使用代理模型评估资源较少语言中的 LLM 鲁棒性

Maciej Chrabszcz*
NASK - National Research Institute,
Warsaw, Poland
maciej.chrabaszcz@nask.pl

Katarzyna Lorenc *
NASK - National Research Institute,
Warsaw, Poland
katarzyna.lorenc@nask.pl

Karolina Seweryn*
NASK - National Research Institute,
Warsaw, Poland
karolina.seweryn@nask.pl

ABSTRACT

近年来，大型语言模型（LLMs）在各类自然语言处理（NLP）任务中展现了令人印象深刻的能力。然而，它们易受越狱和扰动的影响，因此需要进行额外的评估。许多 LLMs 是多语言的，但安全相关的训练数据主要包含英语等高资源语言，这可能使它们对波兰语等低资源语言中的扰动易感。我们展示了如何通过仅仅改变几个字符并使用小的代理模型进行词重要性计算来低成本地生成意外强大的攻击。我们发现，这些字符和词级别的攻击显著改变了不同 LLMs 的预测，表明存在潜在的漏洞，可以用来规避它们的内部安全机制。我们在低资源语言波兰语上验证了我们的攻击构建方法，并发现 LLMs 在这种语言中的潜在漏洞。此外，我们展示了如何将其扩展到其他语言。我们发布了创建的数据集和代码以供进一步研究。

1 介绍

语言模型（LMs）[1, 2] 在自然语言理解（NLU）和自然语言生成（NLG）任务中表现出色，支持许多日常应用。然而，最近的研究 [3, 4, 5, 6, 7] 显示它们易受到模拟人类输入扰动的攻击。因此，测试这些模型对扰动的鲁棒性是至关重要的。

LLMs 研究主要针对高资源语言，例如英语 [2, 8] 。然而，多语言模型 [9, 10, 11] 的出现引入了新漏洞，特别是在包含低资源语言时。由于数据有限，这些语言形成挑战，使得在微调过程中（如监督微调（SFT）和模型对齐）即使是简单的扰动也难以提高其多语言鲁棒性。因此，评估多语言模型对于低资源语言的鲁棒性至关重要。

为解决评估多语言模型鲁棒性的问题，我们提出了一个通过利用代理模型和归因方法来生成扰动数据集的框架。这些数据集可用于评估大语言模型在鲁棒性方面的安全性，使开发人员能够检查模型对特定扰动的鲁棒性，并在发布后降低因简单扰动而导致模型误导的风险。我们的框架利用代理模型，结合归因方法，能够低成本识别特定任务中最重要的词语，从而通过仅扰动最重要的词语来创建评估示例。我们在波兰语上验证了我们的方法，并发现了该语言中大语言模型的潜在漏洞。此方法可以轻松地适应其他语言，所需的语言学努力最小。通过对重要词语进行有针对性的扰动，我们可以严格测试并确保各种语言模型对感兴趣的扰动的鲁棒性。

我们的贡献如下：

- 我们引入了一个框架，该框架生成易于人类理解的扰动示例，以评估 LLMs 对扰动的鲁棒性。
- 我们整理了波兰语数据集，根据它们训练代理模型，并进行扰动，从而创建了该波兰语数据集，可用于评估大型语言模型在波兰语中的鲁棒性。
- 我们利用创建的数据集对 LLMs 的稳健性进行广泛评估。我们识别出导致性能最严重下降的扰动。从这一分析中获得的见解可以帮助模型开发者提高其模型的稳健性。

*Equal contribution.

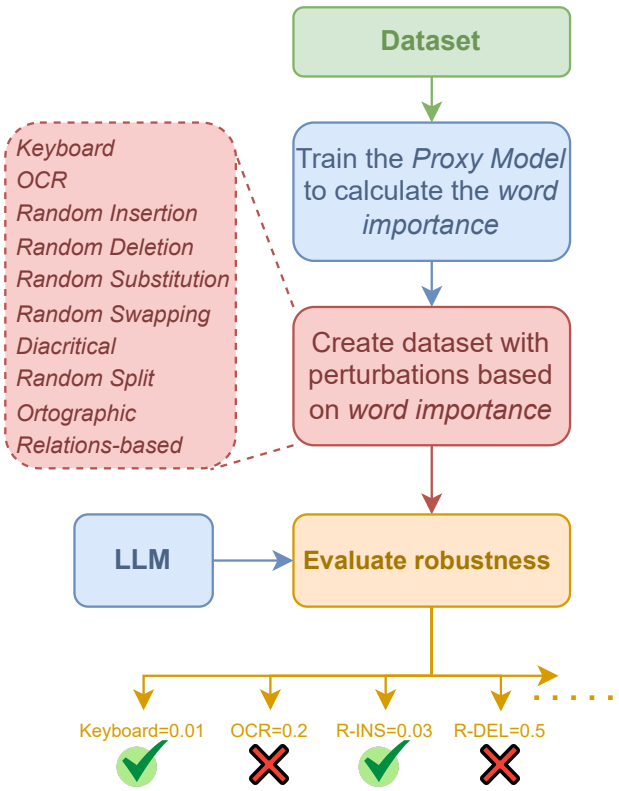

Figure 1: 所提出框架的概述。我们可以使用在目标数据集上训练的代理模型来计算词语的重要性。接下来，借助词语重要性，我们用感兴趣的修改来扰动最重要的词语。然后，我们可以评估大规模语言模型在这些扰动下的鲁棒性。通过设置阈值，我们可以创建自动鲁棒性检查，以突出模型开发过程中存在的问题。

2 相关工作

2.1 安全性和鲁棒性

确保 AI 模型的鲁棒性对于维护模型的安全性至关重要。最近的研究表明，语言模型可能易于受到扰动影响，导致错误的类别预测 [3, 12, 6] 或产生不良的文本，即使模型已很好地对齐 [13] 。这些发现强调了评估模型对扰动的鲁棒性的重要性。

在生成扰动示例时，保持文本的原意是至关重要的，这可以通过添加拼写错误 [12] 、同义词 [3] 、基于 BERT 的替换 [4] 来实现。因此，人们一直致力于开发对抗性数据集，如用于英语数据的 AdvGLUE [14] ，并已在 DecodingTrust 框架中扩展用于 LLM。然而，据我们所知，对于波兰语和许多其他资源较少的语言，目前并没有可用的、可比的数据集。

2.2 归因方法

归因方法旨在识别对模型预测最有影响的输入部分。尽管先前关于攻击的研究 [12, 4] 通过观察当一个词被省略时预测的变化或简单的显著性归因 [15, 12] 来计算词的重要性，但这些方法可能会有问题，因为这些方法不忠实于模型。

提出了替代归因方法，以在黑盒和白盒场景中提供更强大的归因。在黑盒场景中，由于模型的内部机制无法访问，方法如 SHAP（Shapley 加性解释） [16] 和 LIME（本地可解释的模型无关解释） [17] 有助于确定词语的重要性。

在白盒场景中，可以访问模型的内部结构，基于梯度的归因方法获得了重要性。值得注意的例子包括 Grad x Input [18] 、Integrated Gradients [19] 和 SmoothGrad [20] 。这些方法利用模型的梯度信息来量化每个输入特征对最终预测的贡献。

Table 1: 例子说明了成功改变模型预测的不同扰动。

Modification	Dataset	Sample (~~Strikethrough~~ = Original Text, 红色 = Perturbation)	Label → Prediction
OCR	AR	PL : ~~saba~~ sa6a jako, typowa podróbka, nie polecam EN : Poor quality, typical fake, I do not recommend.	1 → 2
Rel	AC	PL : klient akceptuje ~~niniejszy~~ 如下 regulamin EN : client accepts these terms and conditions.	0 → 1
R-Sw	AC	PL : Owiadczam, e otrzymaem, ~~zapoznaem~~ zapzoname si i akceptuj EN : I declare that I have received, read, and accept.	1 → 0
Split	P-O	PL : dziki ~~niemu~~ 倪穆 jeszcze jestem studentem ; p EN : thanks to him, I'm still a student ;p	plus m → minus s
Key	P-O	PL : UNIKAC nie ~~polecam~~ poleFXm . . . brak slow (tz sa ale post mi zlikwiduja) EN : AVOID, I don't recommend... I'm at a loss for words (well, I have words, but they'll delete my post).	minus m → zero
R-Del	AC	PL : sd ~~waciwy~~ （无法翻译的文本） dla siedziby powoda. EN : The court competent for the plaintiff	0 → 1
R-Sub	CBD	PL : @anonymized_account Bierz tego @anonymized_account razem ~~jestecie~~ jestewZe mocni EN : @anonymized_account take @anonymized_account together, you are strong.	0 → 1
R-Ins	P-O	PL : Krotko : cala grupa zaliczyla , nie oddal nikomu ani jednego sprawka . Oceny ~~marzenie~~ mqarzenJie . Tyle : -) EN : Shortly: the whole group passed, didn't hand in a single assignment to anyone. Dream grades, just a dream. That's it : -)	plus m → zero
Ort	AC	PL : Za ewentualne zniszczenia odpowiada ~~opiekun~~ opiekón EN : The supervisor is responsible for any potential damage.	0 → 1
Dia	AC	PL : Po opuszczeniu parkingu firma reklamacji nie ~~uwzgldnia~~ 考虑. EN : After leaving the parking lot, the company does not accept any complaints.	0 → 1

在自然语言处理领域中，基于 Transformer 的模型的广泛应用推动了专门为这种架构设计的归因方法的发展。这些方法包括 Attention Rollout [21] 和其他基于注意力的技术 [22, 23] 。

2.3 语言建模

语言建模是自然语言处理中的一项基本任务，涉及预测序列中词或标记的概率分布。这些模型已经经历了显著的发展，从简单的统计方法如 n-gram 模型演变到如今主导该领域的更先进的神经网络架构。基于 Transformer 的语言模型如 GPT [24] 和 BERT [1] 在各种任务中取得了最先进的成果，包括文本分类和生成。

在波兰，已经开发了几个基于 transformer 的模型 [25, 26, 27] 。其中，Bielik [28] 目前是最为突出的波兰专用生成模型，尽管一些多语言模型也提供了对波兰语的支持（LLama3.1 [2] ，OpenChat [29] ，CommandR [10] ）。

LLM 的广泛应用凸显了评估其稳健性的关键需求，因为这些模型可能会被攻击以生成有害或误导性的输出。这可能会产生严重后果，特别是在医疗保健或金融等关键领域。因此，必须仔细评估这些模型，以确保其可靠性和安全性，并在它们在现实场景中被利用之前识别出潜在的漏洞 [30, 14, 31] 。

为了识别最重要的词语，我们可以手动对这些词语进行排名，但手动选择最重要词语不可行。因此，我们使用归因方法来选择最重要的词语。大多数归因方法需要一个模型来计算重要性。预训练的 LLM 可作为这样的模型使用。不幸的是，这些模型本身占用大量 VRAM，并且计算基于梯度的方法对于 LLM 而言可能非常耗时，对于某些方法而言，由于 VRAM 限制甚至是不可能的。

我们训练一个小的代理模型来解决这个问题，之后用于计算单词的重要性。如果该模型在数据集上表现良好，可以认为它是计算单词重要性的良好代理。由于归属方法突出显示了对模型预测重要的内容，在这个步骤中使用高性能的模型是至关重要的。

2.4 词语重要性

为了识别单个单词的重要性，我们首先使用基于梯度和扰动的方法计算标记归属。例如，X 的输入序列和 y 的输出的 Grad x Input 归属计算如下

$$\text{Grad x Input}(X, y) = \nabla_y(X_{emb}) \cdot X_{emb}, \tag{1}$$

其中 $X_{emb} \in \mathcal{R}^{n,d}$ 是 n 输入标记的词嵌入，d 是嵌入大小。

然而，由于分词过程的特性，词元未必总是对应完整的单词。为了解决这个问题，我们将构成每个单词的词元分组并使用简单的均值来汇总它们的归因，如下述重要性方程所示（$\mathcal{I}$ ）：

$$\mathcal{I}(W) = \frac{1}{|S(W)|} \sum_{t \in S(W)} a(t), \tag{2}$$

其中 S 是一个函数，用于将单词 W 拆分为其子词元，$|S(W)|$ 表示单词中的子词元数量，$a(t)$ 代表从选定的归因方法获得的子词元 t 的归因值。此方法使我们能够独立于分词过程来确定每个单词的重要性，从而提供一种更具可解释性和连贯性的单词级别重要性衡量方法。

我们排除代表标点符号的标记，以确保更准确地聚合词的重要性。如果不排除这些标记，我们在计算词的重要性时还会考虑标点的作用。这可能导致词的重要性得分与文本的语义内容不一致。通过关注文本的语义内容而非句法结构，我们能够更清楚地了解单词对预测的重要性。

我们进行的扰动被分为两个不同的层次：字符层和词语层。

在字符级别上，我们探索了多种设计用来引入印刷错误的技术。具体的扰动包括：

- 键盘错误（键）：模拟由相邻按键导致的常见输入错误。
- 光学字符识别（OCR）错误：由于字符之间的图形相似性引入的错误，这在 OCR 系统中通常会观察到。
- 随机字符插入（R-Ins）：在单词中插入随机字符。
- 字符删除（R-Del）：从单词中移除字符。
- 字符替换（R-Sub）：用随机字符替换字符。
- 字符互换 (R-Sw)：重新排序相邻字符。
- 变音错误（Dia）：省略变音符号，这可能会极大地改变词语的含义。例如，将 “kt”（角）改为 “kat”（刽子手）或将 “jzyk”（语言）改为 “jeyk”（小刺猬）展示了此类错误如何显著影响诸如波兰语等语言中的词语解释。

在其他语言中实施字符级扰动需要更新词典，以包含该语言常见的音标错误。其他扰动则不需要更改。在词汇级别，我们应用了几种扰动方法来评估它们对模型性能的影响：为了将词汇级扰动扩展到其他语言，我们必须更新特定于目标语言的正字错误。此外，基于词汇关系进行更改需要访问该语言的词汇关系网络。

需要注意，并非所有的扰动都能轻易转移到其他语言。我们认为许多特定语言的扰动在波兰语中可能不起作用。因此，谨慎地在我们的框架中添加和删除扰动对于使用特定语言的变化来准确评估模型至关重要。

Table 2: 扰动方法对代理模型的攻击成功率 (ASR)，涵盖所有数据集，并考虑归因方法和更改词语数量的影响。

	Data	Diac	Key	OCR	Ort	R-Del	R-Ins	R-Sub	R-Sw	Rel	Split	Avg
PolBERT	AC	0.01	0.11	0.13	0.02	0.08	0.11	0.11	0.09	0.02	0.07	0.08
	AR	0.03	0.22	0.23	0.06	0.24	0.23	0.23	0.24	0.12	0.24	0.18
	CBD	0.01	0.04	0.04	0.01	0.04	0.03	0.03	0.05	0.01	0.05	0.03
	P-I	0.00	0.04	0.04	0.01	0.04	0.04	0.05	0.04	0.02	0.04	0.03
	P-O	0.01	0.16	0.19	0.02	0.14	0.15	0.16	0.14	0.04	0.15	0.12
HerBERT	AC	0.01	0.11	0.12	0.02	0.04	0.09	0.11	0.05	0.02	0.05	0.06
	AR	0.02	0.14	0.14	0.04	0.12	0.12	0.14	0.12	0.07	0.12	0.10
	CBD	-	-	-	-	-	-	-	-	-	-	-
	P-I	0.00	0.03	0.03	0.00	0.02	0.03	0.03	0.02	0.01	0.02	0.02
	P-O	0.00	0.13	0.12	0.02	0.11	0.11	0.13	0.12	0.04	0.09	0.09
RoBERTa	AC	0.04	0.21	0.24	0.06	0.16	0.20	0.20	0.17	0.05	0.14	0.15
	AR	0.03	0.22	0.23	0.05	0.21	0.21	0.22	0.21	0.10	0.20	0.17
	CBD	0.00	0.02	0.02	0.01	0.02	0.02	0.02	0.03	0.01	0.03	0.02
	P-I	0.00	0.04	0.05	0.01	0.04	0.04	0.04	0.05	0.02	0.04	0.03
	P-O	0.01	0.12	0.13	0.03	0.12	0.11	0.12	0.13	0.04	0.11	0.09

Table 3: 代理模型在所有数据集测试集上的表现。

Model	Data	AUROC	F1	ACC
PolBERT	AC	0.922	0.828	0.843
	AR	0.836	0.480	0.580
	CBD	0.837	0.684	0.893
	P-I	0.969	0.816	0.856
	P-O	0.896	0.537	0.678
HerBERT	AC	0.926	0.836	0.851
	AR	0.887	0.572	0.653
	CBD	0.547	0.464	0.866
	P-I	0.970	0.856	0.892
	P-O	0.910	0.529	0.711
RoBERTa	AC	0.926	0.832	0.848
	AR	0.863	0.542	0.624
	CBD	0.879	0.660	0.887
	P-I	0.972	0.860	0.886
	P-O	0.904	0.529	0.715

3 实验

3.1 数据集

在实验中，我们使用了来自 KLEJ 基准的数据集 [32] 。KLEJ 是波兰语中对文本分析相当于英语中的 GLUE 基准。

AC - Polish Abusive Clauses Dataset [33] 用于检测法律协议中的不当条款，以保护消费者免受不公平条款的影响。它包含两个类别：不当条款和正确的协议陈述。

AR - Allegro Reviews [32] 包含来自波兰电子商务平台 Allegro 的情感标注评论。每个评论都被分配一个值，从 1 到 5，表示情感得分。

CBD - 网络欺凌检测 [34] 数据集包含推特消息，用于预测给定消息是否包含网络欺凌或有害内容。

P-I & P-O - PolEmo2.0 [35] 是来自医药和酒店领域的在线评论集合。任务是预测评论的情感。两个独立的测试集可以进行域内（医药和酒店）和域外（产品和大学）评估。

这些数据集提供了多样化的自然语言处理任务，包括情感分析、辱骂内容检测和特定领域的文本分类，这对于评估模型在各种波兰语言理解任务中的性能至关重要。有关训练和测试划分大小的详细信息可在附录 A 中找到。

3.2 模型

在我们的实验中，我们使用了基于 Transformer 的分类器 HerBERT [25] 、PolBERT [26] 和 Polish RoBERTa [27] 作为代理模型。这些模型在每个分析数据集的训练集上针对特定任务进行了微调。训练过程在 A100（40GB）GPU 上进行了 20 个周期，采用早停策略，批量大小设置为 16（Allegro 评论为 8），模型性能见表 3 。训练后，这些模型还通过测试集对其抗扰动鲁棒性进行评估。我们观察到 HerBERT 模型在 CBD 数据集上训练失败，导致它为所有样本预测相同的标签。因此，我们在后续实验中不再使用该模型。

对于大语言模型（LLMs），我们使用了能够处理波兰语的开源生成模型，例如 Bielik [28] 、Mistral-7B-Instruct [36] 和 Llama-3.1-8 B [2] 。我们对这些模型采用了一种零样本学习方法，其中模型在被查询时使用了原始和修改后的提示。

作为词语重要性的最终代理模型，我们使用了基于 polish-roberta-base-v2 分类器的模型，该模型在所有数据集上表现最佳，并且使用 SHAP 方法作为归因方法，因为 SHAP 在代理模型扰动成功率中表现最高。

为了评估扰动的有效性，我们使用了攻击成功率（Attack Success Rate, ASR）指标。该指标作为评估对抗攻击性能的标准度量，通过计算攻击达到预期目标的情况比例来实现。较高的 ASR 表示攻击更有效，而较低的 ASR 则表明模型对对抗性操纵具有更大的抵抗力。形式上，ASR 可以定义为：

$$ASR = \sum_{(x,y)\in\mathcal{D}} \frac{\mathbb{1}[f(\mathcal{A}(x)) \neq y] \cdot \mathbb{1}[f(x) = y]}{\sum_{(x',y')\in\mathcal{D}} \mathbb{1}[f(x') = y']}, \tag{3}$$

其中，$\mathcal{D}$ 表示输入样本 x 及其对应的真实标签 y 的数据集，$f(x)$ 代表模型给定输入 x 的预测，$\mathcal{A}(x)$ 是由扰动 $\mathcal{A}$ 生成的输入 x 的对抗性扰动版本，而 $\mathbb{1}[]$ 是指示器。

4 结果

4.1 扰动类型的影响

表格 2 的结果突出显示了我们的语言模型在面临简单字符扰动时，如 OCR、R-Del、R-Ins、R-sub、R-Sw 和 Split，被愚弄改变其预测。这些模型在训练数据中没有见过这样的错误，有可能在遇到这些扰动的例子时表现不佳。

在表格 4 中，我们可以观察到，即使 LLM 在互联网上的数据中进行训练，应该有包含字符级错误的例子，那些模型仍然容易被这样的扰动误导。这应引起开发者的注意，以修复此问题。这些是简单的分类任务，但如果模型在面对这些任务中的扰动时出错更多，那么这些类型的扰动可能被用来欺骗模型生成有害内容的风险很高。

Table 4: 所有数据集上，针对较小模型的扰动方法的 ASR，汇总了改变词数。扰动是通过使用 SHAP 词重要性分数和 RoBERTa 分类器产生的。超过阈值的稳健性值用红色突出显示，红色越明显表示稳健性越低。低于阈值的值用绿色突出显示，绿色越明显表示稳健性越高。

	Data	Diac	Key	OCR	Ort	R-Del	R-Ins	R-Sub	R-Sw	Rel	Split
Bielik v1	AC	0.145	0.285	0.267	0.170	0.261	0.261	0.288	0.283	0.182	0.224
	AR	0.082	0.196	0.200	0.083	0.176	0.170	0.191	0.173	0.145	0.151
	CBD	0.087	0.452	0.488	0.156	0.318	0.441	0.477	0.360	0.173	0.383
	P-I	0.134	0.302	0.308	0.167	0.264	0.263	0.286	0.265	0.242	0.273
	P-O	0.149	0.432	0.417	0.174	0.380	0.427	0.414	0.389	0.307	0.361
Mistral 7B	AC	0.010	0.076	0.087	0.014	0.041	0.068	0.076	0.056	0.022	0.042
	AR	0.027	0.175	0.202	0.048	0.156	0.141	0.177	0.156	0.103	0.146
	CBD	0.013	0.348	0.368	0.029	0.156	0.335	0.355	0.213	0.056	0.214
	P-I	0.003	0.028	0.029	0.010	0.027	0.032	0.032	0.026	0.016	0.025
	P-O	0.015	0.075	0.073	0.024	0.074	0.058	0.078	0.061	0.034	0.065
Llama3 8B	AC	0.219	0.333	0.334	0.241	0.314	0.332	0.337	0.325	0.239	0.302
	AR	0.075	0.346	0.346	0.166	0.345	0.354	0.329	0.296	0.221	0.261
	CBD	0.041	0.148	0.232	0.058	0.102	0.159	0.157	0.127	0.062	0.093
	P-I	0.004	0.051	0.044	0.005	0.035	0.043	0.052	0.037	0.013	0.018
	P-O	0.031	0.137	0.146	0.042	0.111	0.131	0.136	0.118	0.070	0.098

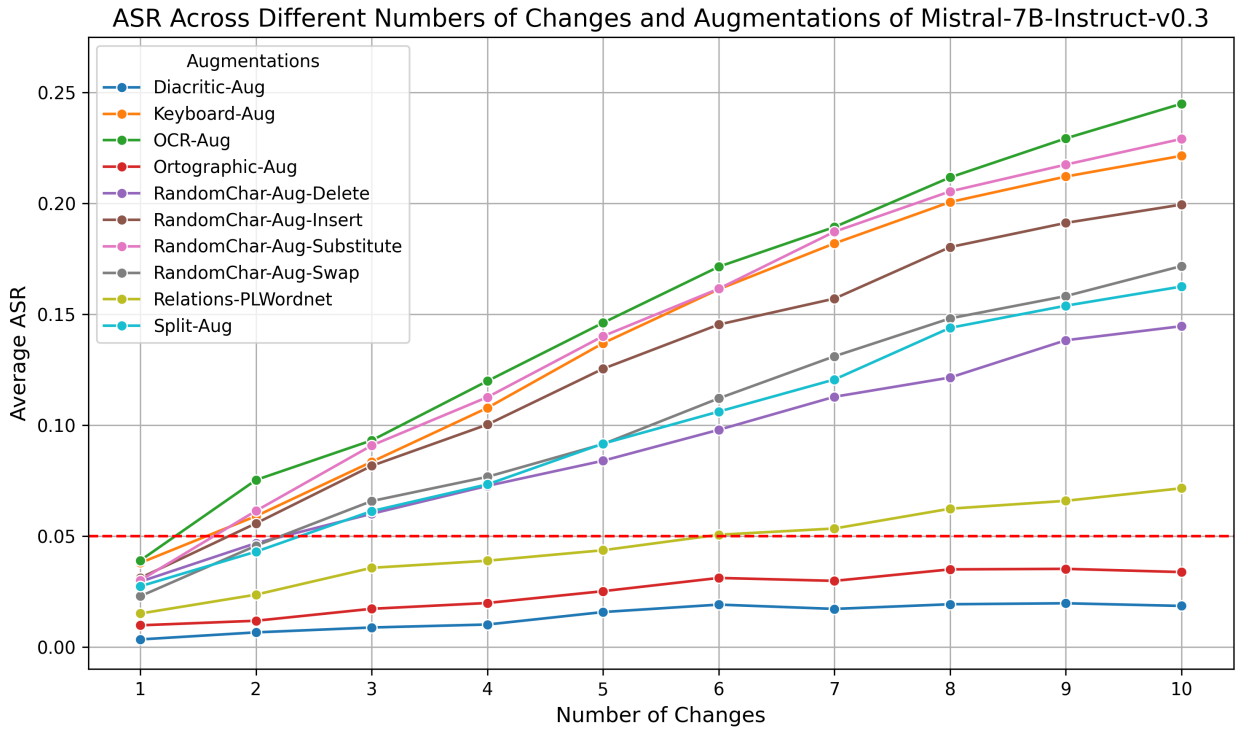

Figure 2: Mistral 上的 ASR 与不同归因方法的扰动词数之间的关系。虚线表示模型被认为对简单扰动不够稳健的稳健性临界值。

4.2 归因方法的影响

根据图 **??** 的结果，我们可以观察到对于所有更改的单词数量，SHAP 与其他方法相比给出了最高的 ASR。Vanilla Gradient 和 SmoothGrad 在 ASR 方面表现最差，其表现与选择随机单词非常相似。这表明在我们的情况下，当选择应扰动的单词时，使用 SHAP 提取单词重要性是有益的。

4.3 大型语言模型的鲁棒性

图 2 和表 4 展示了 Mistral 模型建议框架的一个示例用法。红线表示相当于模型在 5% 的示例中被这些扰动欺骗的阈值（0.05 ），我们决定当模型如此频繁地被欺骗时，这可能表明它在与此类扰动相关方面具有脆弱性。

分析表明，虽然模型可以有效处理符号和正字法的扰动，但它容易受到其他修改的影响，例如在词汇层面插入额外的字符。这个发现从人工智能安全的角度来看尤为重要，因为这表明虽然模型可能在其标准形式上抵抗有害的提示，但细微的改动，例如在关键术语中插入空格，可能会增加生成有害响应的可能性。

5 结论

在这项工作中，我们通过利用代理模型计算的扰动和词重要性引入了一个评估大型语言模型（LLM）稳健性的框架。我们通过精心制作一组可用于评估 LLM 在波兰语中稳健性的扰动数据集，验证了此框架。

为了评估对扰动的鲁棒性，我们执行以下步骤：

1. 创建代理模型：在目标数据集上训练一个小型模型。
2. 排名：基于代理模型计算和聚合归因得分，以创建词语重要性排名。
3. 扰动：根据排名扰动最重要的词。
4. 评估：在原始和扰动后的数据集上评估目标大语言模型，以评估它们的稳健性。

在假定属性计算是可行的情况下，与各种扰动方法和任务的兼容性是这个框架的一个关键特性。使用通过我们的框架准备的扰动数据集在 LLM 开发中尤为有利。通过仔细准备具有特定扰动的数据集，开发者可以有效地评估和控制他们所创建模型的鲁棒性。如果某个扰动导致 LLM 在分类任务上失败，这就突显出一个需要关注的脆弱性，因为这样的扰动有可能对在安全关键生成任务中（如生成被禁止活动的指令）对模型性能产生负面影响。

我们使用波兰语言框架进行的研究表明，这种可以用最少资源实施的扰动在欺骗大型语言模型（LLM）方面意外地有效。我们重点强调了那些在愚弄这些模型中最成功的扰动，并指出了需要关注的领域。

6 局限性

我们的词语重要性依赖于归因方法，这些方法可能无法真实地反映模型的内部推理过程。另一方面，用于生成归因结果的模型的选择也可能影响所选择的词语是否真正重要，这些词语是反映了真正的语义重要性，还是仅仅是代理模型训练中特有的伪影。

该分析依赖于从词汇重要性排名中得出的启发性扰动。这种方法并不包括针对大型语言模型的所有潜在攻击，例如通过对抗性优化技术生成的攻击。

另一个需要考虑的因素是，扰动重要词语的过程可能会使文本的关键部分变得难以理解，即便是从人类的角度来看也是如此。这种语义退化可能会混淆评估，使将模型的鲁棒性问题与输入被破坏的影响分离开来变得具有挑战性。

此外，依赖代理模型生成词语重要性评分是一种近似。直接从目标 LLM 得出的重要性排名可能会有所不同，并可能为识别那些其扰动对 LLM 性能产生最大影响的词语提供更精确的依据。

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. Abhishek Kadian. The llama 3 herd of models, 2024.

[3] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI Conference on Artificial Intelligence*, 2019.

[4] Linyang Li, Ruotian Ma, Qipeng Guo, X. Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *ArXiv*, abs/2004.09984, 2020.

[5] Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6134–6150, Online, November 2020. Association for Computational Linguistics.

[6] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online, July 2020. Association for Computational Linguistics.

[7] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 80079–80110. Curran Associates, Inc., 2023.

[8] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[10] CohereForAI. Command r +, 2024. Accessed: 2024-08-29.

[11] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[12] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society, 2019.

[13] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[14] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems*, 2021.

[15] K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014.

[16] M Scott, Lee Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" : Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[18] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *ArXiv*, abs/1605.01713, 2016.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[20] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825, 2017.

[21] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[22] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, 2021.

[23] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

[24] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[25] Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine, April 2021. Association for Computational Linguistics.

[26] Dariusz Keczek. Polbert: Attacking polish nlp tasks with transformers. In Maciej Ogrodniczuk and ukasz Kobyliski, editors, *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, 2020.

[27] Sawomir Dadas and Magorzata Grbowiec. Assessing generalization capability of text ranking models in polish, 2024. arXiv:2402.14318 [cs.CL].

[28] Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. Bielik 7b v0. 1: A polish language model–development, insights, and evaluation. *arXiv preprint arXiv:2410.18565*, 2024.

[29] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

[30] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

[31] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.

[32] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. Klej: Comprehensive benchmark for polish language understanding. *arXiv preprint arXiv:2005.00630*, 2020.

[33] Lukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Arkadiusz Janz, Piotr Szymański, Marcin Wątroba, Mikoł aj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. This is the way: designing and compiling lepiszcze, a comprehensive nlp benchmark for polish. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21805–21818. Curran Associates, Inc., 2022.

[34] Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval 2019 Workshop*, page 89, 2019.

[35] Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China, November 2019. Association for Computational Linguistics.

[36] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

A 模型和数据集

关于数据集大小的信息可在表格 5 中查阅，其中包括 LLM 在原始测试集上的表现。

Table 5: 使用数据集的训练和测试集划分大小。

Dataset	Train	Test
Polish Abusive Clauses	4284	3453
Allegro Reviews	9577	1006
Cyberbullying	10041	1000
PolEmo2.0 In	5783	722
PolEmo2.0 Out	5783	494

所有的字符级扰动随机选择要更改的字符数量在 1 和 $\min(len(word) \cdot 0.15, 4)$ 之间。它确保字符级的更改使扰动后的单词对于人类来说容易理解。单词级扰动总是尝试修改提供的单词，但某些方法可能会失败。为了计算 SmoothGrad 和 Integrated Gradients 的归因，我们使用了 50 个步骤。对于 SHAP 归因，我们使用 SHAP [16] 库提供的默认参数。

B 扰动攻击成功率

Table 6: 归因方法针对所有数据集的小模型的 ASR。

Model	Data	AGR	AR	G	IG	SH	SG	Avg
PolBERT	AC	0.08	0.09	0.07	0.07	0.08	0.07	0.08
	AR	0.28	0.16	0.12	0.20	0.22	0.12	0.18
	CBD	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	P-I	0.07	0.03	0.02	0.03	0.05	0.01	0.04
	P-O	0.18	0.12	0.07	0.11	0.15	0.07	0.12
HerBERT	AC	0.06	0.09	0.05	0.05	0.07	0.05	0.06
	AR	0.11	0.09	0.08	0.10	0.14	0.08	0.10
	CBD	-	-	-	-	-	-	-
	P-I	0.03	0.02	0.02	0.02	0.03	0.02	0.02
	P-O	0.12	0.08	0.06	0.10	0.11	0.07	0.09
RoBERTa	AC	0.15	0.15	0.13	0.16	0.15	0.13	0.14
	AR	0.21	0.14	0.10	0.21	0.25	0.10	0.17
	CBD	0.02	0.02	0.01	0.02	0.03	0.01	0.02
	P-I	0.04	0.02	0.02	0.04	0.06	0.02	0.03
	P-O	0.11	0.07	0.06	0.11	0.14	0.07	0.09

ASR 分解为归因方法的结果在表格 6 中给出。

表格 7 、8 、9 、10 、11 和 12 展示了按模型、增强方法和分类器中更改的单词数量划分的 ASR。

图 3 表明，与在攻击过程中选择随机词相比，使用归因方法（有针对性攻击）显著提高了对语言模型的攻击效果。

图 4 , 5 显示了对 Llama 和 Bielik 的稳健性检验。

Table 7: 使用 SHAP 作为归因方法时，不同数量单词的 ASR 在数据集上进行汇总分析。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.004	0.006	0.007	0.010	0.012	0.013	0.014	0.015	0.019	0.022
	Key	0.057	0.091	0.110	0.124	0.139	0.152	0.164	0.173	0.180	0.188
	OCR	0.066	0.101	0.125	0.142	0.159	0.170	0.189	0.195	0.204	0.205
	Ort	0.010	0.015	0.020	0.027	0.030	0.035	0.039	0.042	0.047	0.051
	R-Del	0.052	0.087	0.107	0.123	0.135	0.146	0.157	0.166	0.172	0.179
	R-Ins	0.058	0.091	0.115	0.126	0.138	0.146	0.160	0.169	0.177	0.186
	R-Sub	0.063	0.093	0.117	0.132	0.145	0.156	0.168	0.178	0.181	0.187
	R-Sw	0.056	0.088	0.112	0.127	0.145	0.154	0.164	0.176	0.184	0.191
	Rel	0.023	0.033	0.041	0.046	0.049	0.054	0.062	0.065	0.071	0.074
	Split	0.053	0.083	0.107	0.123	0.134	0.148	0.159	0.167	0.175	0.179
HerBERT	Diac	0.003	0.004	0.005	0.006	0.007	0.008	0.011	0.012	0.013	0.016
	Key	0.033	0.056	0.078	0.093	0.111	0.120	0.134	0.145	0.153	0.163
	OCR	0.033	0.053	0.075	0.090	0.101	0.117	0.126	0.133	0.141	0.149
	Ort	0.006	0.010	0.012	0.015	0.015	0.016	0.021	0.024	0.025	0.028
	R-Del	0.025	0.042	0.055	0.068	0.074	0.085	0.090	0.103	0.112	0.121
	R-Ins	0.027	0.049	0.064	0.075	0.086	0.100	0.111	0.116	0.124	0.128
	R-Sub	0.034	0.056	0.078	0.096	0.108	0.121	0.130	0.144	0.152	0.165
	R-Sw	0.027	0.045	0.061	0.075	0.083	0.092	0.100	0.109	0.117	0.123
	Rel	0.011	0.018	0.023	0.028	0.034	0.039	0.043	0.050	0.055	0.054
	Split	0.023	0.038	0.052	0.061	0.071	0.074	0.080	0.088	0.093	0.098
RoBERTa	Diac	0.007	0.008	0.010	0.011	0.013	0.018	0.022	0.026	0.032	0.037
	Key	0.053	0.096	0.127	0.149	0.168	0.193	0.211	0.225	0.238	0.249
	OCR	0.058	0.107	0.140	0.167	0.182	0.201	0.213	0.228	0.242	0.253
	Ort	0.013	0.017	0.023	0.026	0.033	0.042	0.049	0.057	0.061	0.067
	R-Del	0.045	0.082	0.111	0.132	0.156	0.180	0.199	0.213	0.229	0.241
	R-Ins	0.043	0.085	0.116	0.139	0.155	0.175	0.188	0.208	0.223	0.232
	R-Sub	0.050	0.094	0.122	0.146	0.167	0.186	0.203	0.220	0.231	0.246
	R-Sw	0.048	0.093	0.123	0.144	0.161	0.185	0.200	0.219	0.239	0.245
	Rel	0.021	0.035	0.044	0.051	0.056	0.067	0.072	0.080	0.091	0.098
	Split	0.045	0.081	0.104	0.125	0.136	0.158	0.174	0.191	0.207	0.218

Table 8: 使用梯度作为归因方法时，针对不同数量的单词的 ASR，汇总于数据集。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.005	0.007	0.008	0.009	0.010	0.011	0.013	0.015	0.018	0.021
	Key	0.031	0.046	0.060	0.073	0.079	0.083	0.089	0.096	0.102	0.106
	OCR	0.035	0.054	0.069	0.082	0.087	0.094	0.103	0.112	0.119	0.125
	Ort	0.007	0.011	0.013	0.016	0.020	0.024	0.029	0.032	0.034	0.038
	R-Del	0.027	0.040	0.054	0.064	0.072	0.080	0.084	0.090	0.096	0.099
	R-Ins	0.033	0.046	0.058	0.067	0.075	0.080	0.090	0.092	0.100	0.107
	R-Sub	0.032	0.046	0.060	0.071	0.078	0.084	0.090	0.095	0.103	0.107
	R-Sw	0.033	0.050	0.058	0.073	0.079	0.085	0.090	0.096	0.107	0.110
	Rel	0.014	0.022	0.027	0.029	0.034	0.036	0.040	0.043	0.048	0.052
	Split	0.031	0.048	0.057	0.070	0.075	0.082	0.087	0.092	0.095	0.100
HerBERT	Diac	0.003	0.004	0.005	0.007	0.007	0.008	0.009	0.010	0.012	0.014
	Key	0.027	0.036	0.043	0.054	0.063	0.070	0.071	0.076	0.081	0.088
	OCR	0.026	0.039	0.047	0.057	0.061	0.068	0.072	0.078	0.084	0.087
	Ort	0.006	0.008	0.011	0.012	0.013	0.013	0.015	0.017	0.021	0.023
	R-Del	0.018	0.030	0.033	0.037	0.044	0.047	0.048	0.054	0.060	0.065
	R-Ins	0.022	0.035	0.040	0.048	0.053	0.062	0.065	0.068	0.072	0.077
	R-Sub	0.028	0.041	0.046	0.053	0.057	0.067	0.072	0.081	0.085	0.093
	R-Sw	0.020	0.031	0.035	0.045	0.048	0.052	0.053	0.054	0.059	0.066
	Rel	0.011	0.015	0.017	0.021	0.024	0.026	0.029	0.033	0.038	0.040
	Split	0.020	0.032	0.035	0.038	0.041	0.044	0.049	0.050	0.057	0.059
RoBERTa	Diac	0.005	0.006	0.008	0.009	0.011	0.015	0.020	0.026	0.031	0.036
	Key	0.029	0.044	0.062	0.075	0.084	0.096	0.108	0.119	0.125	0.135
	OCR	0.036	0.056	0.073	0.085	0.095	0.109	0.119	0.125	0.136	0.143
	Ort	0.006	0.009	0.011	0.013	0.017	0.023	0.029	0.034	0.042	0.048
	R-Del	0.026	0.038	0.050	0.062	0.072	0.085	0.097	0.106	0.113	0.124
	R-Ins	0.027	0.041	0.055	0.065	0.073	0.091	0.100	0.111	0.120	0.129
	R-Sub	0.030	0.043	0.058	0.070	0.078	0.094	0.104	0.116	0.125	0.134
	R-Sw	0.030	0.041	0.055	0.067	0.076	0.091	0.102	0.112	0.123	0.132
	Rel	0.009	0.011	0.016	0.020	0.023	0.031	0.037	0.043	0.051	0.058
	Split	0.025	0.037	0.049	0.059	0.066	0.078	0.086	0.096	0.102	0.110

Table 9: 对于不同数量的单词更改的 ASR 在使用 SmoothGrad 作为归因方法时，汇总在各个数据集上。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.004	0.006	0.007	0.008	0.008	0.010	0.013	0.015	0.019	0.021
	Key	0.030	0.046	0.060	0.073	0.078	0.083	0.088	0.096	0.102	0.111
	OCR	0.035	0.054	0.066	0.080	0.090	0.096	0.105	0.112	0.121	0.126
	Ort	0.008	0.010	0.012	0.015	0.019	0.023	0.028	0.030	0.033	0.037
	R-Del	0.029	0.041	0.052	0.066	0.074	0.078	0.086	0.093	0.097	0.101
	R-Ins	0.033	0.048	0.058	0.072	0.079	0.088	0.094	0.098	0.102	0.108
	R-Sub	0.033	0.048	0.062	0.072	0.075	0.081	0.085	0.093	0.100	0.108
	R-Sw	0.032	0.048	0.054	0.067	0.074	0.081	0.088	0.093	0.102	0.107
	Rel	0.016	0.021	0.023	0.029	0.030	0.035	0.040	0.043	0.050	0.052
	Split	0.029	0.045	0.053	0.065	0.071	0.078	0.083	0.091	0.096	0.099
HerBERT	Diac	0.003	0.005	0.005	0.006	0.006	0.007	0.009	0.010	0.012	0.014
	Key	0.026	0.041	0.049	0.055	0.060	0.062	0.067	0.077	0.081	0.087
	OCR	0.027	0.038	0.053	0.057	0.063	0.067	0.073	0.078	0.086	0.095
	Ort	0.006	0.008	0.010	0.011	0.013	0.015	0.016	0.018	0.021	0.026
	R-Del	0.019	0.032	0.036	0.042	0.043	0.050	0.053	0.058	0.062	0.070
	R-Ins	0.022	0.037	0.043	0.052	0.058	0.063	0.067	0.070	0.075	0.080
	R-Sub	0.025	0.039	0.048	0.057	0.065	0.069	0.076	0.081	0.086	0.095
	R-Sw	0.020	0.033	0.041	0.045	0.051	0.053	0.057	0.060	0.064	0.071
	Rel	0.010	0.016	0.018	0.022	0.024	0.027	0.032	0.035	0.039	0.043
	Split	0.018	0.027	0.033	0.037	0.040	0.044	0.047	0.051	0.055	0.060
RoBERTa	Diac	0.006	0.007	0.008	0.010	0.011	0.015	0.021	0.025	0.030	0.037
	Key	0.031	0.049	0.060	0.071	0.080	0.092	0.100	0.112	0.123	0.132
	OCR	0.037	0.058	0.073	0.085	0.092	0.107	0.115	0.124	0.135	0.142
	Ort	0.007	0.010	0.012	0.014	0.018	0.025	0.030	0.037	0.044	0.051
	R-Del	0.027	0.043	0.055	0.061	0.068	0.083	0.090	0.100	0.108	0.118
	R-Ins	0.029	0.047	0.058	0.068	0.077	0.091	0.100	0.110	0.120	0.128
	R-Sub	0.031	0.046	0.059	0.068	0.076	0.088	0.099	0.111	0.120	0.130
	R-Sw	0.030	0.043	0.059	0.071	0.077	0.090	0.103	0.109	0.119	0.130
	Rel	0.011	0.017	0.021	0.024	0.027	0.033	0.038	0.044	0.051	0.059
	Split	0.027	0.042	0.053	0.063	0.071	0.080	0.083	0.091	0.098	0.109

Table 10: 使用 IntegratedGradients 作为归因方法时，针对不同数量的更改词汇的 ASR，汇总在数据集上。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.003	0.006	0.006	0.008	0.010	0.011	0.013	0.015	0.019	0.021
	Key	0.044	0.073	0.088	0.101	0.113	0.126	0.132	0.138	0.143	0.149
	OCR	0.045	0.080	0.098	0.118	0.134	0.147	0.160	0.165	0.170	0.174
	Ort	0.009	0.014	0.016	0.019	0.023	0.025	0.030	0.034	0.037	0.041
	R-Del	0.043	0.068	0.085	0.104	0.119	0.130	0.134	0.142	0.149	0.155
	R-Ins	0.045	0.073	0.085	0.103	0.117	0.124	0.131	0.145	0.150	0.155
	R-Sub	0.043	0.072	0.086	0.105	0.117	0.129	0.137	0.144	0.145	0.148
	R-Sw	0.042	0.068	0.082	0.099	0.110	0.124	0.138	0.142	0.149	0.157
	Rel	0.015	0.027	0.031	0.040	0.044	0.050	0.055	0.057	0.064	0.070
	Split	0.048	0.074	0.087	0.104	0.120	0.129	0.139	0.146	0.150	0.153
HerBERT	Diac	0.003	0.003	0.004	0.005	0.005	0.007	0.009	0.011	0.012	0.014
	Key	0.029	0.046	0.061	0.072	0.082	0.090	0.098	0.109	0.118	0.126
	OCR	0.028	0.048	0.058	0.067	0.078	0.085	0.091	0.101	0.111	0.115
	Ort	0.005	0.009	0.013	0.014	0.018	0.019	0.022	0.024	0.026	0.029
	R-Del	0.023	0.036	0.046	0.055	0.060	0.070	0.078	0.082	0.088	0.095
	R-Ins	0.026	0.041	0.054	0.065	0.073	0.078	0.087	0.093	0.097	0.102
	R-Sub	0.026	0.043	0.057	0.070	0.081	0.091	0.100	0.112	0.117	0.125
	R-Sw	0.024	0.040	0.049	0.059	0.065	0.075	0.081	0.083	0.089	0.099
	Rel	0.010	0.018	0.026	0.032	0.034	0.037	0.042	0.043	0.044	0.048
	Split	0.019	0.029	0.039	0.046	0.050	0.055	0.062	0.073	0.077	0.083
RoBERTa	Diac	0.003	0.006	0.008	0.010	0.013	0.017	0.021	0.025	0.031	0.037
	Key	0.045	0.079	0.101	0.124	0.143	0.160	0.175	0.195	0.208	0.219
	OCR	0.055	0.093	0.119	0.136	0.154	0.171	0.188	0.207	0.219	0.230
	Ort	0.009	0.016	0.019	0.023	0.029	0.035	0.042	0.048	0.055	0.062
	R-Del	0.040	0.070	0.094	0.115	0.131	0.148	0.164	0.177	0.192	0.209
	R-Ins	0.043	0.076	0.099	0.120	0.137	0.153	0.170	0.187	0.204	0.212
	R-Sub	0.047	0.083	0.102	0.127	0.143	0.160	0.175	0.194	0.209	0.219
	R-Sw	0.040	0.074	0.094	0.116	0.138	0.158	0.174	0.194	0.208	0.218
	Rel	0.016	0.026	0.033	0.041	0.049	0.059	0.066	0.073	0.083	0.089
	Split	0.037	0.069	0.089	0.105	0.117	0.134	0.148	0.165	0.178	0.184

Table 11: 在使用注意力展开作为归因方法时，对于不同数量的单词的 ASR，汇总于数据集之上。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.004	0.005	0.006	0.007	0.009	0.011	0.013	0.015	0.019	0.021
	Key	0.030	0.057	0.070	0.090	0.111	0.127	0.143	0.156	0.163	0.171
	OCR	0.032	0.063	0.082	0.100	0.117	0.138	0.153	0.167	0.181	0.190
	Ort	0.004	0.008	0.012	0.014	0.018	0.022	0.025	0.030	0.034	0.038
	R-Del	0.025	0.048	0.064	0.084	0.106	0.122	0.143	0.151	0.165	0.171
	R-Ins	0.033	0.058	0.074	0.092	0.110	0.127	0.138	0.147	0.159	0.167
	R-Sub	0.032	0.055	0.071	0.089	0.109	0.127	0.139	0.152	0.164	0.173
	R-Sw	0.029	0.052	0.066	0.091	0.107	0.130	0.144	0.150	0.163	0.177
	Rel	0.004	0.008	0.014	0.023	0.030	0.039	0.048	0.053	0.058	0.062
	Split	0.026	0.051	0.064	0.083	0.106	0.124	0.139	0.149	0.162	0.171
HerBERT	Diac	0.003	0.003	0.004	0.005	0.006	0.006	0.008	0.009	0.010	0.013
	Key	0.024	0.049	0.068	0.077	0.086	0.099	0.105	0.115	0.123	0.130
	OCR	0.026	0.051	0.065	0.078	0.094	0.100	0.109	0.114	0.117	0.120
	Ort	0.005	0.006	0.009	0.009	0.011	0.013	0.015	0.018	0.019	0.022
	R-Del	0.018	0.033	0.041	0.050	0.053	0.060	0.066	0.072	0.076	0.087
	R-Ins	0.024	0.044	0.056	0.070	0.075	0.082	0.090	0.098	0.102	0.107
	R-Sub	0.028	0.049	0.066	0.078	0.085	0.099	0.102	0.109	0.119	0.125
	R-Sw	0.014	0.033	0.042	0.049	0.057	0.064	0.070	0.075	0.081	0.089
	Rel	0.008	0.011	0.015	0.019	0.024	0.026	0.031	0.035	0.038	0.045
	Split	0.017	0.033	0.041	0.048	0.059	0.064	0.072	0.072	0.077	0.079
RoBERTa	Diac	0.003	0.005	0.006	0.007	0.010	0.015	0.020	0.025	0.030	0.036
	Key	0.027	0.048	0.060	0.072	0.090	0.113	0.132	0.140	0.158	0.174
	OCR	0.036	0.065	0.082	0.099	0.113	0.125	0.142	0.154	0.171	0.180
	Ort	0.008	0.012	0.013	0.016	0.023	0.030	0.037	0.044	0.052	0.057
	R-Del	0.022	0.036	0.046	0.063	0.081	0.103	0.119	0.130	0.147	0.158
	R-Ins	0.028	0.047	0.059	0.075	0.090	0.113	0.129	0.142	0.157	0.168
	R-Sub	0.030	0.048	0.063	0.075	0.094	0.110	0.128	0.142	0.157	0.167
	R-Sw	0.021	0.039	0.056	0.065	0.078	0.100	0.121	0.137	0.155	0.170
	Rel	0.004	0.008	0.013	0.015	0.024	0.033	0.042	0.052	0.060	0.071
	Split	0.024	0.040	0.052	0.065	0.080	0.102	0.119	0.134	0.150	0.156

Table 12: 在使用注意力梯度展开作为归因方法时，改变不同单词数量的 ASR，汇总于数据集。

Model	Aug	1	2	3	4	5	6	7	8	9	10
PolBERT	Diac	0.006	0.007	0.008	0.012	0.014	0.015	0.018	0.019	0.022	0.026
	Key	0.070	0.106	0.132	0.153	0.176	0.193	0.200	0.208	0.220	0.230
	OCR	0.076	0.122	0.145	0.169	0.188	0.201	0.215	0.222	0.231	0.239
	Ort	0.010	0.014	0.020	0.024	0.027	0.029	0.033	0.036	0.040	0.045
	R-Del	0.062	0.104	0.131	0.157	0.168	0.181	0.194	0.204	0.215	0.224
	R-Ins	0.066	0.106	0.129	0.147	0.165	0.179	0.188	0.200	0.212	0.219
	R-Sub	0.072	0.115	0.137	0.158	0.180	0.188	0.206	0.216	0.221	0.228
	R-Sw	0.061	0.099	0.131	0.151	0.168	0.186	0.197	0.204	0.216	0.226
	Rel	0.019	0.039	0.049	0.060	0.065	0.071	0.076	0.078	0.084	0.089
	Split	0.064	0.097	0.127	0.155	0.172	0.185	0.196	0.211	0.218	0.228
HerBERT	Diac	0.003	0.004	0.005	0.005	0.006	0.008	0.009	0.011	0.014	0.016
	Key	0.033	0.056	0.073	0.083	0.096	0.107	0.114	0.122	0.130	0.139
	OCR	0.034	0.053	0.075	0.085	0.099	0.110	0.115	0.127	0.132	0.141
	Ort	0.005	0.008	0.009	0.011	0.013	0.014	0.018	0.020	0.022	0.024
	R-Del	0.026	0.041	0.056	0.066	0.076	0.078	0.085	0.093	0.100	0.104
	R-Ins	0.026	0.044	0.059	0.067	0.074	0.083	0.091	0.100	0.104	0.111
	R-Sub	0.031	0.054	0.070	0.082	0.088	0.102	0.115	0.124	0.130	0.144
	R-Sw	0.027	0.042	0.055	0.062	0.071	0.082	0.093	0.099	0.105	0.115
	Rel	0.007	0.015	0.020	0.024	0.029	0.033	0.038	0.041	0.044	0.048
	Split	0.023	0.040	0.054	0.062	0.069	0.076	0.082	0.086	0.091	0.096
RoBERTa	Diac	0.005	0.008	0.010	0.011	0.014	0.019	0.024	0.028	0.033	0.039
	Key	0.038	0.067	0.098	0.119	0.138	0.156	0.169	0.184	0.204	0.216
	OCR	0.045	0.087	0.115	0.132	0.152	0.169	0.184	0.198	0.218	0.231
	Ort	0.008	0.012	0.020	0.022	0.026	0.033	0.043	0.049	0.057	0.063
	R-Del	0.035	0.060	0.090	0.112	0.129	0.154	0.172	0.183	0.198	0.211
	R-Ins	0.034	0.061	0.089	0.105	0.127	0.150	0.166	0.178	0.194	0.205
	R-Sub	0.037	0.070	0.099	0.121	0.139	0.162	0.174	0.185	0.203	0.215
	R-Sw	0.035	0.066	0.096	0.119	0.134	0.154	0.171	0.187	0.208	0.223
	Rel	0.010	0.019	0.029	0.037	0.044	0.054	0.062	0.072	0.082	0.092
	Split	0.032	0.059	0.085	0.100	0.116	0.133	0.149	0.161	0.178	0.190

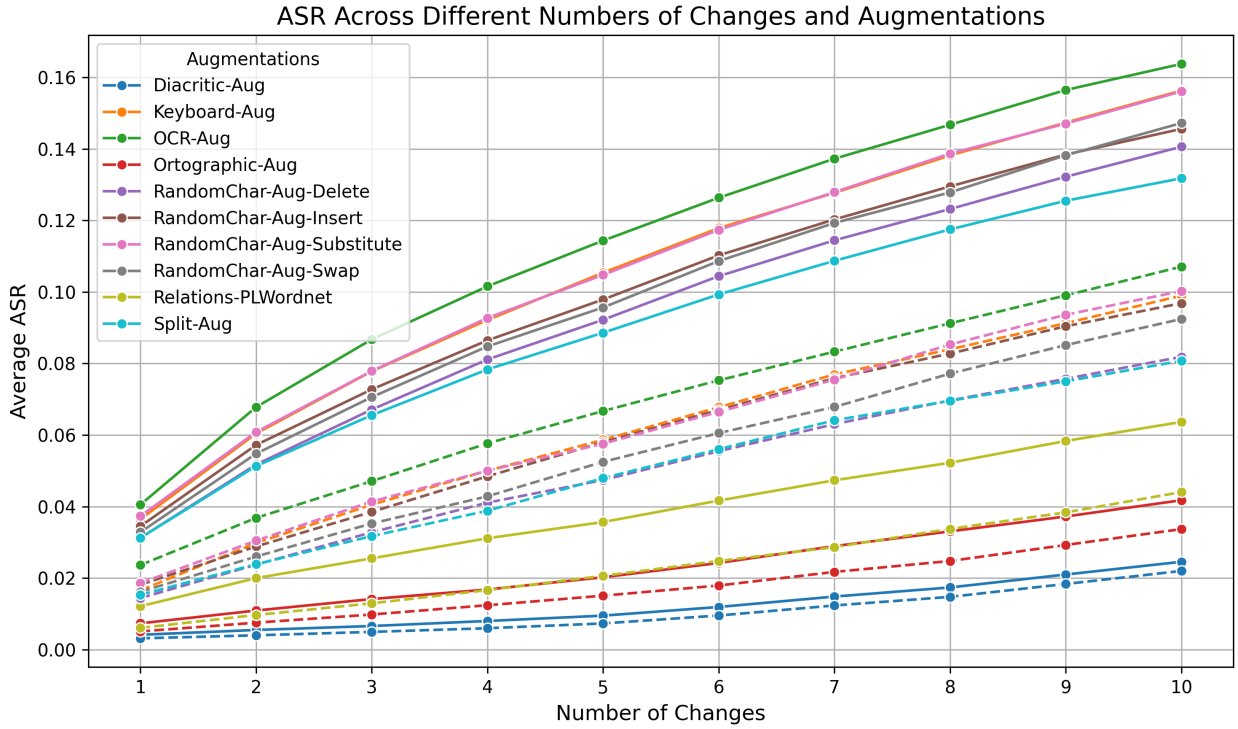

Figure 3: ASR 在较小的语言模型和不同归因方法中受扰动词数量的关系。虚线表示的是选择随机词进行扰动时的 ASR，而不是基于词重要性选择词。

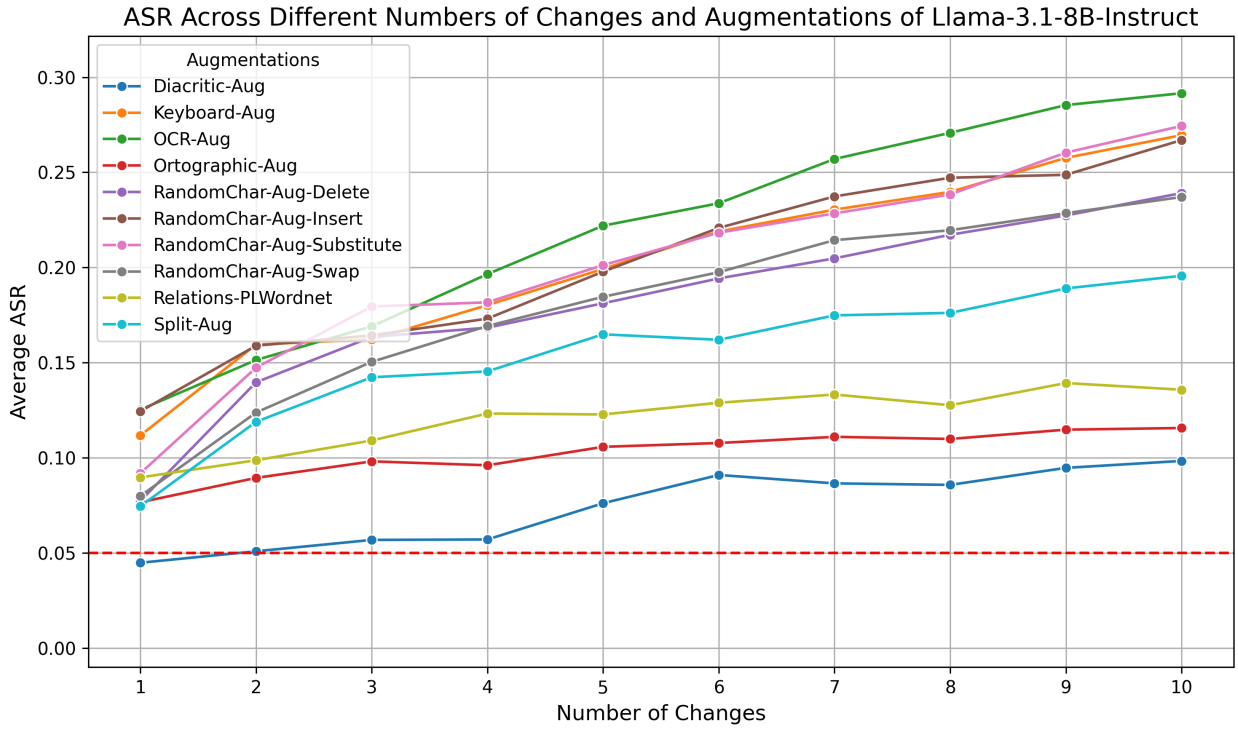

Figure 4: ASR 在 Llama 上的关系以及不同归因方法的扰动词数量。虚线表示稳健性临界值，超过该值模型被认为对简单扰动不稳健。

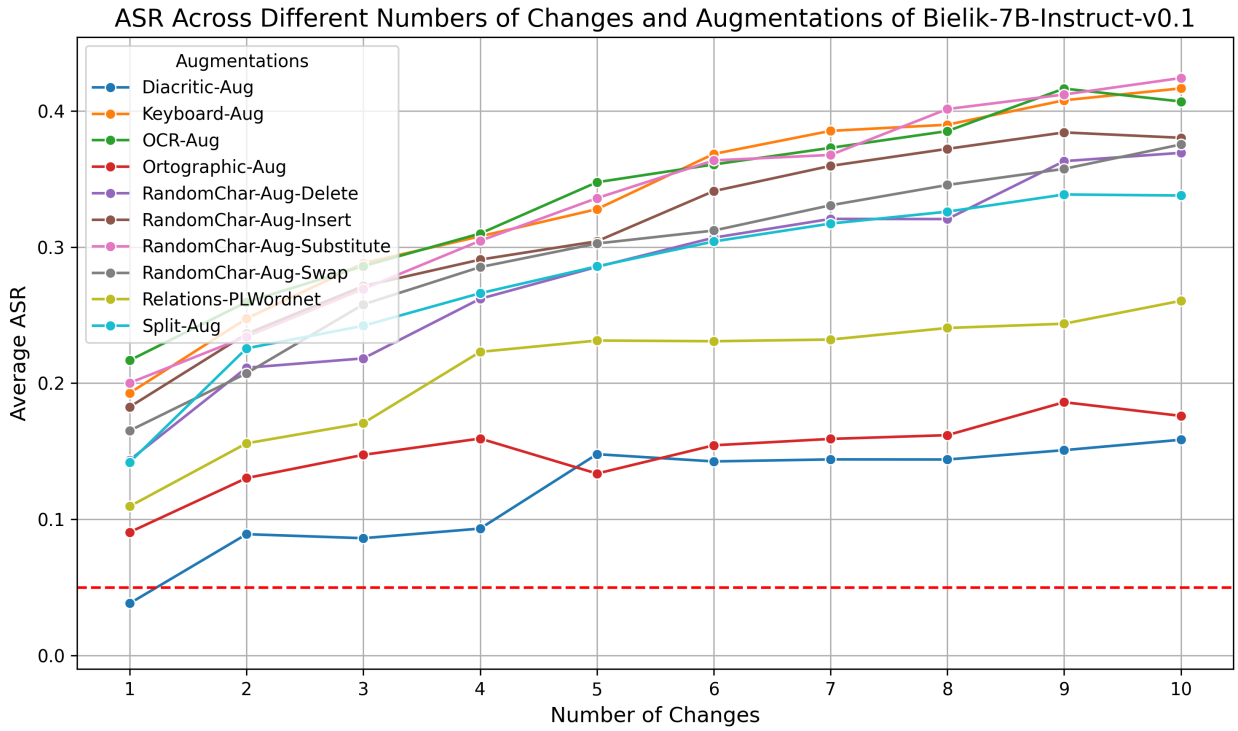

Figure 5: Bielik 上的 ASR 与不同归因方法的扰动词数量之间的关系。虚线表示稳健性临界值，超过这个值模型被认为对简单扰动不稳健。