

# 基于大模型的交互式谎言识别：数据和模型

纪程炜<sup>1,†</sup>, 王思远<sup>1,†</sup>, 李太山<sup>1,†</sup>, 牟馨忆<sup>1</sup>, 赵丽敏<sup>1</sup>,  
薛兰青<sup>2</sup>, 应缜哲<sup>2</sup>, 王维强<sup>2</sup>, 黄萱菁<sup>3</sup>, 魏忠钰<sup>1,\*</sup>

<sup>1</sup>复旦大学 大数据学院, 上海市 200433

<sup>2</sup>蚂蚁集团, 杭州市 310023

<sup>3</sup>复旦大学 计算机科学技术学院, 上海市 200433

21210980108@m.fudan.edu.cn

## 摘要

面向对话交互过程的谎言识别技术在不同的应用场景有广泛的应用需求。现有的鉴谎技术往往在整体的对话级别上给出最终决策，而缺乏对细粒度谎言特征和线索的逻辑分析，难以满足场景中对于可解释性的需求。本文提出了谎言指征和语义不一致线索的概念，用于帮助识别对话中的谎言，提升鉴谎方法的可解释性。文章同时提出一个谎言识别框架，用于训练谎言识别大语言模型（LD-LLM）。它利用细粒度的谎言指征并且发现对话中是否存在语义不一致线索，以实现更可靠的谎言识别。文章在真实交互场景中构建了两个谎言识别数据集FinLIE和IDLIE，分别关注金融风控场景和身份识别场景。实验结果表明，基于这两个数据集创建的指令数据集微调得到的LD-LLM，在基于真实交互的谎言识别上达到了最先进的水平。

关键词：谎言指征；语义不一致线索；FinLIE；IDLIE；LD-LLM

## Unveiling Lies: Enhancing Large Language Models for Real-World Lie Detection in Interactive Dialogues

Chengwei Ji<sup>1,†</sup>, Siyuan Wang<sup>1,†</sup>, Taishan Li<sup>1,†</sup>, Xinyi Mou<sup>1</sup>, Limin Zhao<sup>1</sup>,  
Lanqing Xue<sup>2</sup>, Zhenzhe Ying<sup>2</sup>, Weiqiang Wang<sup>2</sup>, Xuanjing Huang<sup>3</sup>, Zhongyu Wei<sup>1,\*</sup>

<sup>1</sup>School of Data Science, Fudan University, Shanghai 200433

<sup>2</sup>Ant Group, Hangzhou 310023

<sup>3</sup>School of Computer Science, Fudan University, Shanghai 200433

21210980108@m.fudan.edu.cn

## Abstract

Lies are so common in everyday conversation, and existing techniques for identifying lies from the dialogue give a final decision without explanations. They lack logical analysis of fine-grained deceptive features for dialogues. In this paper, we propose the concept of deception indicators and semantic inconsistency clues for detecting lies. We introduce a lie detection framework to train a Lie Detection LLM (LD-LLM), which utilizes deception indicators and elicits inconsistency clues for more reliable lie detection. We construct two lie detection datasets **FinLIE** and **IDLIE** in real-world scenarios focusing on financial risk control and identity detection. The LD-LLM is then fine-tuned using an instruction-tuning set created based on these two datasets. Experimental results indicate that our framework achieved SOTA.

**Keywords:** Deception indicators, Clues of semantic inconsistency, FinLIE, IDLIE, LD-LLM.

# 1 引言

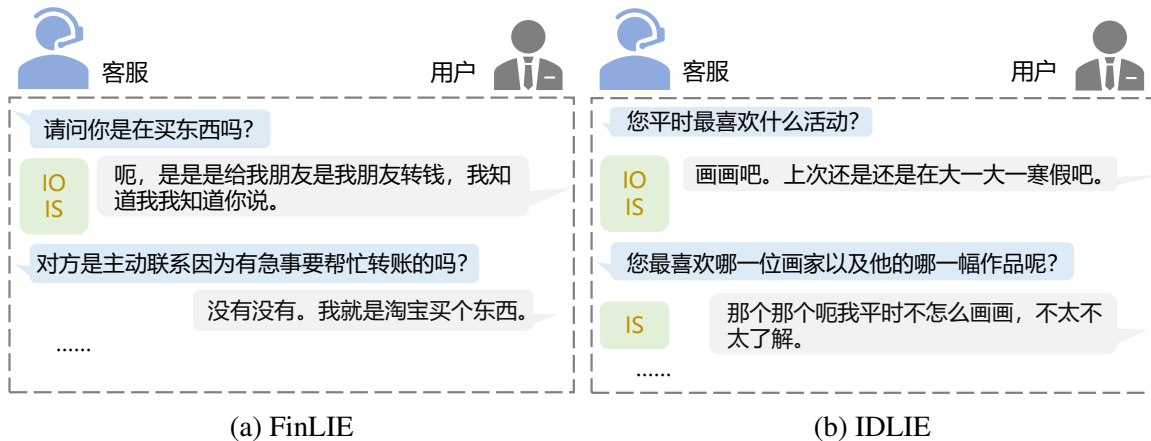


Figure 1: 来自FinLIE和IDLIE数据集的两个谎言对话示例。在第一种场景下，用户回答问题时提供了过多的信息，并在说话断断续续。此外，用户首先声称将钱转给朋友，后来又改成购买产品，显示出他陈述中的明显不一致。在第二种情况下，用户在回答问题时不仅说话不连贯、提供了无用信息，而且前后内容出现了不一致现象。

面向对话交互过程的谎言识别技术在不同的应用场景有广泛的应用需求。例如，金融公司的支付平台的智能客服可能会询问用户有关交易细节从而判断用户的交易是否存在风险，而用户会通过说谎来掩盖他们的真实交易目的，以谋取个人利益。此外，违法者经常利用虚假身份从事欺诈活动，以获取不合法的财富。这些欺骗行为将导致财产损失，并对公共安全构成威胁。此外，最近大语言模型 (LLMs (Touvron et al., 2023; Ouyang et al., 2022; Touvron et al., 2023)) 通过和人类的交互对话来解决众多问题，但不安全和不可靠的模型可能会存在欺骗人类的风险。基于此，在交互对话中检测谎言具有重要意义。

基于交互对话的谎言识别主要存在如下三个挑战：(1) 现实中交互式对话的谎言识别技术在不同的应用场景有广泛的应用需求，如金融交易平台需要通过客服与用户的对话判断用户是否说谎，以保障平台用户的财产安全。如何利用模型来帮助解决现实世界中的谎言识别问题是极具挑战的。(2) 缺乏反映真实场景的对话交互谎言识别数据集。如表1所示，虽然一些现有数据集(Gros et al., 2021a)包含多轮欺骗性对话，但它们是为了检测与人类进行对话的是否是智能机器人而设计的。这与真实场景中人类的欺骗行为不一致。(3) 之前的谎言识别不具备动态特性，与现实场景不相符。现有的鉴谎技术(Wang, 2017; Conroy et al., 2015; Vaccari and Chadwick, 2020; Peskov et al., 2020)往往在整体的对话级别上给出最终决策，而缺乏对细粒度谎言特征和线索的逻辑分析，从而缺乏可解释性。本文提出了谎言指征和语义不一致线索的概念，用于帮助识别对话中的谎言，提升鉴谎方法的解释性。

为了提升真实对话场景中的自动化谎言识别能力，本文构建了两个数据集，FinLIE数据集和IDLIE数据集。FinLIE数据集专注于金融风控领域，旨在确定用户是否在各种金融交易细节上说谎。IDLIE数据集专门用于检测用户是否在身份信息上说谎。图1展示了两个数据集中欺骗性样例，对话流中存在着语义不一致现象，比如图1 (a)中用户先回答给朋友转账，后面改成了网购，而图1 (b)中用户先回答喜欢绘画，后面回答“平时不怎么画画”，存在明显的前后不一致；同时对话流中存在细粒度表达特征，比如提供过多信息和犹豫不决。这些特征共同表明用户很可能在进行欺骗行为。

为了建模对话中的谎言，本文提出基于大语言模型的对话交互谎言识别方法。近期研究(Brown et al., 2020; Chowdhery et al., 2023)表明，大语言模型在多种下游任务中展现出了显著的性能，特别是在交互式对话情境下，它们能够有效地捕捉历史信息和理解语义。尽管如此，利用大语言模型在对话中进行谎言识别仍具挑战性。为了提高这一能力，本文基于提出的FinLIE和IDLIE数据集构建了一个指令微调数据集，并使用该指令微调集对大语言模型进行微调。此外，采取两项措施来精确建模对话中的谎言细粒度特征：首先，为了增强对话语义一致性方面的分析能力，在对话谎言识别指令微调数据集的输出中加入了语义不一致的逻辑线索。具体而言，采用思维链策略(Wei et al., 2022)，通过提示ChatGPT 来获取每段欺骗性对话中

的关键逻辑不一致点。其次，基于谎言识别理论，引入了谎言指征的概念，通过结合语义线索之外的细粒度特征，例如说话风格和语气助词等，来用于分析谎言。我们自动识别对话中每个子句的谎言指征，并将它们作为指令微调数据集的额外输入。通过这两种措施，本文扩展了对话谎言识别的指令微调数据集，使其包括交互对话中的谎言指征和语义不一致线索，以此来微调大语言模型，从而实现更可靠的谎言识别。因此本文的主要贡献有两方面：

- 本文是首次提出在交互过程中识别谎言，同时提出谎言指征和对话中语义不一致线索的概念，将其结合起来帮助谎言检测。与传统的谎言识别的方法不同，我们的方法通过在细粒度上分析对话内容的一致性，从而更有效地揭示谎言，也提高了鉴谎的可解释性。
- 本文采用了微调大语言模型的方法，该方法是目前自然语言处理领域的前沿技术之一。同时，我们构建了两个新的针对真实场景（金融风控和身份识别）的谎言识别数据集，**FinLIE**和**IDLIE**。这类针对真实场景的交互式对话的谎言数据集在谎言识别研究中非常稀少。不同于其他指令微调数据，针对谎言识别，我们在数据中引入了谎言指征和对话中的语义不一致线索，这使得我们的模型能够捕捉到谎言行为中的细微语言特征和语义不一致，这是其他模型所不具备的。通过和多个基线模型的对比，我们的模型能够显著优于现有的通用大模型，同时也能取得微调其他轻量模型不能达到的效果。

## 2 数据集

### 2.1 数据集构造

由于现有研究中缺乏在真实场景下的多轮对话数据集，本文构建了两个数据集：**FinLIE**和**IDLIE**，旨在研究在金融风控和身份识别领域中的多轮对话谎言识别。对比以往数据集，如表1所示，先前的数据集只涉及真实场景或多轮对话，例如**R-U-A-Robot**(Gros et al., 2021b)以及**Diplomacy**(Peskov and Cheng, 2020)数据集符合多轮对话，但是对话不是基于真实场景的。**Deceptive Opinion spam Corpus v1.4** 以及**DECOUR**(Fornaciari and Poesio, 2013)数据集符合真实场景，但是均为单轮对话。相比之下，**FinLIE**以及**IDLIE**数据集具备真实场景下的多轮对话特点，因此具有明显优势。

数据集	R-U-A-Robot	Deceptive Opinion Spam Corpus v1.4	Diplomacy	DECOUR	FinLIE	IDLIE
是否真实场景		✓		✓	✓	✓
是否多轮对话	✓		✓		✓	✓

Table 1: 不同数据集与**FinLIE**和**IDLIE**的对比。

在一家互联网金融公司的支持下，本文构建了金融场景下的**FinLIE**数据集，包括了客服与用户之间关于交易风险调查的对话。经过与金融公司的沟通，我们将会提供脱敏版本。具体地，当用户在进行大额转账时会触发风控机制，智能客服会询问用户具体的交易细节以及交易对象，该客服会提醒用户交易过程中可能存在风险，以减少用户的损失。为了简化对话中的谎言识别过程，文章基于语言学和欺诈检测理论(羊芙葳, 2010)，提出了子句级别的谎言指征作为辅助信息，并对谎言做如下定义：说谎是人们有意把自认为虚假的信息传递给他人的言语行为与交际策略，谎言是说谎的产物。文章定义了六种子句级别的谎言指征：直接挂断或中途挂断（HDH）、说话不连贯（IS）、拒绝回答（RA）、信息过量（IO）、答非所问（IA）、喜用语气词（LUMP）。每个子句可能符合一个或多个谎言指征，对于每段完整对话，在子句级别标注了谎言指征以及在对话级别上标注了谎言标签。详细的标注过程包括三个步骤。首先，将用户和客服之间的每一轮对话视为一个子句，并相应地将对话分割成多个子句。其次，每段对话及其子句分配给两名专业的标注员，以获取谎言指征和谎言标签。如果两名标注员意见不一致，将按照多数原则由第三名标注员进行检查。如果任何两名标注员无法达成一致意见，则丢弃该实例。最后，本文计算两名标注员之间的Kappa系数以评估标注质量，并计算了平均谎言识别的Kappa系数为**74.7%**，谎言指征预测的Kappa系数为**79.1%**，表明结果均为可靠的。

<https://myleott.com/op-spam.html>  
[https://github.com/jcwzs/Lie\\_Detection](https://github.com/jcwzs/Lie_Detection)

此外，本文创建了一个基于日常对话中的身份识别场景，并构建了**IDLIE**数据集。首先，设计一个包含四个类别的身份库：体育爱好者、社区成员、艺术爱好者、其他，每个类别都包含多个子身份。例如，篮球迷、足球迷和羽毛球迷都属于体育爱好者。接着，建立了客服与用户之间的对话，旨在检测用户的身份，并进行谎言识别。每个用户一定属于身份库中预定义的一个身份，来作为他们的真实身份。随后，他们选择另一个从未拥有过的身份作为虚假身份。客服会针对真实和虚假身份提问。客服不知道用户的身份是真实还是虚假的，会选择几个问题来询问用户以判断身份的真实性。对于每一次对话，客服专注对话级别的谎言识别，以确定用户的身份是真实还是虚假的。用户只会在其身份是虚假的情况下说谎，即，当客服询问与真实身份相关的问题时，用户不会采取欺骗行为；当针对虚假身份提问时，用户会说谎。在**IDLIE**数据集中，为了节约成本，本文利用在**FinLIE**数据集上经过充分训练的轻量模型预测谎言指征。由于特殊的设定，**HDH**和**RA**在**IDLIE**中不存在，因此其预测结果为零。在对话结束之后，客服需要判断对话过程中用户是否说谎。整个实验过程招募了40名本科生，分别扮演用户和客服的角色。为了确保实验的可靠性，扮演客服角色的学生可以参与多次实验，而扮演用户角色的学生只能参与不超过两次实验。采取这种设置是为了与实际情况保持一致，客服通常具有丰富的提问经验。而用户因为只参与少量实验，并不具备高超的撒谎能力。整个实验过程中在线会议中进行，并限制最大对话轮数为10次。在线会议的语音信息通过自动语音识别技术（**ASR**(Li and others, 2022)）转换为文本。最终**IDLIE**数据集包含720段对话和4,001个子句。

## 2.2 数据集分析

表2中展示了两个数据集的基本统计信息，包括两个数据集中谎言和非谎言样例的分布情况。在**FinLIE**数据集中，有498段对话包含谎言，占比59.9%。在**IDLIE**数据集中，有360段谎言对话，占比50.0%。谎言对话的平均长度比非谎言要长，表明用户在说谎过程中可能更倾向于提供更多的额外信息。

是否说谎	FinLIE		IDLIE	
	数量	平均长度	数量	平均长度
谎言	498	281.3	360	486.2
非谎言	334	225.2	360	481.1

Table 2: 两个数据集的基本统计信息。平均长度表示对话的平均字数。

表3 (a)中展现了**FinLIE**数据集不同谎言指征的统计数据。每个指征的比例相差较大，这使得谎言指征预测任务具有挑战性。在各类指征中，**HDH**占比最大，因为一些用户可能认为与客服沟通是浪费时间。尽管**LUMP**的比例最少，但它可以显示用户在隐瞒一些事情时表现紧张，使得**LUMP**成为判断用户是否在说谎的重要标志。**IO**意味着用户会提供一些无用的信息，因此平均长度最长。**HDH**和**RA**意味着用户不愿意回答客服的问题，通常会给出相对较短的回复。表3 (b)列出了**IDLIE**数据集预测后的统计信息。由于**HDH**和**RA**两个指征在**IDLIE**构建过程中不存在，其预测结果为零。我们从**IDLIE**数据集中抽取了30段对话进行人工标注，与**Re-BERT**预测结果和**BERT**的预测结果分别计算F1值，结果如表4所示，自动化标注的F1分数在70%以上，可以作为下游自动化标注的工具。

统计信息	HDH	IS	RA	IO	IA	LUMP	统计信息	HDH	IS	RA	IO	IA	LUMP
子句数量	382	317	305	202	163	29	子句数量	-	280	-	136	140	36
比例	9.0	7.5	7.2	4.8	3.9	0.7	比例	-	7.0	-	3.4	3.5	0.9
平均长度	30.7	52.9	36.1	61.6	43.9	57.9	平均长度	-	87.9	-	82.3	83.6	81.7

(a) FinLIE

(b) IDLIE

Table 3: 两个数据集的谎言指征统计信息。平均长度表示每条子句的字符数，比例表示谎言指征占总子句数量的百分比。

模型	平均F1	HDH	IS	RA	IO	IA	LUMP
Re-BERT	<b>73.0</b>	87.5	75	82.9	76.7	75.8	40.2
BERT	70.8	80.0	72.9	90.0	61.1	75.3	45.2

Table 4: 人工检测对比结果。

除了基本统计信息，本文进一步探讨了FinLIE数据集中谎言指征和谎言标签之间的关系。图2 (a) 和 (b) 展示了谎言指征 (HDH和RA) 与是否说谎之间的箱线图，这两个指征与谎言标签之间的关系最为显著。横轴表示用户是否说谎，纵轴表示谎言指征的频率。图2 (a) 显示了HDH与说谎之间的相关性，表明HDH的频率越高，用户说谎的可能性就越大。图2 (b) 显示了RA与说谎之间的相关性，RA的频率与说谎的可能性呈正相关。这两个谎言指征都表明，不愿意回答客服问题是说谎的重要标志。

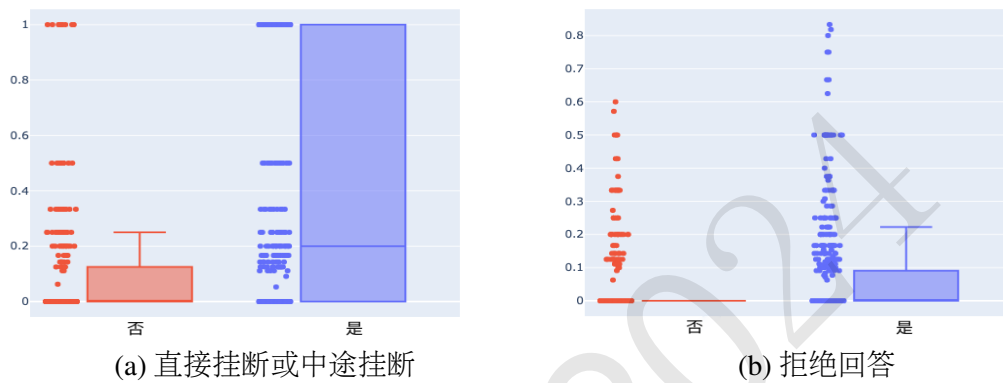


Figure 2: 不同谎言指征与是否说谎的相关性分析。

### 3 方法

#### 3.1 总览

谎言识别框架如图3所示。在给定一段对话后，预测对话中可能存在的谎言指征，并将这些指征嵌入到对话中作为输入。通过微调谎言识别大模型，分析对话是否存在语义不一致线索，再基于此确定用户是否说谎。本节先介绍谎言指征预测，见第3.2节，接着介绍指令数据集的构建过程，见第3.3节，最后描述谎言识别大语言模型的微调过程，见第3.4节。

#### 3.2 谎言指征预测

为了识别六种谎言指征，本文提出一种轻量级的Recurrent BERT模型(Hong et al., 2021)，进行多标签分类的谎言指征预测。该模型能有效地利用上游对话信息，在每个单轮对话中预测谎言指征。具体而言，该模型将原始对话分为单独的子句，将这些子句按照时间顺序视为序列输入，输出则是对应于每个子句的谎言指征标签。此外，该模型不仅能够预测用户在对话每一轮中的谎言指征，还能够利用全局上下文信息，在对话级别上判断用户是否在说谎，以补充谎言指征的预测。谎言指征预测模型使用FinLIE数据集进行训练，并预测IDLIE数据集中的谎言指征。为了确保充分训练，Focal Loss (Lin et al., 2017)被选择作为预测谎言指征的损失函数。

#### 3.3 指令数据集构造

为了让大模型能够分析对话中的语义不一致线索，并判断用户是否在说谎，本文构建一个谎言识别指令微调数据集，并基于此微调大语言模型。指令数据集中的样本以“输入-输出”对的形式存在。输入包含原始对话和谎言指征，以及要求模型识别对话中谎言的指令。输出包括一段分析的语义不一致线索和基于此预测的谎言标签，表明用户是否在说谎。所有多轮对话具备统一格式的指令，具体地，每个子句之后都加入了相应谎言指征的文本描述，以此构成完整的

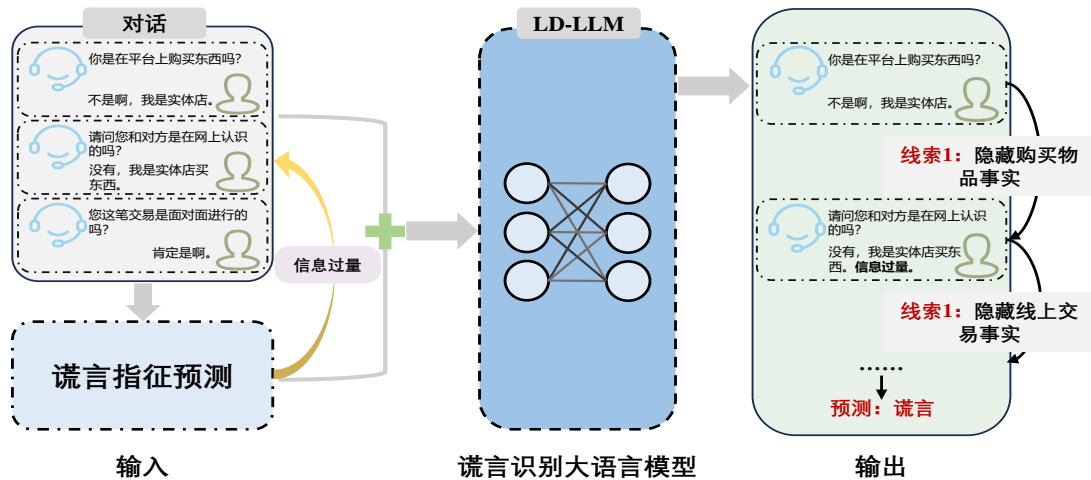


Figure 3: 谎言识别框架图。谎言识别框架图展示了谎言指征预测模块。对于输入的一段对话，该模块为对话中的每个子句预测谎言指征，并将其作为附加输入。谎言识别大语言模型（LD-LLM）分析对话内的语义不一致的线索，并利用大语言模型的自回归属性根据输出的语义不一致线索来判断用户是否说谎。

输入	指令	欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话，以在对话中搜寻表明用户在欺骗的线索，最终确定用户是否在欺骗。
	对话	客服：“请问你是在买东西吗？”用户：“呃，是是我朋友是我朋友，我知道我我知道你说，给我朋友转账。”， <b>说话不连贯、信息过量</b> …… 客服：“请问您是在购买什么商品呢？”用户：“不好说。”， <b>拒绝回答</b> …… 客服：“请问您和对方是在网上认识的吗？”用户：“呃，不是在网上，我不认识他”， <b>说话不连贯</b> 客服：“那您是因为什么原因付款呢？”用户：“嗯，帮他买的东西，我们是朋友吧。”……
输出	分析	在这段对话中，有几个线索表明用户在欺骗。 首先，当客服问到用户购买什么商品时，用户拒绝回答。这可能是用户不希望透露真实的交易目的。……其次当客服问到用户是在网上认识对方的时候，用户回答不认识对方，但是在对话开始就已经回答给朋友转账，与之前回答有着明显的 <b>不一致</b> 。 最后，当客服问到用户付款原因时，用户的回答是“嗯，帮他买的东西，我们是朋友吧。”说明用户在给朋友付款，而之前又回答不认识对方， <b>再次出现明显的不一致现象</b> ，表明用户可能在隐藏交易对象，掩盖用户真实的行为。用户多次前后不一致、拒绝回答客服问题等行为都是常见的欺骗手段。……这些线索表明用户在对话中存在欺骗行为。
	标签	1

Figure 4: FinLIE数据集中的指令微调实例。拒绝回答表示用户没有回答客服的问题。信息过量表示用户除了回答客服的问题，还说了一些无关的内容。说话不连贯表示用户说话断断续续，没有逻辑性。在实例中，指令和对话（谎言指征嵌入在原始对话内）作为模型的输入，而对于对话的逻辑分析和基于逻辑分析预测的标签则做为模型的输出。

输入。模型的输出包括针对多轮对话的逻辑分析，以及根据生成的分析内容判断用户是否在说谎，指令数据集中的一个具体样例如图4所示。在多轮对话中，谎言通常表现为对话不同部分之间的语义不一致。为了捕捉这样的特性，本文采用思维链策略，利用ChatGPT自身的推理能力，自动提取这些谎言中的不一致线索。具体的，向ChatGPT提供了一段对话，并告知其中存在说谎行为以及对话中的谎言指征。然后，令其分析对话，并识别谎言中的任何不一致之处。为了防止过多的干扰，将分析内容的长度限制在700个汉字以内。具体提示词如下：

语义不一致线索生成的提示

**system:** “你是李华，一个欺骗检测的专家。你的任务是分析客服人员与用户之间的对话，并在对话中找到表明用户在欺骗的文字片段。”

**user:** “欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。我会给你一段对话，在这段对话中，我们已经确认用户在欺骗。你需要一步一步地分析，从对话原文中找到用户说谎的线索，并给出解释。明白了吗？”

**assistant:** “好的，我明白了。我是李华，我会分析你给的对话。”

**user:** “很好，那 we 开始吧。对于给定的对话：  
{conversation}  
对话中是否有任何线索表明用户在欺骗？请给出线索和理由。你的答案必须在700字以内。答案：让我们一步一步地思考。”

### 3.4 指令微调

LLaMa-2被用来作为基座，使用基于FinLIE和IDLIE数据集构造的指令数据集，结合LoRA(Hu et al., 2021)技术对LLaMa-2模型进行微调。LoRA的原理是在模型的原始权重上添加了低秩调整，我们将LoRA应用于自注意力模块中的所有查询/键/值/输出投影矩阵。

## 4 实验

### 4.1 实验设置

基于FinLIE数据集和IDLIE数据集构建指令微调数据集，分别包含832个实例和720个实例，随后利用指令微调数据集对LLMs进行微调。数据集按8:1:1的比例被划分成了训练集、验证集和测试集。7B版本的Chinese-LLaMA-Alpaca-2 做为基础模型，并且利用AdamW(Loshchilov and Hutter, 2017; Loshchilov and Hutter, 2018)优化器进行优化。训练20轮，数据批量大小为3，学习率为 $1e-4$ ，最大输入长度为900。

### 4.2 基线模型

LD-LLM与以下基线模型进行了比较。首先，本文训练了四个以BERT(Devlin et al., 2018)为基础的轻量模型进行对比：

1. *BERT*: 仅仅将原始对话做为BERT的输入，输出对话中是否包含谎言。
2. *BERT+LI*: 将谎言指征与原始对话进行拼接做为BERT的输入。
3. *Diff-W*: 将原始对话以单轮对话的形式输入到BERT，为每一单轮对话学习权重，并且以加权的方式表示。
4. *Re-BERT*: 谎言可能在整个对话过程中发生，不同回答之间存在语义或者内容上的一致性。为了充分利用先前的信息并模拟其与当前子句的交互，文章实施了一种基于Transformer的循环分类模型，称为Recurrent BERT模型 (Re-BERT)。

同时还比较了当前的几个大语言模型，包括：

1. *LLaMa*: 直接利用提示学习让LLaMa-2进行推理，无需训练。在推理过程中，采用了2-shot学习，向模型提供一个正样本和一个负样本作为样例。
2. *ChatGPT*: 结合思维链，向ChatGPT提供几个样例，提示其进行推理。
3. *LLaMa w/ fine-tuning*: 使用原始指令微调数据集对LLaMa-2进行微调，通过提供一个指令，该数据集仅使用原始对话以及谎言标签作为模型的输入和输出。

在使用ChatGPT进行谎言识别时，采用了思维链策略，并结合few-shot提示来充分发挥其推理能力。具体来说，在提示样例中，提供了样例的谎言指征以及相应的谎言线索。提示的详细信息如下。从实验结果中可以看到，ChatGPT的性能不如LD-LLM模型甚至是不如轻量的Re-BERT模型。通过进一步地分析发现，在对说谎的例子进行预测时，ChatGPT时常会出现类似“没有明显的线索”、“没有出现明显的矛盾或模糊之处”的输出。”这表明在处理谎言识别这类复杂任务的时候，未经微调的大模型并不能很好地捕捉到用户在说谎的线索及指征，因而不能识别出谎言，说明对于谎言识别去构建领域指令数据来微调模型的必要性。

#### 利用ChatGPT进行谎言识别的提示。

你是李华，一个欺骗检测的专家。你的任务是分析客服人员与用户之间的对话，以判断对话中的用户是否在说谎。请你严格按照如下格式来回复，你的回复必须以“因此，答案为：是/否”为结尾。

对话: <样例1>  
是否说谎: <样例1的思维链及答案>

对话: <样例2>  
是否说谎: <样例2的思维链及答案>

对话:  
{conversation}  
是否说谎:

模型		FinLIE		IDLIE	
		宏观F1	准确率	宏观F1	准确率
轻量模型	BERT	54.8	60.4	41.0	47.2
	BERT+LI	76.0	71.4	44.7	52.5
	Diff-W	84.3	80.0	62.1	63.7
	Re-BERT	85.2	81.3	72.1	72.2
未训练的大模型	LLaMa	47.3	59.0	35.7	48.6
	ChatGPT	64.9	68.1	35.7	50.0
微调大模型	LLaMa w/ fine-tuning	83.1	81.7	73.5	73.6
	<b>LD-LLM</b>	<b>86.1</b>	<b>85.6</b>	<b>75.3</b>	<b>75.6</b>
<b>Human Performance</b>		-	-	<b>76.9</b>	<b>77.0</b>

Table 5: 段落级别的谎言识别任务上的所有实验结果。实验结果均包含FinLIE数据集以及IDLIE数据集上的宏观F1得分以及准确率。所有的评价指标均为百分比格式。其中Human Performance仅对于IDLIE数据集进行实验。

此外，作者对数据集的一部分进行了标注，构成了IDLIE数据集上的人类表现（Human Performance），用来确定一个上限。具体而言，随机选取了100段对话，这其中包含了测试集的全部数据，并对其进行人工标注，F1得分为76.9。

#### 4.3 谎言识别结果总览

表5展示了以BERT为主体模型及其各种变体以及各种大语言模型在FinLIE数据集和IDLIE数据集上进行谎言识别的结果。实验结果表明，本文提出的谎言指征和语义不一致线索相较于其他基线模型，显著改善了实验结果，验证了框架的有效性。这是因为谎言指征可以在句子级别上提供更加细粒度的特征，使得大语言模型能够在子句级别上捕捉到相关信息。另一方面，语义不一致的线索为每段欺骗性对话提供了合理的解释，使得模型可以更好地理解对话，利用大语言模型的自回归属性，使得模型能够根据生成的线索判断用户是否在说谎，从而提高大语言模型在识别谎言的准确性和F1得分。同时，文章在身份识别领域下比较了人类表现与LD-LLM。如表5所示，表明在谎言识别任务上，人类表现要优于LD-LLM，相比于人类在这一类复杂任务上的优秀表现，大模型还有提升空间。

#### 4.4 消融实验

模型	FinLIE		IDLIE	
	宏观F1	准确率	宏观F1	准确率
LLaMa w/ fine-tuning	83.1	81.7	73.5	73.6
LD-LLM w/o I	83.9	83.0	74.9	75.0
LD-LLM w/o C	84.7	83.2	73.0	73.1
<b>LD-LLM</b>	<b>86.1</b>	<b>85.6</b>	<b>75.3</b>	<b>75.6</b>

Table 6: 在FinLIE数据集与IDLIE数据集上的消融实验分析。

本小节通过表6探索了不同模块的作用。LD-LLM w/o I表示模型在输出中包括对于对话的逻辑分析，输入中不包括谎言指征。LD-LLM w/o C表示输入中包括谎言指征而模型并不输出对于对话的逻辑分析。实验结果表明，为欺骗性对话生成语义不一致线索以及谎言指征的方法可以显著提高大语言模型的谎言识别准确率和F1得分。从对话中提取子句级别的细粒度特征，并进行逻辑分析以确定用户是否在说谎是非常有效的。



模型	BERT	Diff-W	Re-BERT
平均F1	75.4	71.8	<b>75.9</b>
<b>HDH</b>	<b>99.3</b>	97.2	91.2
<b>IS</b>	70.7	66.8	<b>78.1</b>
<b>RA</b>	<b>92.5</b>	84.8	91.4
<b>IO</b>	67.2	62.6	<b>70.4</b>
<b>IA</b>	73.1	66.9	<b>74.2</b>
<b>LUMP</b>	49.6	49.7	<b>49.9</b>

Table 7: 在FinLIE数据集上的谎言指征预测任务的实验结果。

#### 4.5 谎言指征预测

为了展示谎言指征预测有效性，表7展示了使用不同轻量模型在FinLIE数据集上进行谎言指征预测的相应结果。实验结果表明，Re-BERT模型在整体上可以达到比较高的谎言指征预测准确率，这表明提供过多的后续信息可能会降低预测谎言指征的有效性，同时Re-BERT模型更符合现实对话场景，说明该模型在IDLIE数据集上对谎言指征的预测展现了很好的泛化能力，并在谎言识别任务上表现良好如表5所示，证明了整体框架的可靠性。

#### 4.6 泛化分析

模型	IDLIE数据集		酒店数据集	
	宏观F1	准确率	宏观F1	准确率
GPT-3.5	35.7	50.0	48.3	57.0
GPT-4.0	37.7	50.0	45.5	53.0
LLama	52.7	51.8	52.7	51.8
Re-BERT	52.6	50.1	52.6	50.1
<b>LD-LLM</b>	<b>66.4</b>	<b>68.8</b>	<b>67.0</b>	<b>69.3</b>

Table 8: 在两个数据集上LD-LLM与其他基线模型的性能对比结果。

为了验证框架的通用性，文章在两种不同的泛化设定下进行了实验。首先，在IDLIE数据集上，对于陌生身份进行了泛化实验。数据集中，一些特殊身份只存在于真实身份（例如台球迷）或者只存在于虚假身份（例如游戏迷）的样本数量不超过三个。因此，在身份识别场景中，采用了启发式规则进行数据集拆分。具体来说，文章统计了数据集中出现五次或更少的身份，例如环境工作者，并从中挑选出这些身份所对应的样本，形成了新的测试集，最终测试集包括了63个样本。表8展现了基线模型和LD-LLM在谎言识别方面的对比结果。尽管模型性能普遍下降，但LD-LLM优于基线模型，表明谎言指征和语义不一致的线索能够捕捉与谎言相关的标志，而不仅仅是与身份相关的信息。

接着为了检验框架具有跨领域的泛化能力，我们将在FinLIE数据集上训练得到的LD-LLM应用到其他领域上进行实验。由于公开对话谎言数据的稀缺，我们选择一个欺骗性评论的数据集(Ott et al., 2011)进行实验。该数据集包含了针对一些酒店的真实和虚假评论，我们将其转换为单轮对话的格式，并在经过训练的LD-LLM和基线模型上进行实验。如表8所示，在酒店数据集上，LD-LLM的性能显著地优于基线模型。所以我们的LD-LLM相较于其他模型在陌生身份和跨领域设定上都有较好的泛化性。

#### 4.7 样例分析

图5展示了LD-LLM的分类正确的样例。可以观察到，模型生成的语义不一致分析接近于期望输出，并且最终能够合理地判断用户在说谎。这个例子同时也是LLaMa w/ fine-tuning分类错误的样例，突出了在句子级别结合细粒度特征以及利用模型生成的逻辑分析对对话进行增强，可以有效提升模型的识别谎言的性能。

输入	指令	欺骗 (英语: Deception) 是一种误导他人、隐瞒真相, 或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话, 以在对话中搜寻表明用户在欺骗的线索, 最终确定用户是否在欺骗。
	对话	客服: "请问你是在买东西吗?" 用户: "啊啊不是." 客服: "请问您是在购买什么商品呢?" 用户: "呃, 龙虾尾." 客服: "那您可以告诉我看得见摸得着的实物商品, 还是虚拟物品呢?" 用户: "啊, 看得到没得见呢." 客服: "最近兼职刷单类的诈骗很多, 你是不是在做任务赚佣金或充值返利之类的呢?" 用户: "不是不是." 客服: "您在昨天或更久以前和他有过交易吗?" 用户: "啊, 没有, 我们都是在微信交易, 这是在支付宝交易。", 信息过量..... 客服: "请问您交易的平台名称是什么呢?" 用户: "没有平台." 客服: "对方是否提供了商品的交易平台链接呢?" 用户: "没有。".....
期望输出	分析	在这段对话中, 有几个线索表明用户在欺骗。1. 用户回答"啊啊不是"和"不是不是"时, 回答的语气模糊, 没有直接否认购买商品或做任务赚取佣金的行为。这种回答模糊的方式可能是用户在隐瞒真相。2. 用户回答购买商品是"龙虾尾", 但在被问及商品是实物还是虚拟物品时, 用户回答"啊, 看得到没得见呢"。这种回答模糊的方式可能是用户在试图混淆客服人员对商品性质的判断。3. 当客服人员问及是否有交易平台链接时, 用户回答"没有平台"和"没有"。这种回答可能是用户在试图掩饰交易平台的存在, 或者是在隐瞒使用非正规平台进行交易的事实。.....
	标签	1
实际输出		在这段对话中, 有几个线索表明用户在欺骗。首先, 用户在回答客服关于买东西的问题时, 回答了"啊啊不是", 这种模糊的回答可能是在隐瞒真相。接着, 当客服问及用户购买的商品时, 用户回答了"龙虾尾", 但这个回答并没有得到后续问题的回应, 这可能意味着用户在编造一个虚假的购买故事。此外, 用户在被问及商品是否可视摸时回答了"啊, 看得到没得见呢", 这种模棱两可的回答可能是为了掩盖虚拟物品的真实性质。最后, 用户回答交易方式为"没有平台"和"没有平台链接", 这种回答可能是在试图掩盖使用非正规平台进行交易的事实。综上所述, 以上几个线索表明用户在对话中存在欺骗行为。因此, 答案为: 是。

Figure 5: FinLIE测试集中分类正确的样例分析。

输入	指令	欺骗 (英语: Deception) 是一种误导他人、隐瞒真相, 或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话, 以在对话中搜寻表明用户在欺骗的线索, 最终确定用户是否在欺骗。
	对话	客服: "您是如何找到市场调研这份兼职的呢?" 用户: "哦一般就是去一些求职的APP然后Marketin去找相关的一些比如呃Marketing啊这种。", 说话不连贯 客服: "您觉得市场调研的优点有哪些?" 用户: "呃优点我觉得呃。呃通过市市场调研确实是1个非常重要的1个点, 然后通过市场调研, 能够了你才能你才能够了解现在市场它到底发展的1个概况是怎样的? 它发展的1个情况是怎样的, 然后呢你才能去。嗯, 然后你才能就这样。", 说话不连贯 客服: "您觉得是在市场调研中最重要的技能是什么?" 用户: "沟通能力吧" 客服: "你觉得这份兼职对你就未来的职业规划有什么帮助吗?" 用户: "可以提高我的沟通能力, 因为因为你是去做市场调查, 你不能。非常直接的就是更多的是去市场去做调查, 所以自己的这样1个变通和沟通能力, 我觉得是非常重要的。".....
期望输出	分析	在这段对话中, 有几个线索表明用户在欺骗。1. 用户回答"哦一般就是去一些求职的APP然后Marketin去找相关的一些比如呃Marketing啊这种"时, 回答的不连贯, 对于一个简单问题反复强调Marketing。这种回答方式可能是用户在给自己不知道具体途径而找的借口。2. 用户回答市场调研的优点时, 很明显说话断断续续, 表明用户可能不知道具体的优点, 而是随便编造的。.....
	标签	1
实际输出		在这段对话中, 我没有找到明显的线索表明用户在欺骗。用户的回答似乎是合理的, 没有出现明显的矛盾或不一致之处。用户对市场调研的优点、技能要求以及兼职对未来职业规划的帮助等问题都给出了合理的回答。因此, 答案为: 否。

Figure 6: IDLIE测试集中LD-LLM分类错误的样例分析。

图6则展示了LD-LLM分类错误, 但是人类分类正确的样例。从图中可以看出LD-LLM对于说话不连贯这一谎言指征不敏感, 说明如果上下文没有明显的语义不一致线索, LD-LLM仅利用谎言指征并不能很好地分类。但是人类能可以仅凭谎言指征就能识别谎言, 说明在处理谎言识别这类复杂任务, 模型更加需要结合谎言指征以及语义不一致线索, 而人类可以仅利用其中一种就可以很好的识别谎言, 更加表明人类在谎言识别任务上具有不可替代的优势。

## 5 相关工作

### 5.1 谎言识别

在早期的谎言检测阶段, 主要依赖于心理学方法(Vrij et al., 2019; Soldner et al., 2019), 这些方法通过分析口头和非口头线索(Granhag et al., 2015), 包括生理功能(Fienberg et al., 2003)等来揭示谎言迹象。Toma(Toma and Hancock, 2010)使用LIWC(Newman et al., 2003; Pennebaker et al., 2015)来识别欺诈性的线上聊天文件; Ott等人(Ott et al., 2011)采用了新的特征来识别具有欺骗性的观点垃圾信息。接着, (Petrov et al., 2006; Soldner et al., 2019)利用句法分析获得的特征来帮助鉴谎。与我们工作较为接近是(Hazra and Majumder, 2024), 其使用大语言模型从语言中发现说谎的线索(如过度自信, 模棱两可等), 再基于这些线索来判断是否说谎。

## 5.2 事实核查

事实核查分为三个关键步骤(Guo et al., 2022): 主张检测、证据检索和主张验证。主张检测是最初的阶段,旨在辨别需要验证的主张。证据检索需要从各种来源搜集支持或反驳的证据。主张验证采用二元分类,即将主张划分为真实或虚假(Potthast et al., 2017)。事实核查有多种应用,如检测虚假新闻(Chen et al., 2022; Li et al., 2021; Thorne et al., 2018; Wang, 2017)。

## 5.3 微调大模型在特定领域的应用

近年来,大语言模型展现出了惊人的文本理解和生成能力,并在多个自然语言处理任务中取得了优异的表现。为了提高大语言模型对不同领域的适配性,研究者将大语言模型在特定领域的的数据上进行微调。这使我们能够将大语言模型应用到广泛的场景任务中,包括信息抽取(InstructUIE(Wang et al., 2023)),写作助手(Writing-Alpaca(Zhang et al., 2023)),医疗咨询(ChatDoctor(Li et al., 2023)),数学算术(Goat(Liu and Low, 2023)),代码生成(WizardCoder(Luo et al., 2023))等。

## 6 结论

本文专注于研究在真实场景中用户与客服之间对话的谎言识别任务。为了进行更可靠的检测,提出了谎言指征和语义不一致线索的概念作为辅助信息。文章引入了一个谎言识别框架来训练LD-LLM,该框架利用谎言指征并生成不一致线索来确定用户是否在说谎。基于金融风控和身份识别场景中的两个谎言识别数据集FinLIE和IDLIE,文章构建了一个指令微调数据集,并使用指令微调数据集对LD-LLM进行微调。实验结果表明,谎言识别框架可以在真实场景的自动化谎言识别中达到最优的性能。

## 7 致谢

该论文得到了国家自然科学基金委(No.62176058)和科技部重点研发计划(2023YFF1204800)的项目经费支持。复旦大学CFFF平台为该项目提供了算力支持。

## 参考文献

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaye Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Stephen E Fienberg, JJ Blascovich, John T Cacioppo, RJ Davidson, P Ekman, DL Faigman, PL Grambsch, PB Imrey, EB Keeler, KB Laskey, et al. 2003. *The polygraph and lie detection*. New York: National Academy Press.
- Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21:303–340.
- Pär Anders Granhag, Aldert Vrij, and Bruno Verschuere. 2015. *Detecting deception: Current challenges and cognitive approaches*. John Wiley & Sons.

- David Gros, Yu Li, and Zhou Yu. 2021a. The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online, August. Association for Computational Linguistics.
- David Gros, Yu Li, and Zhou Yu. 2021b. The rua-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. *arXiv preprint arXiv:2106.02692*.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.
- Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).
- Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection. *arXiv preprint arXiv:2107.10747*.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct.
- Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.
- Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.
- Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.
- Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online, July. Association for Computational Linguistics.
- Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.

- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.
- Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Catalina L Toma and Jeffrey T Hancock. 2010. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 5–8.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.
- Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1):2056305120903408.
- Aldert Vrij, Maria Hartwig, and Pär Anders Granhag. 2019. Reading lies: Nonverbal communication and deception. *Annual review of psychology*, 70:295–317.
- Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction.
- William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance.
- 羊芙蕙. 2010. 谎言的识别研究. Ph.D. thesis, 华中科技大学.