大型语言模型的道德鸿沟

Maciej Skórski
maciej.skorski@gmail.com

Alina Landowska
alina.landowska@gmail.com

Abstract

道德基础检测对于分析社会话语和开发符合伦理的人工智能系统至关重要。尽管大型语言模型在各种任务中表现出色，但其在专门的道德推理方面的表现仍不明朗。本研究首次提供了对最新 LLMs 和经过微调的变压器在 Twitter 和 Reddit 数据集上的全面比较，使用 ROC、PR 和 DET 曲线分析。结果显示存在显著的性能差距，虽然进行了提示工程，但 LLMs 表现出较高的假阴性率和系统性地未能检测到道德内容。这些发现表明，在道德推理应用中，任务特定的微调仍然优于提示。

1 介绍

1.1 动机与背景

道德基础理论（MFT）提供了一个框架，用于理解各文化中的道德推理和判断的心理基础 (?) 。MFT 的发展是为了解释社会内部和社会之间道德直觉的差异，主张道德判断由一组进化的道德基础——例如关爱、公平、忠诚、权威和圣洁——所塑造，这些基础反映了潜在的心理机制 (??) 。

MFT 已经在多个领域得到了广泛应用，包括政治意识形态 (?) 、环境态度 (?) 、疫苗犹豫 (?) 、社会规范 (?) 、新闻框架分析 (?) 、社交媒体言论 (?) 、日常道德困境 (?) 和论证评估 (??) 。

Social Media Post	Moral Foundation
"My heart breaks seeing children separated from families at the border"	关心
"Everyone deserves equal access to healthcare regardless of income"	公平
"Respect your elders and follow traditional values that built this nation"	权威
"Stand with our troops - they sacrifice everything for our freedom"	忠诚
"Marriage is sacred and should be protected from secular corruption"	神圣

Table 1: **帖子及其相关的道德基础**

用于检测道德基础的计算方法已经从基于词典的方法 (??) 发展到变压器模型 (??) ，最近又发展到大型语言模型 (?) 。虽然 Anthropic 和 OpenAI 的 LLMs 提供了吸引人的可访问性，但没有研究严格比较过它们在道德基础检测方面的性能与专门的变压器模型的性能。这一差距使研究人员在工具选择上缺乏基于证据的指导，因此对于在道德敏感的应用中进行知情部署，系统评估是必不可少的。本文通过跨多个数据集和评估指标的全面实证比较来解决这一需求。

1.2 贡献

对计算道德心理学的关键贡献包括：

(1) 综合评估：首次对比最先进的大规模语言模型（Claude Sonnet 4, GPT-o1-mini）与经过微调的变压器模型（DeBERTa, RoBERTa）在推特和 Reddit 数据集上的道德基础检测，建立了域内和跨域场景的性能基准。
(2) 严谨的方法论：通过使用诊断曲线谱（ROC、PR 和 DET）建立稳健的评估框架，以提供全面的性能洞察，并解决之前仅依赖 ROC 分析的工作中存在的类别不平衡限制。
(3) 错误分析与实践指南：识别系统性的 LLM 局限性，包括高假阴性率（58-90 %）、特定基础的失效模式和保守的预测偏见，同时提供基于证据的建议，用于模型选择、提示工程限制以及在道德内容分析中的部署注意事项。

1.3 相关工作

早期的道德基础检测计算方法主要依赖于基于字典的方法，使用诸如道德基础词典 (?) 及其扩展 (??) 等手工制作的词典。虽然由于可解释性和计算效率仍然非常受欢迎，但这些方法的准确性非常低。

深度学习的出现，特别是 transformer 架构的出现，标志着道德内容分析的显著进步。? 首次将深度学习模型应用于道德基础分类，展示了相对于词典基础方法的显著改进。研究人员通过微调基于 transformer 的模型（如 BERT 和 RoBERTa (???) ）进一步提高了准确性，这些模型目前达到最先进的性能。

最近提出了将大型语言模型应用于道德内容分类的建议 (?) 展现了潜力，但存在方法上的局限性。该研究仅使用了一个数据集，去除了模糊（较难）的实例，并且处理注释者意见分歧的方式与该领域的标准实践不同，导致了评估的偏倚。

2 数据与方法

2.1 数据集

我们评估模型在两个已建立的道德基础数据集上的表现。Twitter 数据集（MFTC）(?) 包含 34,987 条推文，涵盖七个与社会相关的话题，由经过培训的标注人员标记道德基础及其情感。我们合并美德/恶德标签（每个基础的正面/负面方面，例如，“纯洁” + “堕落” → “神圣”）。Reddit 数据集（MFRC）(?) 包含来自 12 个子版块的 17,886 条评论，涵盖美国政治、法国政治和日常道德生活。我们将原始的平等/比例划分重新合并为公平，并将“薄道德”案例视为不存在基础。两个数据集都使用二进制标签表示五个道德基础（权威、关怀、公平、忠诚、神圣），采用包容性标注方案（只要有一个标注者同意即为正面）。数据在 Table 2 和 Figure 1 中汇总。

Table 2: **数据集摘要**

Dataset	Category/Topic	Count	%
MFTC (Twitter)			
	All Lives Matter	4,988	14.3 %
	Black Lives Matter	4,990	14.3 %
	2016 US Presidential Election	4,987	14.3 %
	Hate Speech	4,989	14.3 %
	Hurricane Sandy	4,990	14.3 %
	# MeToo	4,995	14.3 %
	Baltimore Protests	4,985	14.2 %
	MFTC Subtotal	34,924	100 %
MFRC (Reddit)			
	US Politics	6,968	39.0 %
	French Politics	3,984	22.3 %
	Everyday Moral Life	6,933	38.7 %
	MFRC Subtotal	17,885	100 %
TOTAL ALL DATASETS			52,809

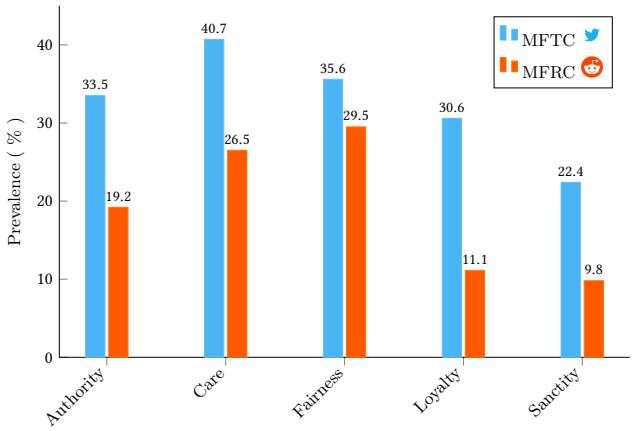

Figure 1: **数据集中的道德基础普及率**

2.2 模型

大型语言模型 . 我们基于成本效益和推理能力选择了四个最近的模型（Table 3 ）。这些模型包括通用模型（Haiku, GPT-4o-mini）和专门用于推理的模型（Sonnet, GPT-o1-mini），它们来自主要供应商。模型代表了最新的技术发布：GPT-o1-mini（2025 年 4 月），Claude Sonnet（3.5 版本来自 2025 年 2 月，4 版本来自 2025 年 5 月）。所有模型都通过 Python API 访问，并使用一致的提示策略。

Table 3: **大型语言模型规范**

Model	Vendor	Context	Output	Price ($ /1M)
Claude 3.5 Haiku	Anthropic	200K	4K	$ 0.25
Claude Sonnet 4	Anthropic	200K	8K	$ 3.00
GPT-4o-mini	OpenAI	128K	16K	$ 0.15
GPT-o4-mini	OpenAI	128K	65K	$ 3.00

Context and output in tokens. Pricing for input tokens (May 2025).

Transformer 模型 . 虽然之前的研究确立了 BERT 模型，尤其是 RoBERTa，在道德基础检测方面是最先进的，但我们评估了更新的 Transformer 架构。本文报告的结果中，我们使用了微软的 DeBERTa-v3-base，训练时用的学习率是 2e-5，并在前三个 epoch 中冻结了前两层（在所有结果中，为简洁起见，“BERT” 均指此 DeBERTa 模型）。对于域内评估，我们使用了 4:1 的训练-测试划分，而对于域外评估，我们在完整的训练集上进行了训练并在完整的测试集上进行了评估。

2.3 技术

预测指标 . 一系列的度量标准被用于连续概率分数和二元预测，以全面评估模型性能。缩写 TP 、FP 、TN 和 FN 分别代表真正例、假正例、真负例和假负例，从而得到：

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

和聚合的度量标准

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}, \quad BER = \frac{1}{2}\left(\frac{FN}{TP + FN} + \frac{FP}{FP + TN}\right)$$

连续分数在各种阈值下进行评估，这些阈值通过三个互补的曲线呈现不同的性能权衡：

- ROC 曲线：绘制 TPR 与 FPR 的关系，AUC 范围从 0.5（随机）到 1.0（完美）(?)
- PR 曲线：绘制精确度与召回率，考虑 AUC 对不平衡数据集更具有信息性 (?)
- DET 曲线：在正态偏差刻度上绘制 (1-TPR) 与 FPR 的关系图，强调错误率 (?)

我们开发了一种新颖的色盲友好色彩方案，能准确反映如 Figure 2 中所展示的道德基础理论结构。该调色板代表着从个体主义基础（冷色调）到集体主义基础（暖色调）的渐变：绿色代表关怀，作为一种包容、养育的基础；蓝色代表公平，反映与正义和神圣审判的联系；红色代表忠诚，象征联系和忠诚度；紫色代表权威，代表传统权力和等级制度；金色代表圣洁，象征神圣和纯净。

Figure 2: **新颖的道德基础色彩调色板**

3 结果

3.1 道德内容检测

如先前的研究所示，在不识别其维度的情况下，检测是否存在任何道德内容，对于使用文本嵌入模型来说相对简单。我们再现了（并稍微改进了）之前通过 transformers 实现的结果。然而，对大规模语言模型（LLMs）的评估揭示了即使在这个更简单的任务上也存在显著的性能差距。如性能曲线所示，大规模语言模型的性能始终较低——它们位于 transformer 曲线内，表明是系统的表现不足而不是阈值依赖的差异。这证明了在道德内容检测能力上，与经过微调的 transformers 相比，大规模语言模型存在根本性的限制。

见 Figure 3 用于 ROC 分析，见 Figure 4 用于 PR 曲线。在下一节中评估基础设施特定性能时，数值指标作为额外的“任何”维度进行报告（Tables 4 and 5 ）。

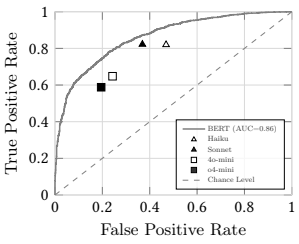

Figure 3(a) MFRC **数据集**

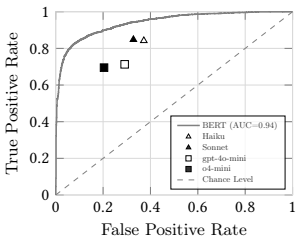

Figure 3(b) MFTC **数据集**

Figure 3: ROC **预测性能**。

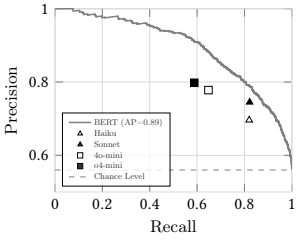

Figure 4(a) MFRC **数据集**

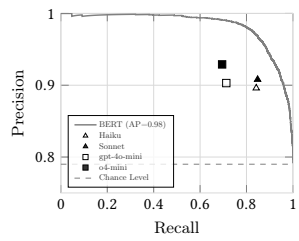

Figure 4(b) MFTC **数据集**

Figure 4: PR **预测性能**

3.2 道德价值观的分类

我们评估的结果，在 Table 4 和 Table 5 中展示，揭示了微调后的 transformers 与 LLMs 之间存在显著的性能差距。BERT 始终优于所有的 LLMs，其 ROC-AUC 分数为 0.83-0.90，F1 分数为 0.38-0.80，而 Claude-3.5-Sonnet 仅能达到 0.21-0.78 的 F1 分数。LLMs 在忠诚和神圣性方面表现最差，显示出高假阴性率（0.58-0.90），错过了大多数正例。虽然 LLMs 对一般道德内容的检测表现尚可，但它们无法识别具体基础，这表明对于道德推理任务，任务专用微调仍然优于提示。

3.2.1 ROC **分析** . 图 ?? 和 ?? 显示了在所有分类阈值下绘制的真实率与假阳性率的 ROC 曲线。这些曲线在视觉上确认了表中报告的显著性能差距，BERT 始终在获得更高的 AUC 值，而 LLM 曲线则系统性地位于 transformer 性能包络线内。

Table 4: **简单提示下** MFRC **的性能指标**

dataset	moral dimension	model	FPR	FNR	precision	recall	f1	ROC-AUC	AP
MFRC	any	BERT	0.225	0.221	0.761	0.862	0.808	0.862	0.893
		BERT-OOD	0.247	0.306	0.691	0.892	0.779	0.800	0.829
		Haiku 3.5	0.469	0.180	0.696	0.820	0.753		
		Sonnet 3.7	0.374	0.184	0.741	0.816	0.777		
		Sonnet 4	0.369	0.178	0.745	0.822	0.782		
		gpt-4o-mini	0.243	0.352	0.778	0.648	0.707		
		o4-mini	0.195	0.412	0.798	0.588	0.677		
	authority	BERT	0.276	0.137	0.549	0.638	0.590	0.869	0.605
		BERT-OOD	0.280	0.297	0.422	0.600	0.495	0.775	0.453
		Haiku 3.5	0.229	0.529	0.330	0.471	0.388		
		Sonnet 3.7	0.143	0.523	0.444	0.477	0.460		
		Sonnet 4	0.144	0.573	0.416	0.427	0.421		
		gpt-4o-mini	0.146	0.623	0.383	0.377	0.380		
		o4-mini	0.091	0.769	0.379	0.231	0.287		
	care	BERT	0.217	0.156	0.691	0.720	0.706	0.896	0.775
		BERT-OOD	0.265	0.234	0.533	0.727	0.615	0.823	0.626
		Haiku 3.5	0.255	0.405	0.451	0.595	0.513		
		Sonnet 3.7	0.155	0.359	0.593	0.641	0.616		
		Sonnet 4	0.192	0.316	0.556	0.684	0.613		
		gpt-4o-mini	0.137	0.477	0.573	0.523	0.547		
		o4-mini	0.155	0.471	0.546	0.529	0.537		
	fairness	BERT	0.275	0.215	0.544	0.785	0.643	0.830	0.711
		BERT-OOD	0.368	0.266	0.455	0.734	0.562	0.745	0.533
		Haiku 3.5	0.362	0.416	0.419	0.584	0.488		
		Sonnet 3.7	0.264	0.387	0.509	0.613	0.556		
		Sonnet 4	0.201	0.493	0.530	0.507	0.518		
		gpt-4o-mini	0.125	0.765	0.456	0.235	0.310		
		o4-mini	0.150	0.618	0.532	0.382	0.445		
	loyalty	BERT	0.194	0.202	0.471	0.639	0.542	0.878	0.563
		BERT-OOD	0.243	0.357	0.372	0.415	0.392	0.777	0.395
		Haiku 3.5	0.150	0.685	0.199	0.315	0.244		
		Sonnet 3.7	0.091	0.584	0.351	0.416	0.381		
		Sonnet 4	0.100	0.535	0.354	0.465	0.402		
		gpt-4o-mini	0.052	0.892	0.198	0.108	0.140		
		o4-mini	0.072	0.801	0.245	0.199	0.220		
	sanctity	BERT	0.293	0.131	0.358	0.566	0.439	0.859	0.444
		BERT-OOD	0.283	0.258	0.305	0.531	0.388	0.804	0.376
		Haiku 3.5	0.160	0.715	0.171	0.285	0.213		
		Sonnet 3.7	0.067	0.695	0.345	0.305	0.323		
		Sonnet 4	0.064	0.695	0.353	0.305	0.327		
		gpt-4o-mini	0.060	0.903	0.157	0.097	0.120		
		o4-mini	0.076	0.857	0.178	0.143	0.158		

Table 5: **简单提示下的** MFTC **性能指标**

dataset	moral dimension	model	FPR	FNR	precision	recall	f1	ROC-AUC	AP
MFTC	any	BERT	0.075	0.176	0.902	0.959	0.930	0.940	0.983
		BERT-OOD	0.189	0.238	0.867	0.934	0.899	0.860	0.957
		Haiku 3.5	0.372	0.158	0.896	0.842	0.868		
		Sonnet 3.7	0.321	0.155	0.909	0.845	0.876		
		Sonnet 4	0.328	0.152	0.908	0.848	0.877		
		gpt-4o-mini	0.291	0.287	0.903	0.713	0.797		
		o4-mini	0.204	0.305	0.929	0.695	0.795		
	authority	BERT	0.204	0.173	0.743	0.742	0.742	0.899	0.830
		BERT-OOD	0.205	0.395	0.567	0.645	0.603	0.749	0.645
		Haiku 3.5	0.151	0.611	0.551	0.389	0.456		
		Sonnet 3.7	0.136	0.553	0.610	0.447	0.516		
		Sonnet 4	0.139	0.583	0.589	0.417	0.489		
		gpt-4o-mini	0.201	0.603	0.485	0.397	0.436		
		o4-mini	0.146	0.698	0.498	0.302	0.376		
	care	BERT	0.115	0.244	0.805	0.769	0.787	0.897	0.881
		BERT-OOD	0.252	0.261	0.663	0.748	0.703	0.811	0.765
		Haiku 3.5	0.314	0.343	0.594	0.657	0.624		
		Sonnet 3.7	0.295	0.298	0.624	0.702	0.661		
		Sonnet 4	0.339	0.254	0.606	0.746	0.668		
		gpt-4o-mini	0.251	0.447	0.606	0.553	0.579		
		o4-mini	0.224	0.400	0.651	0.600	0.624		
	fairness	BERT	0.111	0.232	0.792	0.768	0.780	0.906	0.874
		BERT-OOD	0.317	0.286	0.512	0.819	0.630	0.770	0.659
		Haiku 3.5	0.243	0.338	0.601	0.662	0.630		
		Sonnet 3.7	0.204	0.287	0.659	0.713	0.685		
		Sonnet 4	0.197	0.322	0.655	0.678	0.666		
		gpt-4o-mini	0.128	0.734	0.534	0.266	0.355		
		o4-mini	0.190	0.437	0.620	0.563	0.590		
	loyalty	BERT	0.159	0.267	0.670	0.733	0.700	0.866	0.782
		BERT-OOD	0.228	0.460	0.442	0.680	0.536	0.713	0.551
		Haiku 3.5	0.083	0.818	0.485	0.182	0.265		
		Sonnet 3.7	0.071	0.711	0.636	0.289	0.398		
		Sonnet 4	0.083	0.676	0.627	0.324	0.427		
		gpt-4o-mini	0.094	0.880	0.355	0.120	0.179		
		o4-mini	0.124	0.786	0.425	0.214	0.284		
	sanctity	BERT	0.138	0.262	0.688	0.654	0.671	0.875	0.740
		BERT-OOD	0.340	0.271	0.433	0.616	0.509	0.761	0.523
		Haiku 3.5	0.122	0.762	0.360	0.238	0.286		
		Sonnet 3.7	0.080	0.674	0.541	0.326	0.407		
		Sonnet 4	0.095	0.646	0.517	0.354	0.420		
		gpt-4o-mini	0.097	0.839	0.325	0.161	0.216		
		o4-mini	0.115	0.782	0.354	0.218	0.270		

3.2.2 PR **分析** . 图 ?? 和 ?? 展示了精确率-召回率曲线，对于像道德基础分类这样不平衡的数据集，这些曲线特别具有参考价值。这些图表证实了表格结果，显示经过微调的变换器在所有道德基础上相比于大型语言模型保持更优的精确率-召回率权衡。

3.2.3 DET **曲线** . ?? 和 ?? 中的结果显示检测错误权衡曲线，这些曲线在标准偏差刻度上绘制了假阴性与假阳性率。这些曲线强调了错误率的观点，并证实了表中记录的大型语言模型的高假阴性率，特别是在忠诚和纯洁基础领域。