 和 3 中进行了微调。最后，使用 DeCLaRatiVE 风格测量对表现最佳的模型进行了可解释性分析，以解释结果。

| 模型 | 超参数 | 价值 |
|---|---|---|
| FLAN-T5 小号 | 学习率 | 5e−4 |
| | 权重衰减系数 | 0.01 |
| | 批量大小 | 2 |
| | 历元数 | 3 |
| FLAN-T5底座 | 学习率 | 5e−5 |
| | 权重衰减系数 | 0.01 |
| | 批量大小 | 2 |
| | 历元数 | 3 |

表 4. 小型和基本尺寸版本的 FLAN-T5 超参数配置。每个场景的初始学习率为小模型的 5e−4 和基本模型的 5e−5。这一选择是由初步实验结果推动的，较小的模型（但不是基础模型）通常在较高的学习率下表现更好。所有模型和场景中的权重衰减系数均设置为 0.01。出于计算原因，批量大小设置为 2，特别是为了避免耗尽可用内存，尽管已知较大的批量大小通常会带来更好的性能。最后，经过初步实验，在没有过拟合的情况下，第三个epoch后的测试精度达到最大，将epoch数设置为3。

对于记忆和意图数据集，我们对 26 个语言特征的独立样本计算了排列 t 检验 (n = 10,000)，以概述真实文本和欺骗性文本之间的显着差异。

对于意见数据集，我们的分析过程如下。首先，我们计算了所有受试者意见的 DeCLaRatiVE 风格测量技术。这产生了 2500（意见）× 26（语言特征）矩阵。然后，由于每个受试者提供了五个意见（一半是真实的，一半是欺骗性的），我们分别对真实和欺骗性意见集的风格向量进行平均。这个过程使我们能够获得同一主题的两种不同的平均风格向量，一种代表真实的意见，一种代表欺骗性的意见。重要的是，这种平均过程使我们能够获得独立于主题（例如堕胎或大麻合法化）和主题所采取的立场（例如支持或反对该特定主题）的结果。最后，我们通过配对样本排列检验（n = 10,000）验证了这些差异的统计显着性。每个数据集的结果均经过 Holm-Bonferroni 校正的多重比较进行校正。

效应大小由置信区间为 95% (95% CI) 的通用语言效应大小 (CLES) 表示，这是效应大小的度量，通过提供特定效果的概率，可以更直观地理解。随机挑选的真实陈述中的语言特征将比随机挑选的欺骗性陈述中的得分更高。CLES 的空值是机会水平为 0.5（概率范围为 0 到 1），表示在采样时，一组将大于另一组，且机会相等。还计算了 95% CI 的 Cohen d 效应大小以添加解释。

## 可解释性分析的统计程序

为了检查输入语句的语言风格是否对模型的结果输出产生影响，并为错误的分类输出提供解释，我们对由表现最好的模型正确分类和错误分类的语句进行了 DeCLaRatiVE 风格分析。场景 3（FLAN-T5 基础）。

为此，在交叉验证的每次迭代中，我们将属于测试集的句子及其实际标签与模型预测的标签进行配对。交叉验证结束后，对于 10 个折叠中的每一个以及构成该折叠测试集的句子的 26 个语言特征中的每一个，我们对独立样本（n = 10,000 ）进行以下兴趣比较：

一个。 真实的陈述被错误地分类为欺骗性的（假阴性），欺骗性的陈述被错误地分类为真实的（假阳性）； b. 正确分类为欺骗性的陈述（真阴性）与被错误分类为欺骗性的真实陈述（假阴性）； c. 正确分类为真实的陈述（真阳性）与被错误分类为真实的欺骗性陈述（假阳性）。 d. 模型正确分类的真实与欺骗性陈述（真阳性与真阴性）。

为了计算效应大小，我们计算了 CLES 和 Cohen d 效应大小分数的平均值以及从每次折叠获得的各自的 95% CI。

结果描述性语言分析本节概述了 DeCLaRatiVE 风格计量分析的描述性语言分析结果，以比较三个数据集的语言特征。

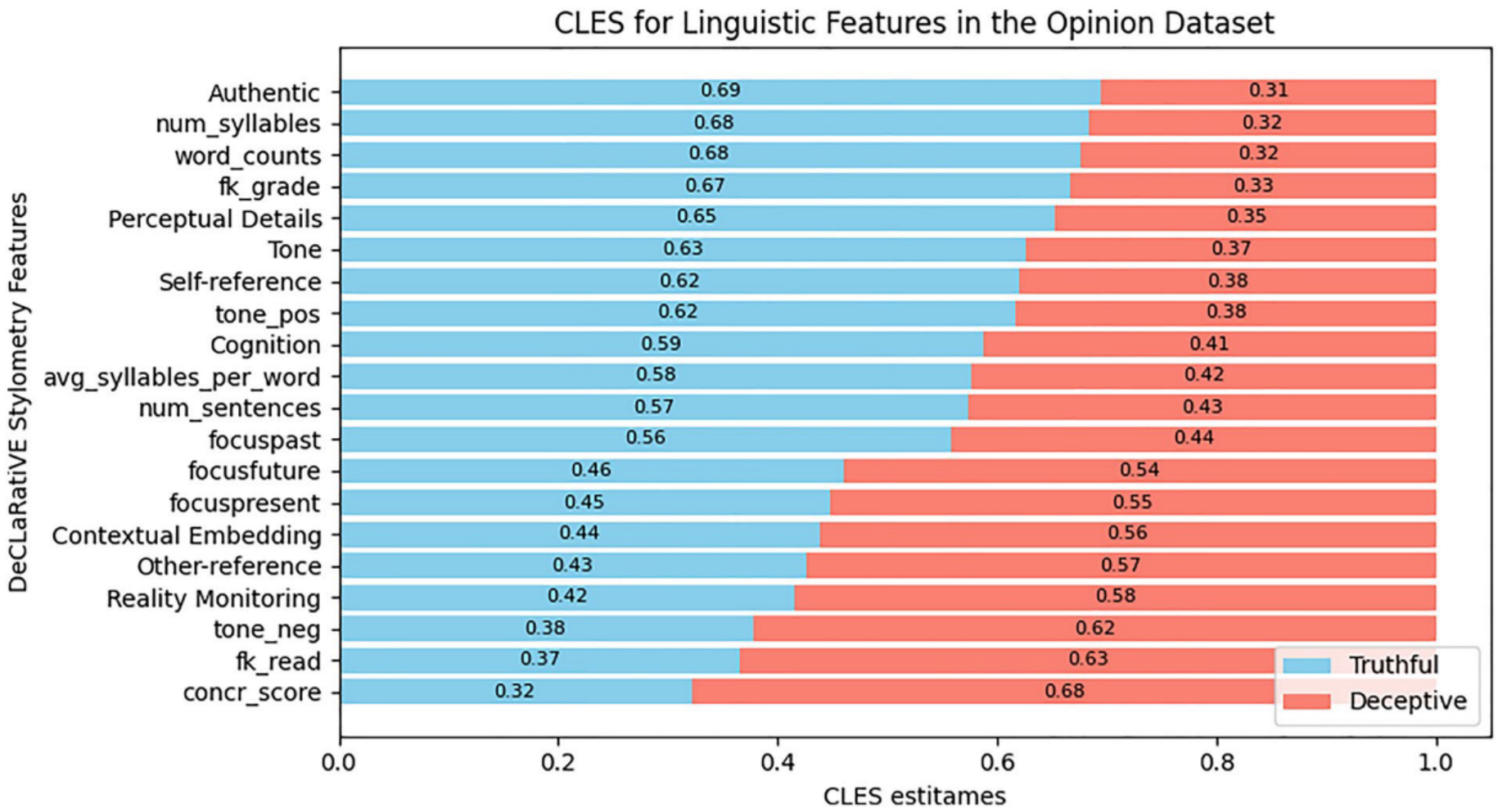


图 4. 水平堆叠条形图显示了意见数据集中经过事后更正的重要语言特征的公共语言效应大小 (CLES) 估计。 CLES 估计代表在真实意见（天蓝色）中找到特定语言特征的概率（范围从 0 到 1），而不是在欺骗性意见（鲑鱼色）中找到特定语言特征的概率（范围从 0 到 1）。真实意见的 CLES 按降序排序，而欺骗性意见的 CLES 按升序排序。

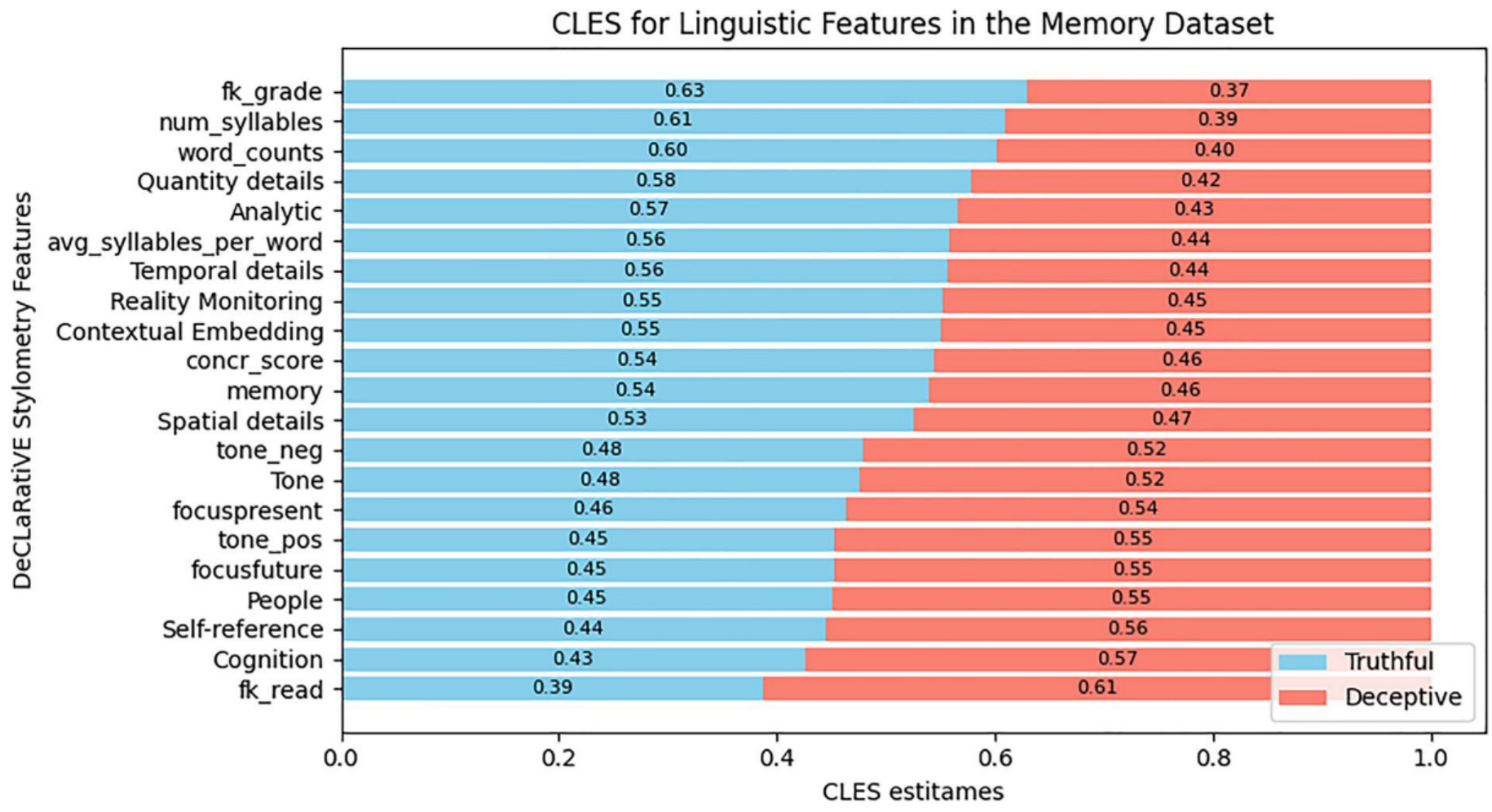


图 5. 水平堆叠条形图显示了记忆数据集中事后更正后幸存的重要语言特征的公共语言效应大小 (CLES) 估计值。 CLES 估计表示在真实记忆（天蓝色）中找到特定语言特征的概率（范围从 0 到 1），而不是在欺骗性记忆（鲑鱼色）中找到特定语言特征的概率（范围从 0 到 1）。真实记忆的 CLES 按降序排序，而欺骗性记忆的 CLES 按升序排序。

对于三个数据集，图图 4、5 和 6 显示了事后修正后幸存的语言特征在数量、类型、CLES 效应大小的大小以及效应方向方面的差异。为了举例说明这些差异，单词的具体性得分（“concr_score”）在意图数据集中针对真实陈述呈现出最大的 CLES（图 6），而在意见数据集中，它针对欺骗性陈述呈现出最大的 CLES声明（图4）。总体而言，与意见和记忆数据集相比，意图数据集在真实和欺骗性陈述之间的语言特征上显示出较少的显着差异。在表 S5（补充信息）中，我们报告了所有语言特征和三个数据集、所有统计数据、校正后的 p 值、CLES 和 Cohen's D 表示的效应量分数（置信区间为 95%）以及方向的效果。

## 测谎分类任务的表现

本节以 10 倍的平均准确度（和标准差）的形式介绍所有场景中小模型和基本模型的最后一个时期之后测试集上的性能。

场景1
表 6 描述了 FLAN-T5 模型的测试精度，按场景 1 中的数据集和模型大小分类。在每种情况下，基本模型平均优于小模型，其中内存数据集显示出最大的改进4%，而意图数据集显示平均准确度仅增加了 0.06%。这些结果表明，较大的模型尺寸通常会提高三个数据集的性能，并且在基础版本中观察到更高的准确性。

场景2
此场景旨在研究我们经过微调的 LLM 在不同欺骗领域的泛化能力。如表 5 所示，这种情况下三个实验的测试准确性显着下降到机会水平，这表明该模型在任何情况下都能够学习检测来自不同上下文的谎言的一般规则。

场景3
在场景 3 中，我们在聚合的意见、记忆和意图数据集上测试了 FLAN-T5 小版本和基础版本的准确性。小尺寸FLAN-T5的平均测试精度为75.45%（st.dev.±1.6），而基本尺寸FLAN-T5的平均测试精度更高，为79.31%（st.dev.±1.3）。
换句话说，基本尺寸模型的性能比小模型高出大约四个百分点。
表 6 中的结果显示了场景 3 中的小型 FLAN-T5 模型和基本 FLAN-T5 模型在各个数据集上的分类性能，并与场景 1 中的对应模型进行了比较。这些比较表明，场景 3 中的 FLAN-T5-small 表现出比场景 3 中的 FLAN-T5-small 更差的性能。在场景 1 中。相反，在场景 3 中，

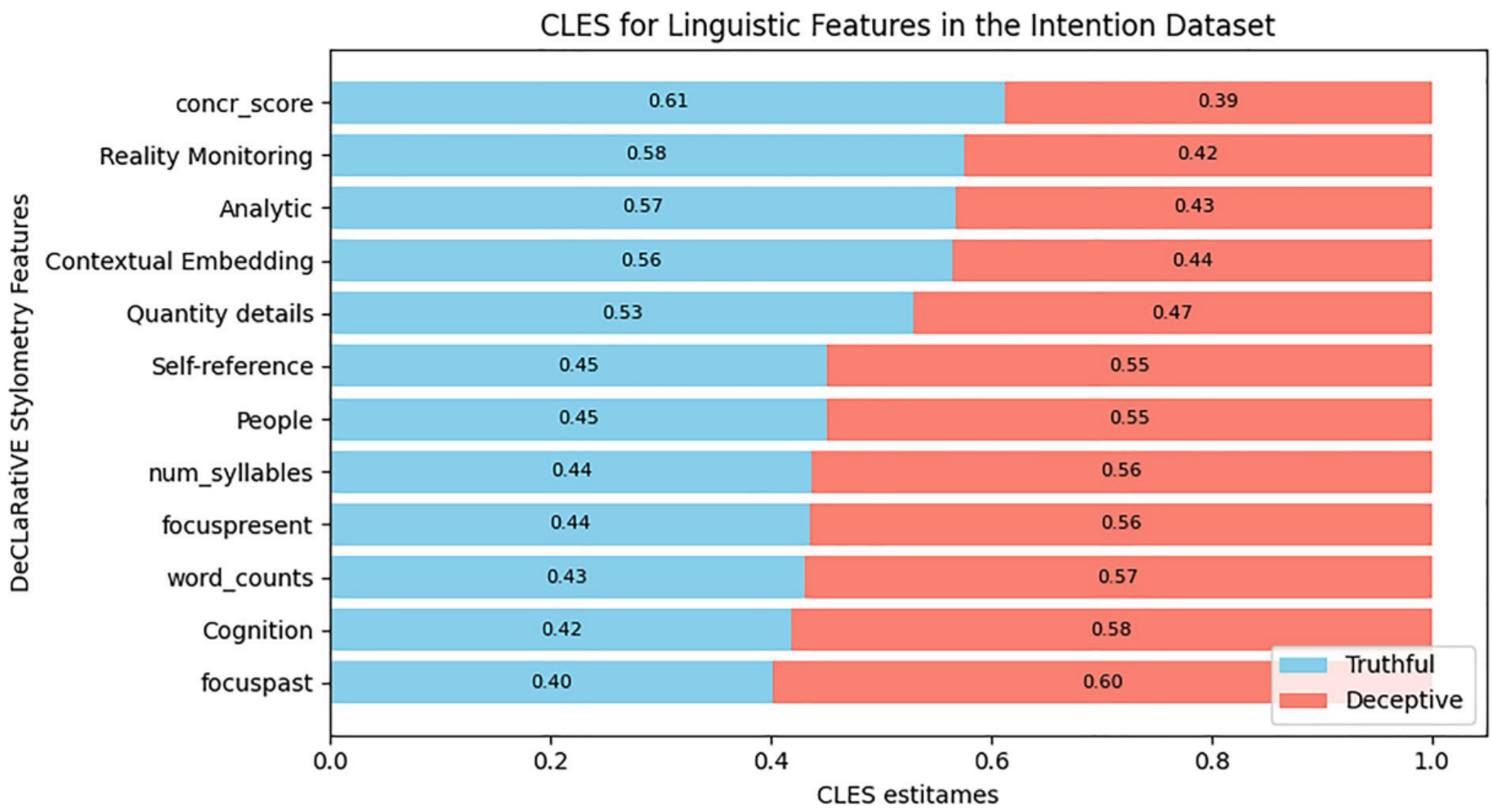


图 6. 水平堆叠条形图显示了对意向数据集中经过事后更正后仍保留的重要语言特征的公共语言效应大小 (CLES) 估计。 CLES 估计表示在真实意图（天蓝色）中找到特定语言特征的概率（范围从 0 到 1），而不是在欺骗性意图（鲑鱼色）中找到特定语言特征的概率（范围从 0 到 1）。真实意图的 CLES 按降序排序，而欺骗意图的 CLES 按升序排序。

| 火车组 | 测试集 | 型号尺寸 | 测试精度 |
|---|---|---|---|
| 意见+记忆 | 意图 | FLAN-T5 小号 | 55.37 |
| | | FLAN-T5底座 | 55.67 |
| 意见+意图 | 记忆 | FLAN-T5 小号 | 55.37 |
| | | FLAN-T5底座 | 54.23 |
| 记忆+意图 | 观点 | FLAN-T5 小号 | 53.12 |
| | | FLAN-T5底座 | 49.40 |

表 5. 场景 2 中 FLAN-5 模型的测试精度（训练集的三种组合）。性能比较是FLAN-T5模型的小型版本和基础版本在训练集的三种组合中的表现：意见+记忆、意见+意图、记忆+意图。

| 模型 | 观点 | 记忆 | 意图 |
|---|---|---|---|
| 词袋基线 | 76.16±2.9% | 57.57±7.66% | 67.07±3.18% |
| 文献基线 | 65.16±5.7% | – | 69.00 [63; 74]%<br>69.86±2.34%<br>70.61±2.58% |
| FLAN-T5 小型 — 场景 1 | 80.64±2.03% | 76.87±2.06% | 71.46±3.65% |
| FLAN-T5 底座—场景 1 | 82.60±3.01% | 80.61±1.41% | 71.52±2.21% |
| FLAN-T5 小型 — 场景 3 | 79±2.11% | 75.67±1.90% | 69.32±3.75% |
| FLAN-T5 底座—场景 3 | 82.72±2.39% | 79.87±1.60% | 72.25±2.86% |

表 6. 测试场景 1 和 3 中三个数据集的 FLAN-T5 模型的准确性。报告值是 10 倍的平均值±标准差。每个评估指标的最佳结果以粗体显示。 Opinion 数据集的文献基线是指来自 FastText Embedding + Transformer 的所有主题内准确度的平均准确度和标准差。 Intention 数据集的文献基线是指使用 LIWC 特征的 Vanilla Random Forest 的准确率（方括号中的置信区间）、RoBERTa + Transformers + Co-Attention 模型和 BERT + co-attention 模型的平均准确度和标准差。

基本模型在意见和意图数据集上的表现仅比场景 1 的对应模型高出不到 1%，并且在内存数据集上的表现略低于场景 1 的对应模型。

我们在场景 3 中将表现最好的模型确定为 FLAN-T5 基础，因为它的整体性能精度更高。该模型 10 次折叠的平均混淆矩阵如图 7 所示。

值得注意的是，无论如何，我们都能够超越词袋+逻辑回归分类器基线以及之前研究中在相同数据集上实现的性能。

## 可解释性分析

本节旨在通过对模型正确分类和错误分类的语句进行 DeCLaRatiVE 风格计量分析，更深入地了解场景 3（FLANT5 基础）中确定的表现最佳的模型。此分析的目的是检查输入语句的语言风格是否对模型的最终输出产生影响，并为错误的分类输出提供解释。为了进行此分析，我们比较了：

一个。 真实的陈述被错误地分类为欺骗性的（假阴性），欺骗性的陈述被错误地分类为真实的（假阳性）； b. 正确分类为欺骗性的陈述（真阴性）与被错误分类为欺骗性的真实陈述（假阴性）； c. 正确分类为真实的陈述（真阳性）与被错误分类为真实的欺骗性陈述（假阳性）。 d. 模型正确分类的真实与欺骗性陈述（真阳性与真阴性）。

报告的统计显着特征在每次折叠中都经过了多重比较的事后校正。总体而言，对于 a)、b) 和 c) 的比较，我们观察到大多数分割的任何语言特征都没有统计显着差异 ($p < 0.05$)，唯一的例外是：

1) 第 1 折叠中的“fk_read”（t = 5.30；p = 0.04，CLES = 0.63 [0.55, 0.71]，d = 0.46 [0.18, 0.75]）和第 6 折叠中的“现实监控”（t = 4.74；p = 0.047， a) 比较的 CLES = 0.62 [0.54, 0.70], d = 0.46 [0.17, 0.75]）；
2) 第 6 折中的“现实监测”（t = −3.39，p = 0.04，CLES = 0.40 [0.34，0.46]，d = −0.34 [−0.55，−0.13]）和“现实监测”（t = −3.16 p = 0.04，CLES = 0.41 [0.34，0.47]，d = −0.34 [−0.56，−0.12]）和“上下文嵌入”（t = −2.11；p = 0.01，CLES = 0.39 [0.33，0.45]，d = − 0.42 [−0.63, −0.2]) 在第 7 倍 b) 比较中；

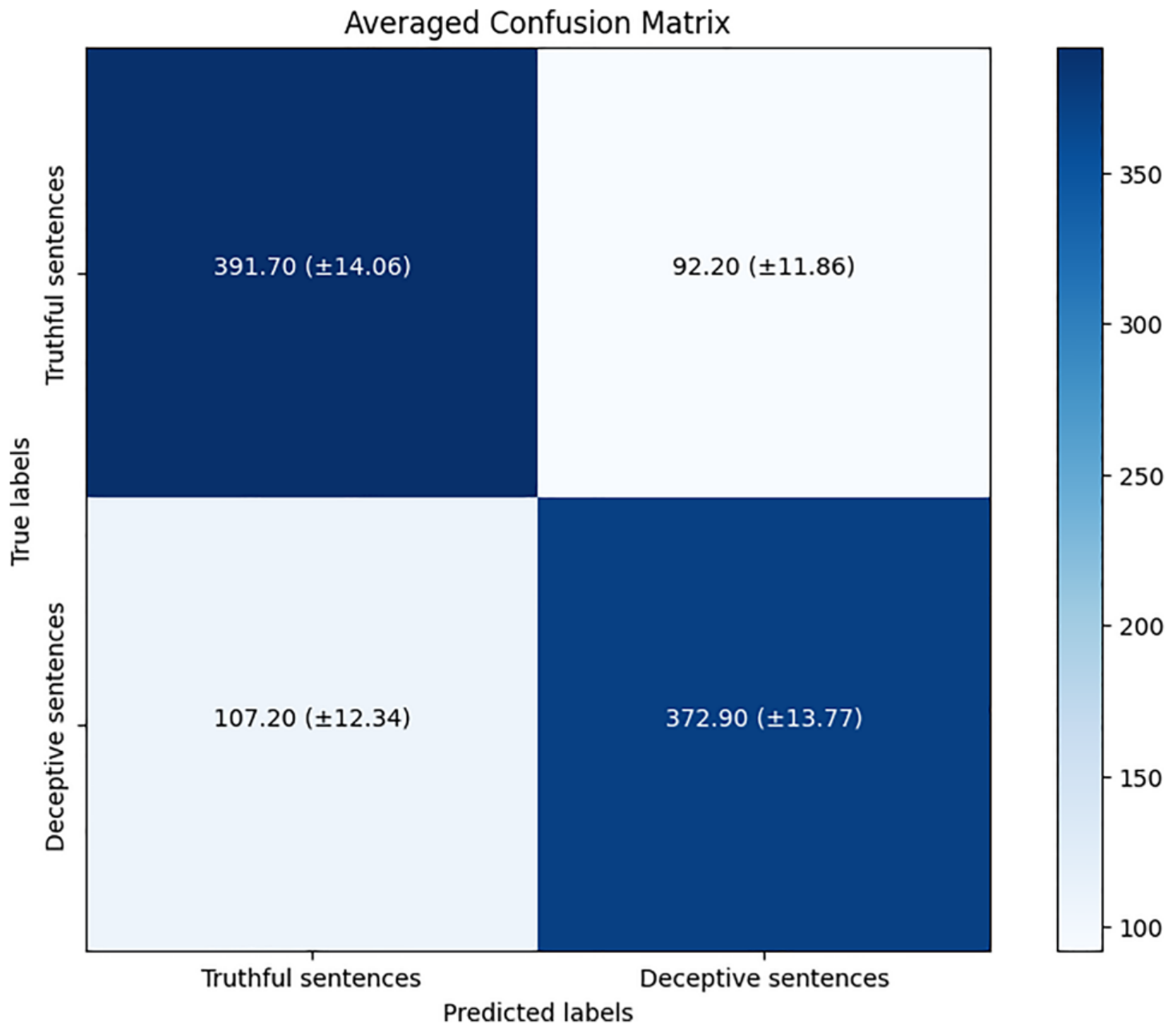


图 7. 场景 3 中被识别为 FLAN-T5 基础的最佳性能模型的平均混淆矩阵。在每个方格中，获得的结果代表 10 倍交叉每次迭代的测试集的平均值（和标准差） -验证。

3) “num_syllables”（t = 76.87，p = 0.01，CLES = 0.64 [0.57，0.7]，d = 0.46 [0.27，0.7]）和“word_counts”（t = 59.63，p = 0.01，CLES = 0.64 [0.57，0.71] ], d = 0.46 [0.21, 0.7]) 在第 9 折中用于 c) 比较。

相反，对于 d) 比较，所有折叠中都出现了几个显着特征，并且在多重比较的校正中幸存下来。图 8 描述了语言特征的 CLES 效应大小分数，根据它们在十个折叠中被发现显着的次数进行排序。图 8 中的前六个特征代表了与认知负荷框架相关的一组语言特征。

## 讨论

在本研究中，我们研究了大型语言模型（特别是小型基础版本的 FLAN-T5）在学习和泛化不同上下文中欺骗的内在语言表示方面的功效。为了实现这一目标，我们使用了三个数据集，其中包含有关个人观点、自传经历和未来意图的真实或捏造的陈述。

### 描述性语言分析

通过探索 DeCLaRatiVE 风格的差异，进行描述性语言分析来比较三个数据集的语言特征，即分析从距离、认知负荷、现实监控和可验证性方法的心理框架中提取的 26 个语言特征。该分析旨在检验假设 5a，该假设假设语言风格存在差异，可以在不同的背景下区分真实的陈述和欺骗性的陈述（即个人观点与自传体记忆与未来意图）。这项分析的结果证实了我们的假设，表明真实和欺骗性陈述之间表现出统计显着差异的语言特征确实在不同数据集中有所不同。这种变化是根据特征的总数和类型、效应大小的大小（从非常小到中等）以及效应的方向来观察的。在以下段落中，将讨论每个数据集的重要语言特征的解释。

## 意见

在使用 DeCLaRatiVE 文体测量法分析真实和欺骗性的观点后，发现与 CL、RM 和 Distancing 的理论框架相关的不同语言特征非常重要。

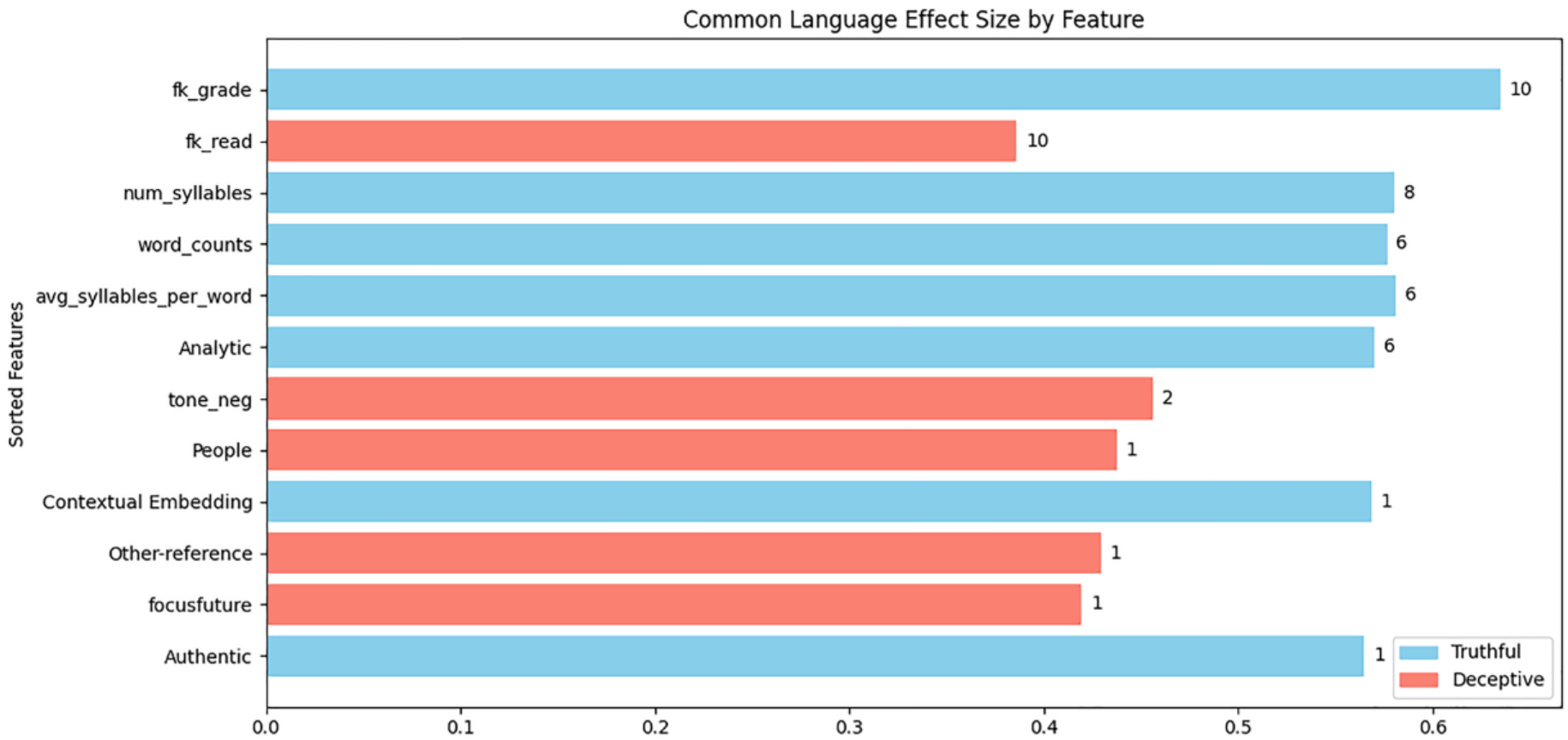


图 8. 场景 3 中由 FLAN-T5 库准确分类的真实和欺骗性陈述中的语言特征。条形图显示了经过事后修正的十倍语言特征中的平均公共语言效应大小。语言特征根据在 10 个折叠中被发现显着的次数按降序排列（显示在每个条形的侧面）。真实文本中平均较高的语言特征以天蓝色显示，而欺骗性文本中平均较高的语言特征以鲑鱼色显示。

根据 CL 框架，我们观察到真实意见的特点是更复杂、更冗长、语言风格更真实。

对于与 RM 框架相关的特征，如前所述，真实意见的特点是具体词数量较少，认知词数量较多；相比之下，欺骗性意见在言语的具体性、上下文细节和现实监控方面得分较高。这些差异一方面反映了说真话的人在评估抽象和有争议的概念（例如堕胎）的利弊时所进行的推理过程，而对于欺骗者来说，这可能表明他们在抽象方面存在困难，从而导致了虚假的观点听起来更贴近现实。

最后，与之前关于疏远框架和欺骗性意见的文献一致，欺骗者使用更多的其他相关词类（“其他参考”）和更少的自我相关词（“自我参考”），证实个人可能倾向于避免表达欺骗性言论时的个人参与。

## 回忆

通过陈述式文体测量法对自传体记忆的真实和欺骗性叙述进行分析后，发现与 CL、RM、VA 和距离理论框架相关的各种语言特征是重要的。

至于观点，根据CL框架，自传体记忆的真实叙述表现出更高程度的复杂性和冗长性，并且在风格上似乎更具分析性。

根据 RM 框架，假设真实的记忆账户倾向于反映经历事件时所涉及的感知过程，而捏造的账户是通过认知操作构建的，我们发现真实的记忆在与记忆相关的单词和相关单词的数量上表现出更高的分数具有空间和时间信息（“上下文嵌入”），以及总体较高的 RM 分数。相反，我们发现欺骗性记忆在与认知过程相关的单词（例如推理、洞察力、因果关系）中表现出较高的分数。此外，根据克莱因伯格的真实具体假设

39、真实记忆的总体特征是具体性得分较高的词语。

与 VA 一样，真实的记忆包含了更多可验证的细节，正如更多关于时间和地点的命名实体所表明的那样。值得注意的是，尽管参与者在低风险场景中撒了谎，但我们还是发现了这种效应。然而，欺骗性记忆出人意料地以大量的自我参照和“人”的命名实体为特征。这一结果与之前关于距离框架的文献形成了鲜明对比。对于这种显着但微小的影响的一种可能解释是，说谎者可能会试图通过培养社会联系感来提高自己的可信度。

## 意图

通过DeCLaRatiVE文体测量法检查未来意图的真实和欺骗性陈述后，发现一些重要的语言特征。我们的发现与之前的研究一致，即真正的意图包含更多的“如何表达”，即仔细计划的指标和对活动的具体描述。相反，虚假意图的特点是“为什么话语”，即解释和理由

为什么有人计划一项活动或以某种方式做某事。事实上，我们发现真实意图更有可能提供有关预期行动的具体而独特的信息，将其陈述扎根于现实世界的经验并提供时间和空间参考。此外，真实意图的特点是更具分析性和更多的数字实体。相比之下，错误意图表现出更多数量的认知词语和表达，并且在时间上面向现在和过去。

此外，我们发现了与以下说法相符的证据：说谎者可能会过度准备他们的陈述，如更冗长的内容所表明的那样。最后，与疏远框架相比，我们发现欺骗性陈述中自我提及和提及他人的比例明显更高。然而，这一发现的影响很小。至于欺骗性记忆，一种可能的解释是，说谎者可能试图通过创造一种社会联系感来显得更可信。

## 测谎任务

为了测试 FLAN-T5 模型在测谎任务上微调的能力，我们开发了三个场景。

在场景 1 中，我们测试了微调 LLMs 是否可以有效地对基于原始文本的简短陈述的准确性进行分类，其性能远高于机会水平（假设 1a）。为此，我们对 FLAN-T5 的小版本进行了微调，以执行谎言检测作为分类任务。我们对三个数据集（即意见、记忆、意图）重复了这个过程。这个微调过程产生了有希望的结果，证实了我们的假设，意见数据集的平均准确度为 80.64%（标准偏差 ± 2.03%），内存数据集的平均准确度为 76.87%（标准偏差 ± 2.06%），意图数据集为 71.46%（标准偏差 ± 3.65%）。

在场景 2 中，我们测试了对欺骗性叙述进行微调 LLM 是否能让模型检测新型欺骗（假设 2）。为了验证这一假设，我们在两个数据集上对 FLAN-T5（小版本）进行了微调，并在第三个数据集上进行了测试（例如，训练：意见 + 记忆；测试：意图）。我们的研究结果表明，该模型在该场景的所有三种组合中都在机会水平上执行，这表明该模型没有可以学习区分真实性和欺骗性陈述的通用规则，从而能够在不同的上下文中泛化任务。事实上，如描述性语言分析部分所示，三个数据集在真实和欺骗性叙述的内容和语言风格方面存在显着差异。因此，该模型努力识别语言欺骗的特定模式，并且似乎进行了特定领域的学习，根据特定的欺骗领域调整其分类能力。

在场景 3 中，我们测试了在多上下文数据集上微调 LLM 是否能让模型在多上下文测试集上获得成功的预测（假设 3）。为此，我们使用三个聚合数据集（即意见+记忆+意图）对FLAN-T5（小版本）进行了微调和测试。小型 FLAN-T5 的平均准确度为 75.45% (st. dev. ± 1.6)。此外，与场景 1 中的对应数据集相比，各个数据集的分类性能仅表现出准确性的小幅下降（约 1%）。这些发现证实了我们的假设，提供了 LLMs 在多上下文数据集上进行微调和发送文本时进行泛化能力的证据，而之前的经验证据表明机器学习模型在多上下文数据集上的性能下降同样的场景。

为了测试使用更大模型时模型性能是否会提高（假设 4），我们使用 FLAN-T5 的基础版本在场景 1、2 和 3 中重复了相同的实验。

在场景1中，我们发现FLAN-T5的基本版本比小版本提供了更高的精度。
在场景 3 中，模型的基础版本实现了 79.31%（st.dev.±1.3）的平均准确度，比小模型高出约四个百分点。此外，与场景 1 中较小模型或 FLAN-T5 基础所实现的效果相比，总体准确性的提高并没有影响任何单个数据集的性能。相比之下，场景 2 中 FLAN-T5 的基础版本仍然获得了机会水平附近的表现。

一方面，从场景 1 和 3 中的基本模型获得的结果证实了模型大小确实影响性能的假设，可能是因为更大的模型能够更好地表示真实和欺骗性叙述的语言模式。具体来说，在场景 3 中，FLAN-T5 库规模较大，具有理解和整合三个不同数据集特征的能力，从而在所有单个数据集上保持一致的性能。相比之下，场景 3 中较小的 FLAN-T5 似乎放弃了某些专门能力，这些能力有利于特定数据集对不同上下文中的欺骗进行分类。

另一方面，场景 2 和 3（使用小型且基础的 FLAN-T5）的结果表明，尽管 LLMs 已经对语言模式有了全面的了解，但仍然需要接触先前的示例来准确分类欺骗性内容。不同领域内的文本。

最后，为了测试我们的方法在言语测谎（假设 1b）方面是否优于经典机器学习和深度学习方法，我们将 FLAN-T5 的小型版本和基本版本获得的结果与逻辑回归器的更简单基线的性能进行了比较基于 BoW 嵌入和先前在意见和意图数据集文献中使用的 Transformer 模型。

具体来说，将 Memory 数据集与逻辑回归基线进行比较时，性能提高了 32%。这种改进可能归因于记忆数据集中的故事更长、更复杂，这对谎言检测任务中基于 BoW 的逻辑回归等更直接的方法的有效性提出了挑战。相比之下，LLMs已经拥有强大的语言表示；因此，微调 LLMs 利用这种表示，专门针对测谎任务调整他们的 NLP 熟练程度，从而产生更高的准确性。

对于意见和意图数据集，通过微调 LLMs 获得的性能不太明显。
对于意见数据集，这可能是由于这些数据集中的分类相对容易，其中更简单

模型已经可以取得良好的性能，留下的改进余地较小。尽管如此，我们的方法与基线之间的差异不可忽视。在 Opinion 数据集中，我们比从头训练的 Transformer 模型的文献基线准确率高出 17%，并且比我们的逻辑回归基线高出 6 个百分点。对于意图数据集，我们的方法比逻辑回归基线提高了 5 个百分点，比最佳文献基线提高了约 1-2%。值得注意的是，意图数据集的最佳文献基线（平均准确度：70.61 ± 2.58%）在使用的模型类型方面与我们的方法类似，涉及基于 Transformer 的模型（BERT + Coattention），这可以解释缩小绩效差距。

除了性能上的差异之外，与以前的研究中使用的方法相比，我们的方法的主要优点是它的简单性和灵活性。微调 LLM 利用现有的语言编码，可以轻松处理任何类型的语句，这与基于 BoW 的逻辑回归或从头开始训练新的基于 Transformer 的模型不同。综合考虑所有这些方面，微调 LLMs 在可行性、灵活性和性能准确性方面更具优势。

## 可解释性分析

为了提高所收集性能的可解释性，我们研究了以真实和欺骗性叙述为特征的语言风格是否会在模型的最终预测中发挥作用（假设 5b）。为此，我们对场景 3 中确定的表现最好的模型（即 FLAN-T5 基础）正确分类和错误分类的语句应用了 DeCLaRatiVE 风格计量分析。

在错误分类的样本中，对于使用 DeCLaRatiVE 文体测量技术提取的任何语言特征，真实和欺骗性的陈述没有显着差异。唯一的例外是第 1 折叠，它显示出文本可读性分数的显着差异，以及第 6 折叠，它显示出“现实监控”分数的显着差异。在被正确分类为欺骗性（真阴性）的欺骗性陈述和被错误分类为欺骗性（假阴性）的真实陈述之间，每个折叠的语言特征均未检测到显着差异，但第 6 折叠和第 7 折叠中的“现实监控”除外以及第 7 折中的“上下文嵌入”分数。最后，被正确分类为真实的真实陈述（真阳性）和被错误分类为真实的欺骗性陈述（假阳性）没有表现出显着差异，除了音节数和单词数之外。折叠中的单词 9. 我们认为，对特定折叠中选定语言特征的显着差异的观察更表明这些发现可能无法概括，并且可能受到所分析的特定折叠的影响。综合来看，大多数分析的折叠显示出语言风格的大量重叠。因此，该模型可能对这些陈述表现出较差的分类性能，因为它们虽然具有欺骗性，但表现出类似于真实陈述的语言风格，反之亦然。

相比之下，正确分类的陈述显示出真实陈述和欺骗性陈述之间的一些显着差异。值得注意的是，图 8 中的前 6 个语言特征在 10 倍中至少有 6 倍具有统计显着性。我们在正确分类的陈述中发现了一致的语言特征模式，但在错误分类的陈述中没有发现这一事实，这一事实为我们的假设提供了证据，表明陈述的语言风格确实在模型的最终预测中发挥了作用。更详细地说，图 8 中描绘的前 6 个语言特征代表了与 CL 框架相关的一组语言线索，特别是与文本的长度、复杂性和分析风格相关的低级特征，这些特征可能使得区分真实陈述和欺骗性陈述。事实上，CL 的语言线索在反映观点、记忆和意图的话语混合数据集中的几个可用特征中幸存下来，这一事实提出了一个问题：CL 线索是否比其他更具体的线索更具有普遍性。特定类型的欺骗。

## 结论、局限性和进一步的工作

截至撰写本文时，据我们所知，这是第一项涉及使用 LLM 进行测谎任务的研究。

LLMs 是基于 Transformer 的模型，在大型文本语料库上进行训练，已被证明可以生成人类自然语言的连贯文本，并且在各种 NLP 任务中具有极高的灵活性。此外，这些模型可以使用较小的特定任务数据集对特定任务进行进一步微调，从而实现最先进的结果。在本研究中，我们测试了经过微调的 LLM (FLAN-T5) 在测谎任务上的能力。

首先，考虑到 LLM 的极端灵活性，我们测试了微调 LLM 是否是检测高于机会水平的原始文本的欺骗行为并优于经典机器和深度学习的有效程序。接近。我们发现，在单个数据集上微调 FLAN-T5 是获得最先进精度的有效过程，这一过程优于基线模型（BoW + 逻辑回归）和之前的工作这一事实证明了这一点在同一数据集上应用机器和深度学习技术。

其次，我们想研究对欺骗性叙述进行微调 LLM 是否能让模型也检测出新型的欺骗性叙述。场景 2 的结果否定了这一假设，表明该模型需要以前不同欺骗性叙述的示例才能在此分类任务中提供足够的准确性。

第三，我们研究了是否有可能在多上下文数据集上成功微调LLM。
场景 3 的结果证实，经过微调的 LLM 可以在检测不同上下文中的欺骗方面提供足够的准确性。我们还发现，相对于在单个数据集上进行微调，对多个数据集进行微调可以提高性能。

此外，我们假设模型性能可能取决于模型大小，因为模型越大，模型形成其语言的内部表示就越好。场景 1 和 3 的结果证实 FLAN-T5 的基本尺寸模型比小尺寸模型具有更高的精度。

最后，通过我们的实验，我们引入了 DeCLaRatiVE 文体测量技术，这是一种基于理论的新文体测量方法，用于从四个心理框架（距离、认知负荷、现实监控和可验证性方法）研究文本中的欺骗行为。我们采用 DeCLaRatiVE 文体测量技术来比较三个数据集的语言特征，我们发现来自不同上下文的捏造陈述表现出不同的欺骗语言线索。我们还采用 DeCLaRatiVE 文体测量技术来进行可解释性分析，并调查真实或欺骗性叙述的语言风格是否是模型在其最终预测中考虑的特征。为此，我们比较了表现最好的模型（场景 3 中的 FLAN-T5 基础）的正确分类和错误分类的语句，发现正确分类的语句具有与认知负荷理论相关的语言特征。相比之下，真实和欺骗性的错误分类陈述在语言风格上并不存在显着差异。

鉴于所取得的结果，我们强调多样化数据集对于实现普遍良好性能的重要性。我们还认为数据集多样性和 LLM 大小之间的平衡至关重要，这表明数据集越多样化，实现更高级别精度所需的模型就越大。我们的方法的主要优点在于它适用于原始文本，而不需要大量的培训或手工制作的功能。

尽管我们的模型取得了成功，但三个重大限制影响了我们研究结果的生态有效性及其在现实生活场景中的实际应用。

第一个显着的局限性与我们研究的狭窄焦点有关，该研究仅集中在三个特定背景下的测谎：个人观点、自传体记忆和未来意图。这种有限的范围限制了在不同领域内准确分类欺骗性文本的可能性。第二个限制是，我们专门考虑了在旨在收集真实且完全捏造的叙述的实验设置中开发的数据集。然而，人们经常在现实生活场景中使用嵌入的谎言，其中很大一部分叙述是真实的，而不是编造一个完全虚构的故事。最后，本研究中使用的数据集是在实验性低风险场景中收集的，其中参与者撒谎和表现得可信的动机较低。由于上述所有问题，我们的模型在现实生活中的应用可能会受到限制，建议在这种情况下解释结果时务必小心。

本研究解决的局限性强调了未来研究需要扩大测谎模型在现实生活中的适用性和普遍性。未来的工作可能会探索包含新数据集，尝试不同的 LLMs （例如最新的 GPT-4）、不同的大小（例如 FLAN-T5 XXL 版本）和不同的微调策略来研究测谎任务中的性能差异。此外，我们的微调方法完全消除了模型先前拥有的功能；因此，未来的工作还应该侧重于不损害模型原始功能的新的微调策略。

## 数据可用性

对于意见数据集，我们在联系相应作者后获得了完全访问权限。内存数据集可通过以下链接下载：https://msropendata.com/datasets/0a83f6f-a759-4a17-aaa2-fac84577318。意图数据集可通过以下链接公开获取：https://osf.io/45z7e/。

## 代码可用性

用于对三个数据集执行语言分析、在三个场景中微调模型以及进行可解释性分析的所有 Colab Notebooks 均位于 https://github.com/robecoder/VerbalLieDetectionWi thLLM.git。

收稿日期：2023 年 6 月 29 日；接受日期：2023 年 12 月 16 日
Published online: 21 December 2023

## 参考

1. Walczyk, J. J.、Harris, L. L.、Duck, T. K. 和 Mulay, D. 用于理解严重谎言的社会认知框架：激活决策-建构-行动理论。新思想心理学。 34、22-36。 https://doi.org/10.1016/j.newideapsych.2014.03.001（2014）。
2. Amado, B. G.、Arce, R. 和 Fariña, F. Undeutsch 假设和基于标准的内容分析：荟萃分析综述。 《欧洲心理应用杂志》法律背景 7, 3–12。 https://doi.org/10.1016/j.ejpal.2014.11.002（2015）。
3. Vrij，A.等人。言语测谎：过去、现在和未来。脑科学 12, 1644。https://doi.org/10.3390/brainsci12121644 (2022)。
4. Vrij, A. 和 Fisher, R. P. 哪些测谎工具可供刑事司法系统使用？。 J.应用程序。资源。内存。认知。 5、302-307。 https://doi.org/10.1016/j.jarmac.2016.06.014（2016）。
5. 德保罗，B.M. 等人。欺骗的暗示。心理。公牛。 129、74–118。 https://doi.org/10.1037/0033-2909.129.1.74（2003）。
6. Bond, C. F. Jr. & DePaulo, B. M. 欺骗判断的准确性。个人的。苏克。心理。修订版 10，214–234。 https://doi.org/10。1207/s15327957pspr1003_2 (2006)。
7. Levine, T. R.、Park, H. S. 和 McCornack, S. A. 检测真相和谎言的准确性：记录“真实性效应”。交流。莫诺格。 66、125–144。 https://doi.org/10.1080/03637759909376468（1999）。
8. Levine, T. R. 真值默认理论 (TDT)。 J·朗.苏克。心理。 33、378-392。 https://doi.org/10.1177/0261927x14535916（2014）。
9. Street, C. N. H. & Masip, J. 真相偏差的来源：启发式处理？。扫描。 J.心理学。 56、254–263。 https://doi.org/10。1111/sjop.12204（2015）。
10. Verschuer，B.，等人。使用最佳启发式方法有利于欺骗检测。纳特。哼。行为。 7、718–728。 https://doi.org/10。 1038/s41562-023-01556-2 (2023)
11. Chen, X.、Hao, P.、Chandramouli, R. 和 Subbalakshmi, K. P. 从电子邮件中检测作者相似性。在模式识别中的机器学习和数据挖掘国际研讨会上。编辑 P. Perner（纽约，纽约：Springer），375–386。https://doi.org/10.1007/978-3-642-23199-5_28（2011）。
12. Chen, H. 暗网：探索和挖掘网络的黑暗面。 2011 年欧洲情报与安全信息学会议，1-2。 IEEE（2011）。
13. Daelemans, W. 计算文体测量学的解释。计算语言学和智能文本处理，451-462。施普林格，柏林。 https://doi.org/10.1007/978-3-642-37256-8_37（2013）。

14. Hauch, V.、Blandón-Gitlin, I.、Masip, J. 和 Sporer, S. L. 计算机是有效的测谎仪吗？对欺骗的语言线索的元分析。个人的。苏克。心理。启示录 19，307-342。 https://doi.org/10.1177/1088868314556539（2015）。
15. Tomas, F.、Dodier, O. 和 Demarchi, S. 欺骗性语言的计算测量：前景和问题。正面。交流。 7 https://doi.org/10.3389/fcomm.2022.792378（2022）。
16. Conroy, N. K.、Rubin, V. L. 和 Chen, Y. 自动欺骗检测：查找假新闻的方法。过程。副教授。信息。科学。技术。52、1-4。 https://doi.org/10.1002/pra2.2015.145052010082（2015）。
17. Pérez-Rosas, V.、Kleinberg, B.、Lefevre, A. 和 Mihalcea, R. 自动检测假新闻。 arXiv 预印本 arXiv:1708.07104 (2017)。

18. Fornaciari, T. & Poesio, M. 意大利法庭案件中的自动欺骗检测。阿蒂夫。英特尔。法律 21, 303–340。 https://doi.org/10。1007/s10506-013-9140-4 (2013)。
19. Yancheva, M., & Rudzicz, F. 使用句法复杂性特征自动检测儿童生成的语音中的欺骗行为。计算语言学协会第 51 届年会记录 1, 944–953, (2013)。
20. Pérez-Rosas, V. 和 Mihalcea, R. 开放域欺骗检测实验。 2015 年自然语言处理经验方法会议论文集。 https://doi.org/10.18653/v1/d15-1133（2015）。
21. Ott, M.、Choi, Y.、Cardie, C. 和 Hancock, J. T. 通过任何想象力寻找欺骗性意见垃圾邮件。 arXiv 预印本 arXiv：1107.4557 (2011)。

22. Fornaciari, T., & Poesio, M. 将虚假亚马逊评论识别为向人群学习。计算语言学协会欧洲分会第 14 届会议记录。 https://doi.org/10.3115/v1/e14-1030n（2014）。
23. Kleinberg, B.、Mozes, M.、Arntz, A. 和 Verschuere, B. 使用命名实体进行计算机自动言语欺骗检测。法医学杂志 63, 714–723。 https://doi.org/10.1111/1556-4029.13645（2017）。
24. Mbaziira, A. V., & Jones, J. H. 用于母语和非母语英语网络犯罪网络的基于文本的混合欺骗模型。国际计算和数据分析会议论文集。 https://doi.org/10.1145/3093241.3093280（2017）。
25. Levitan, S.I.、Maredia, A. 和 Hirschberg, J. 采访对话中欺骗和感知欺骗的语言线索。计算语言学协会北美分会 2018 年会议记录：人类语言技术，1。https://doi.org/10.18653/v1/n18-1176 (2018)。

26. Kleinberg, B.、Nahari, G.、Arntz, A. 和 Verschuere, B. 通过言语欺骗检测来检测飞行欺骗意图的调查。科拉布拉：心理学。 3. https://doi.org/10.1525/colabra.80（2017）。
27. Constâncio, A. S.、Tsunoda, D. F.、Silva, H. de F. N.、Silveira, J. M. da 和 Carvalho, D. R. 使用机器学习进行欺骗检测：系统评价和统计分析。 PLOS ONE，18，e0281323。 https://doi.org/10.1371/journal.pone.0281323（2023）。
28. 赵，W.X.，等。大型语言模型的调查。 arXiv 预印本 arXiv：2303.18223。 （2023）。
29. Newman, M. L.、Pennebaker, J. W.、Berry, D. S. 和 Richards, J. M. 说谎话：从语言风格预测欺骗。个人的。苏克。心理。公牛。 29、665–675。 https://doi.org/10.1177/0146167203029005010（2003）。
30. 莫纳罗，M.等人。使用键盘动态进行隐蔽测谎。科学报告 8，1976 年。https://doi.org/10.1038/s41598-018-20462-6 (2018)。

31. Vrij, A.、Fisher, R. P. 和 Blank, H. 测谎的认知方法：荟萃分析。法律犯罪。心理。 22（1），1-21。https://doi.org/10.1111/lcrp.12088（2015）。
32. Johnson, M.K. & Raye, C.L. 现实监测。心理。修订版 88, 67–85。 https://doi.org/10.1037/0033-295x.88.1.67（1981）。
33. Sporer, S. L. 通往真相的鲜为人知的道路：捏造和自我经历的事件中的欺骗检测中的言语线索。应用。认知。心理。 11（5），373-397。 https://doi.org/10.1002/(SICI)1099-0720(199710)11:5%3c373::AID-ACP461%3e3.0。CO；2-0 (1997)。
34. Sporer, S. L.《法医环境中欺骗的检测》中的现实监测和欺骗检测（剑桥大学出版社），64–102。 https://doi.org/10.1017/cbo9780511490071.004（2004）。
35. Masip, J.、Sporer, S. L.、Garrido, E. 和 Herrero, C. 用现实监测方法检测欺骗：对经验证据的回顾。心理。犯罪法 11(1), 99–122。https://doi.org/10.1080/10683160410001726356（2005）。
36. Amado, B. G.、Arce, R.、Fariña, F. 和 Vilariño, M. 基于标准的内容分析 (CBCA) 成人现实标准：荟萃分析综述。国际。 J.克林。健康心理学。 16（2），201-210。 https://doi.org/10.1016/j.ijchp.2016.01.002（2016）。
37. Gancedo, Y.、Fariña, F.、Seijo, D.、Vilariño, M. 和 Arce, R. 现实监测：法医实践的荟萃分析综述。欧元。J.心理学。应用。法律背景 13(2), 99–110。 https://doi.org/10.5093/ejpalc2021a10（2021）。
38. Vrij，A.等人。言语测谎：它的过去、现在和未来。脑科学。 12(12), 1644。https://doi.org/10.3390/brainsci12121644 (2022)。

39. Kleinberg, B.、van der Vegt, I. 和 Arntz, A. 通过语言具体性检测欺骗性沟通。开放科学中心。 https://doi.org/10.31234/osf.io/p3qjh（2019）。
40. Nahari, G.、Vrij, A. 和 Fisher, R. P. 通过检查细节的可验证性来利用说谎者的言语策略。法律犯罪。心理。19、227-239。 https://doi.org/10.1111/j.2044-8333.2012.02069.x（2012）。
41. Vrij, A. 和 Nahari, G. 可验证性方法。《基于证据的调查性访谈》（第 116-133 页）。劳特利奇。 https://doi.org/10.4324/9781315160276-7（2019）。
42. Pennebaker, J. W.、Francis, M. E. 和 Booth, R. J. 语言查询和字数统计：LIWC 2001。Mahway：Lawrence Erlbaum Associates，71，2001 (2001)。
43. Boyd, R. L.、Ashokkumar, A.、Seraj, S. 和 Pennebaker, J. W. LIWC-22 的开发和心理测量特性。德克萨斯州奥斯汀：德克萨斯大学奥斯汀分校，1-47。 （2022）。
44. Bond, G. D. & Lee, A. Y.《监狱中的谎言语言：囚犯真实和欺骗性自然语言的语言分类》。应用。认知。心理。 19（3），313-329。 https://doi.org/10.1002/acp.1087 (2005)。
45. 邦德，G.D.等人。 “撒谎的泰德”、“狡猾的希拉里”和“欺骗性的唐纳德”：2016 年美国总统辩论中的谎言语言。应用。认知。心理。 31(6), 668–677。 https://doi.org/10.1002/acp.3376（2017）。
46. Bond, G. D.、 Speller, L. F.、 Cockrell, L. L.、 Webb, K. G. 和 Sievers, J. L. 《瞌睡乔》和《唐纳德，弥天大谎之王》：2020 年美国总统选举辩论中的现实监控和言语欺骗。心理。代表。003329412211052。https://doi.org/10.1177/00332941221105212（2022）。

47. Schutte, M.、Bogaard, G.、Mac Giolla, E.、Warmelink, L.、Kleinberg, B. 和 Verschuere, B. 人与机器：比较手册与用于口头测谎的感知和上下文细节的 LIWC 编码。开放科学中心。 https://doi.org/10.31234/osf.io/cth58（2021）。

48. Kleinberg, B.、van der Toolen, Y.、Vrij, A.、Arntz, A. 和 Verschuere, B. 意图的自动口头可信度评估：模型陈述技术和预测建模。应用。认知。心理。 32、354-366。 https://doi.org/10.1002/acp.3407（2018）。
49. Kleinberg, B. 和 Verschuer, B. 人类如何损害自动欺骗检测性能。心理学学报，213，https://doi。org/10.1016/j.actpsy.2020.103250（2021）。
50. Ilias, L.、Soldner, F. 和 Kleinberg, B. 使用 Transformer 进行可解释的言语欺骗检测。 arXiv 预印本 arXiv:2210.03080 (2022)。

51. Capuozzo, P.、Lauriola, I.、Strapparava, C.、Aiolli, F. 和 Sartori, G. DecOp：用于检测打字文本中的欺骗行为的多语言和多领域语料库。第十二届语言资源和评估会议论文集，1423–1430 (2020)。
52. 萨普，M.等人。量化想象与自传故事的叙事流程。过程。国家。阿卡德。科学。 119(45)，e2211715119。 https://doi.org/10.1073/pnas.2211715119（2022）。

53. Hernández-Castañeda, Á.、Calvo, H.、Gelbukh, A. 和 Flores, J. J. G. 使用支持向量网络进行跨域欺骗检测。软计算。 21, 585–595。 https://doi.org/10.1007/s00500-016-2409-2（2016）。
54. Pérez-Rosas, V. 和 Mihalcea, R. 跨文化欺骗检测。计算语言学协会第 52 届年会记录 2。https://doi.org/10.3115/v1/p14-2072 (2014)。
55. Mihalcea, R., & Strapparava, C. 测谎仪：欺骗性语言自动识别的探索。 ACL-IJCNLP 2009 年会议论文集短论文 309-312。 https://doi.org/10.3115/1667583.1667679（2009）。
56. Ríssola, E. A.、Aliannejadi, M. 和 Crestani, F. 超越建模：了解在线社交媒体中的精神障碍。信息检索进展：第 42 届欧洲 IR 研究会议，ECIR 2020，葡萄牙里斯本，2020 年 4 月 14-17 日，会议记录，第 I 部分 42（第 296-310 页）。施普林格（2020）。
57. 钟，H.W.，等人。扩展指令微调语言模型。 arXiv 预印本 arXiv：2210.11416。 （2022）。
58. Zhou, L., Burgoon, J. K., Nunamaker, J. F. & Twitchell, D. 自动检测基于文本的异步计算机介导通信中的欺骗行为的基于语言学的线索。小组决策。洽谈。 13、81-106。 https://doi.org/10.1023/b：grup.0000011944。62889.6f (2004)。
59. Solà-Sales, S.、Alzetta, C.、Moret-Tatay, C. 和 Dell’Orletta, F. 通过自发语言的语言风格分析证人记忆中的欺骗。脑科学。 13, 317。https://doi.org/10.3390/brainsci13020317 (2023)。
60. Sarzynska-Wawer, J.、Pawlak, A.、Szymanowska, J.、Hanusz, K. 和 Wawer, A. 真相还是谎言：探索欺骗的语言。 PLOS ONE 18，e0281179。 https://doi.org/10.1371/journal.pone.0281179（2023）。
61. Brysbaert, M.、Warriner, A. B. 和 Kuperman, V. 对 4 万个众所周知的英语单词引理的具体性评级。行为研究 46, 904–911。 https://doi.org/10.3758/s13428-013-0403-5（2014）。
62. Lin, Y. C.、Chen, S. A.、Liu, J. J. 和 Lin, C. J. 线性分类器：经常被遗忘的文本分类基线。 arXiv 预印本 arXiv:2306.07111 (2023)。
63. Moore, J. H. Bootstrapping、排列测试和替代数据方法。物理。医学。生物。 44（6），L11（1999）。
64. McGraw, K. O. & Wong, S. P. 共同语言效应大小统计。心理。公牛。 111、361。https://doi.org/10.1037/0033-2909。111.2.361 (1992).
65. Hancock, J. T.、Curry, L. E.、Goorha, S. 和 Woodworth, M. 关于说谎和被说谎：计算机介导的通信中欺骗的语言分析。话语过程。45、1-23。 https://doi.org/10.1080/01638530701739181（2007）。

## 致谢

我们要感谢 Bruno Verschuere 和 Bennett Kleinberg 与我们分享他们的意图数据集的完整版本。

## 作者贡献

G.S. 构思了这项研究。 R.L.、R.R.、P.C. 和 G.S. 设计了这项研究。个人电脑。分享了欺骗性意见数据集的更新版本。 R.L. 进行了描述性语言分析和可解释性分析。 R.R. 制定并实施了微调策略。 R.L. 和 R.R. 撰写了这篇论文。 PP G.S. 监督了整个研究的各个方面并提供了重要的修改。

## 利益竞争

作者声明没有竞争利益。

## 附加信息

补充信息 在线版本包含 https://doi 上提供的补充材料。 org/10.1038/s41598-023-50214-0。

信件和材料请求应发送至 R.L.