用户对注意力可视化的感知：对基于证据的医学文档可解释性的影响

Andrés Carvallo[1], Denis Parra[2], Peter Brusilovsky[3], Hernan Valdivieso[2], Gabriel Rada[2], Ivania Donoso[4], and Vladimir Araujo[2]

[1] CENIA –Centro Nacional de Inteligencia Artificial, Chile
afcarvallo@uc.cl
[2] Pontificia Universidad Católica de Chile, Chile
{ dparras,hfvaldivieso,vgaraujo } @uc.cl, gabriel@rada.cl
[3] University of Pittsburgh, United States
peterb@pitt.edu
[4] KU Leuven, Belgium
indonoso@uc.cl

Abstract. 注意力机制是 Transformer 架构的核心组成部分。除了提高性能之外，注意力机制被提出作为通过注意力权重来解释性的方法，这些权重与输入特征（例如，文档中的标记）相关。在这种情况下，较大的注意力权重可能暗示对模型预测更相关的特征。在循证医学中，此类解释可以支持医生对用于分类生物医学文献的 AI 系统的理解和交互。然而，关于注意力权重是否提供有用的解释仍然没有共识。此外，很少有研究探索如何将注意力可视化来影响其作为解释辅助的效用。为了弥合这一差距，我们进行了用户研究，以评估基于注意力的解释是否支持用户进行生物医学文档分类，以及是否存在更优的可视化它们的方式。该研究涉及来自不同学科的医学专家，他们根据研究设计（例如，系统综述、广泛综合、随机和非随机试验）对文章进行分类。我们的研究结果表明，Transformer 模型（XLNet）准确地分类了文件；然而，注意力权重并未被认为对解释预测特别有帮助。然而，这种看法因注意力的可视化方式而显著变化。与 Munzner 提出的视觉效果原则相反，后者偏好像条形长度这种精确编码，用户更喜欢更直观的格式，例如文本亮度或背景颜色。虽然我们的结果未能确认注意力权重对解释的整体效用，但它们表明其感知的帮助程度受其视觉呈现方式的影响。

1 引言

Transformers (Vaswani et al., 2017) 在包括自然语言处理 (NLP) (Canchila et al., 2024) 、计算机视觉 (Khan et al., 2022) 和信息检索 (Wang et al., 2024) 等广泛任务中取得了最先进的成果。

尽管它们取得了成功，Transformer 常常因为缺乏可解释性而受到批评。它们复杂的架构，涉及数百万参数，使得理解其预测背后的推理变得困难。

注意力机制 (Bahdanau et al., 2014) ，最初是为了增强序列到序列模型的性能而引入的，如今已成为 Transformer 中的核心组件。自注意力使模型能通过对输入元素（如文档中的标记）分配权重来捕捉上下文关系。这些注意力权重被提出作为模型预测的一种潜在解释形式 (Parra et al., 2019) 。

然而，关于注意力权重是否真正提供了有意义或可信的解释，仍存在持续的争论 (Jain and Wallace, 2019) 。此外，研究是否有某些方式可视化文本中的注意力被用户视为更有帮助的仍然有限。
这个问题在循证医学（EBM）中尤为重要，在这里，临床医生必须快速评估大量文献以支持医学决策 (Elliott et al., 2014) 。在这种情况下，AI 系统不仅要准确，还需提供直观、值得信赖的解释，以帮助用户更高效地工作。
为了弥补这一差距，我们：

1. 开发了一个系统，该系统对生物医学研究文章进行分类，并基于 Transformer 模型的注意力权重生成视觉解释。
2. 进行了一项用户研究，以评估注意力权重和模型预测概率在生物医学文档分类中是否被认为是有用的解释。
3. 比较了不同的文本注意力可视化方式，并评估了它们的感知实用性是否因所审阅文档的类型而有所不同。

2 相关工作

2.1 作为解释的注意力

使用注意力权重已经被提议作为解释基于 Transformer 模型的一种方法 (Parra et al., 2019)；然而，它们解释预测的能力仍然存在争议。虽然一些人认为注意力权重不反映模型的推理 (Jain and Wallace, 2019)，但另一些人支持在某些条件下它们的效用 (Wiegreffe and Pinter, 2019) 。最近的研究甚至质疑注意力的整体作用，转而指向前馈层 (Geva et al., 2022) 。尽管如此，注意力可视化在生物医学 NLP 中仍具有相关性，其中基于编码器的模型在特定领域语料库上进行微调可以产生可解释的模式 (Roccabruna et al., 2024) 。之前在生物医学文本分类中的工作使用了基于 Transformer 的模型 (Carvallo et al., 2020b,a; Carvallo and Parra, 2019)，并研究了其稳健性 (Araujo et al., 2020a; Aspillaga et al., 2020; Araujo et al., 2021) 。在这项工作中，我们基于这些基础，探索了注意力可视化如何影响医学文献分类中感知的实用性。

2.2 注意力可视化界面

像 `BertViz` (Vig, 2019) 和 `AttentionViz` (Yeh et al., 2023) 这样的工具使用户能够查看不同层和头部的注意力权重。然而，它们很少评估文本中显示的注意力如何影响人类的感知。我们的研究通过进行用户评估，重点关注注意力可视化在不同类型的生物医学证据中的感知有用性，以此补充这一研究方向。我们还在循证医学中的先前应用 (Carvallo et al., 2023) 和生物医学 NLP 任务中的对抗性评估 (Araujo et al., 2020b) 的基础上，评估基于注意力的解释在现实的、高风险环境中的表现。

2.3 关于可解释性的用户研究

以用户为中心的 XAI 研究表明，解释的有效性取决于用户的专业知识、控制和上下文 (Cai et al., 2019; Eiband et al., 2019) 。最近的研究强调了解释设计与特定领域需求和用户特征对齐的重要性，特别是在医疗保健领域 (E. Ihongbe et al.,

2024; Kim et al., 2023) 。在这项工作中，我们通过评估医学专家如何看待不同的注意力可视化解释，以及这些可视化是否支持他们分类生物医学证据的任务，为这一研究方向做出贡献。

3 可解释界面

我们在 Epistemonikos [5] 平台内开发了一个接口，以便与基于 Transformer 的模型进行交互，该模型能够突出显示词汇层级的注意力分数。Epistemonikos 是一个专注于循证医学的非营利组织，广泛用于医生群体。为了保留原有的工作流程，我们将界面作为 Chrome 扩展 [6] 进行部署，该扩展覆盖了视觉解释，而不修改底层系统。

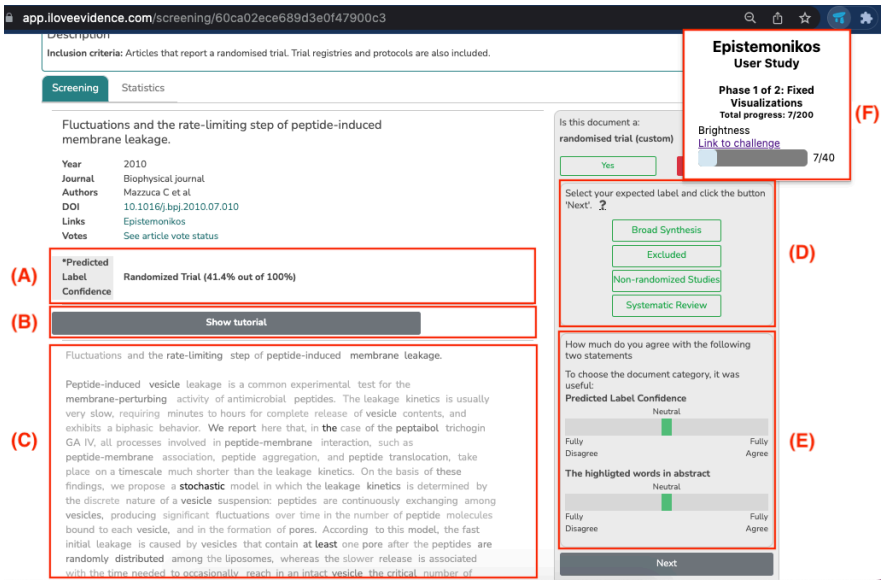

Fig. 1. 用于评价生物医学文档分类中基于注意力解释的 Epistemonikos 用户研究界面的截图。(A) 模型预测的标签和置信度得分。(B) 互动教程开关。(C) 生物医学摘要，带有基于注意力的词高亮（例如，通过亮度）。(D) 文档分类的用户标签选择选项。(E) 对解释组件感知实用性的 Likert 量表反馈。(F) 研究进度和当前可视化条件。

图 1 展示了该界面设计的六个关键组成部分：（A）模型预测的研究类型，（B）帮助/教程按钮，（C）带有词级注意力高亮的摘要，（D）用户标签选择，（E）关于预测标签和高亮词的反馈，以及（F）进度条。该界面使得不同的注意力可视化进行比较，设计遵循信息可视化中的有效性原则，该原则优先考虑准确编码关键信息的感知性 (Midway, 2020) 。

[5] https://www.epistemonikos.org/
[6] https://chromewebstore.google.com/category/extensions

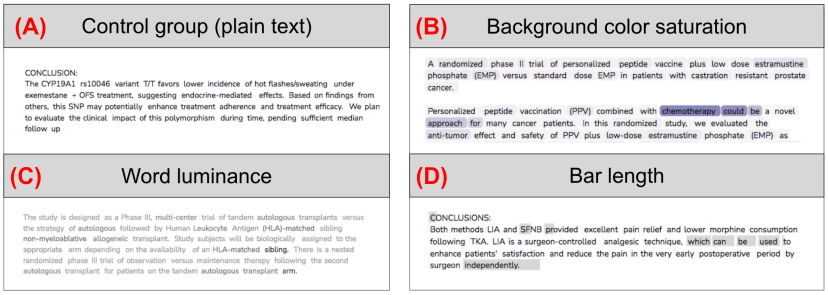

Fig. 2. 本研究中评估的四种文本注意力可视化方法的示例。(A) 纯文本（无注意力提示的对照条件），(B) 背景颜色饱和度，(C) 单词亮度，(D) 每个单词下方的条形长度。

我们测试了三种不同的文本注意力可视化方法，以及一种不使用可视化的对照条件，如图 2 所示。在对照组 (A) 中，摘要显示为普通文本，没有视觉提示。在背景颜色条件 (B) 中，根据每个单词的注意力权重对其背景进行着色——背景越深，该单词越重要。在词亮度条件 (C) 中，文本亮度随注意力变化，使相关单词显得更暗。在条长条件 (D) 中，每个单词下方显示一个水平条，其长度与该单词的重要性成比例。除了注意力可视化，我们还测试了显示模型预测概率（或确定性）是否有助于用户进行分类决策。我们分析了不同类型的循证医学文件的结果——如随机试验、系统综述和广义综合——以评估注意力可视化和模型信心水平如何影响用户的感知有用性和决策。

4 语言模型

我们评估了三种基于注意力机制的编码器模型：BERT (Devlin et al., 2018) 、BioBERT (Lee et al., 2020) 和 XLNet (Yang et al., 2019) 。这些模型在循证医学（EBM）背景下针对医学文献的多类别分类任务进行了微调。分类通过将特殊的 [CLS] 标记传递给一个全连接层，然后经过 softmax 激活，生成五种证据类型的概率分布：综合综述（BS）、排除（EXC）、随机对照试验（PS-RCT）、非随机对照试验（PS-NRCT）和系统综述（SR）。SR 和 PS-RCT 代表医学证据等级体系中的最高层级 (Gopalakrishnan and Ganeshkumar, 2013) 。

我们选择了基于编码器的模型，而不是大型自回归语言模型（LLMs），因为它们在推理过程中的效率更高，微调时更稳定，并可通过注意力权重直接解释。编码器特别适合对固定长度输入进行分类任务，其自注意力机制产生的结构化输出更容易与人类可解释的特征 (Roccabruna et al., 2024) 对齐。

表 1 展示了三个模型在一个大规模的 EBM 数据集上的表现，该数据集由 399,737 份训练文档和 18,854 份测试文档组成，来源于 Epistemonikos。在所有证据类别中，XLNet 在精度、召回率和 F1 分数方面均显著优于 BERT 和 BioBERT。鉴于其卓越表现，我们选择 XLNet 来为可解释界面提供注意力权重。这些权重是从最终编码层提取的，并在多个注意力头中取平均，以生成在可视化中使用的词级重要性分数。

Type	BERT			XLNet			BioBERT		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
BS	0.53	0.37	0.44	0.84	0.77	0.81	0.56	0.69	0.62
EXC	0.86	0.83	0.84	0.97	0.96	0.97	0.90	0.62	0.73
PS-RCT	0.63	0.84	0.72	0.83	0.89	0.86	0.64	0.80	0.71
PS-NRCT	0.91	0.93	0.92	0.99	0.99	0.99	0.82	0.96	0.88
SR	0.90	0.93	0.91	0.94	0.97	0.96	0.94	0.92	0.93
Avg	0.88	0.88	0.88	0.97*	0.97*	0.97*	0.85	0.84	0.84

Table 1. 针对五种生物医学证据类型的文档分类所得结果。表现最佳的数值用粗体表示。* 符号表示基于 Friedman 事后检验的统计显著性。

5 研究设计

我们设计了一项用户研究，以调查不同的解释组件如何影响用户在对生物医学文档进行分类时的感知。具体来说，我们考察了三个因素：(1) 基于注意力的解释是否被认为是有帮助的，(2) 在文本中可视化注意力的某些方式是否更受欢迎或更有效，以及 (3) 模型预测的概率（或确定性）是否支持决策。
这些方面在多种基于证据的医学文章中进行了评估，包括系统评价、随机试验和非随机研究。
研究包括两个阶段：
第一阶段是一个对照实验，参与者使用我们的可解释界面对文章进行分类。在每次分类后，他们使用 5 点李克特量表对模型预测概率的有用性和注意力突出显示的词语进行评分。
第二阶段允许参与者选择他们偏好的注意力可视化方法——或者完全禁用它——并在他们选择的设置下继续对文档进行分类。这个阶段在一个更灵活的互动场景中捕捉用户的偏好。
该研究涉及来自不同专业的五位医学专家，每位专家标记了 200 篇生物医学文章，最终获得了 1,000 条注释记录。此任务反映了一个真实的循证医学（EBM）环境，其中临床医生根据研究设计和质量对文献进行分类。
注意力的可视化呈现有多种格式，如第 2 节所述。在每篇文章之后，参与者回答以下问题：（1）在 1 到 5 的范围内，模型预测概率对分类这篇文章有多大帮助？（2）在 1 到 5 的范围内，摘要中高亮的词汇有多大帮助？

6 结果

我们分析了用户对模型解释的感知帮助性与预测概率和基于注意力的高亮词之间的关系——跨越不同文章类型和视觉编码。使用双因素 ANOVA 来探讨交互效应。
图 3 展示了文章类型和视觉编码之间的交互效应对突出显示词语的感知帮助性（即，基于注意力的解释）的影响。总体而言，用户对突出显示词语的实用性评分相对较低，平均在 5 分制中达到约 3.0 分。
我们发现一个显著的交互作用效应，这表明对解释的感知有用性取决于文章类型和使用的视觉编码。对于系统评价（SR），背景编码被评为比条形编码更有帮助（M=2.58，SD=1.15，而条形编码为 M=1.75，SD=1.09）。在广泛综合（BS）文

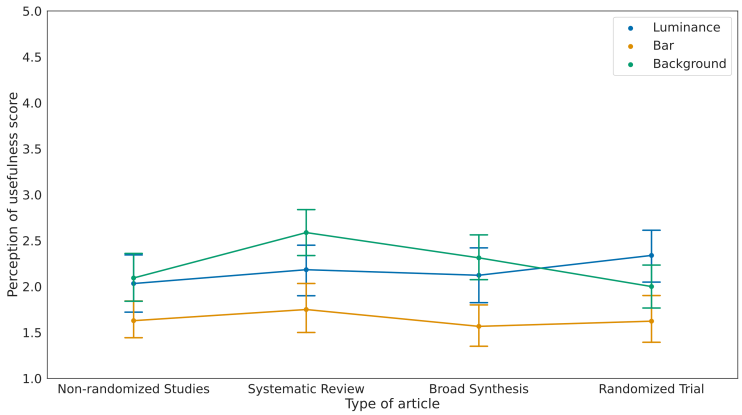

Fig. 3. 按可视化类型和文章类别感知到的加亮词的有用性。一项双因素方差分析显示，不同类型文档的背景颜色和亮度有用性评分存在差异，其中系统综述和广泛综合的评分较高。相比之下，条形编码在所有类型中一致被认为不太有用。

章中，亮度编码（M=2.12，SD=1.15）和背景编码（M=2.32，SD=1.14）都被认为比条形编码更有帮助（M=1.57，SD=0.89）。对于随机对照试验（PS-RCT），亮度编码（M=2.34，SD=1.14）的评分也显著高于条形编码（M=1.62，SD=1.04）。同样，在非随机对照试验（PS-NRCT）中，背景编码（M=2.09，SD=1.14）和亮度编码（M=2.03，SD=1.22）被评为比条形编码更有帮助（M=1.63，SD=0.82）。

这些结果表明，尽管条形长度是知觉上最精确的通道，但用户普遍觉得它不如背景或亮度等更直观的编码有用。这与可视化文献中预期的有效性层次结构相矛盾。

图 4 展示了第二次 ANOVA 的结果，分析了模型预测概率的感知有用性。与基于注意的解释相比，预测概率在所有文章类型中都被一致评为非常有用，平均得分超过 4.0。重要的是，我们发现视觉编码（包括无可视化条件）在预测概率感知有用性方面没有统计显著性差异。这表明尽管用户依赖概率信息，但他们对其效用的感知不受其他解释视觉表现的影响。总体而言，这些发现表明基于注意的解释对其视觉表现和基础内容的性质更为敏感。相比之下，预测概率无论在何种视觉上下文中都被认为始终有帮助。

Visual encoding	Mental	Physical	Temporal	Performance	Effort	Frustration
No visualization	46.1 (25.15)	25.3 (11.08)	44.2 (23.54)	61.6 (16.87)	49.70 (27.32)	27.8 (16.44)
Background color	37.2 (26.81)	24.4 (19.74)	36.5 (26.12)	55.3 (28.59)	42.4 (27.73)	30.3 (25.05)
Word luminance	49.1 (30.54)	35.1 (26.13)	49.4 (27.51)	50.6 (25.57)	54.3 (30.71)	43.3 (27.98)
Bar length	48.5 (24.28)	35.4 (22.82)	52.5 (23.58)	59.1 (15.58)	56.5 (24.95)	49.4 (21.66)

Table 2. NASA-TLX 子量表在视觉编码条件下的均值（标准差）。加粗的值表示最佳（最低工作负载或最高性能）分数。

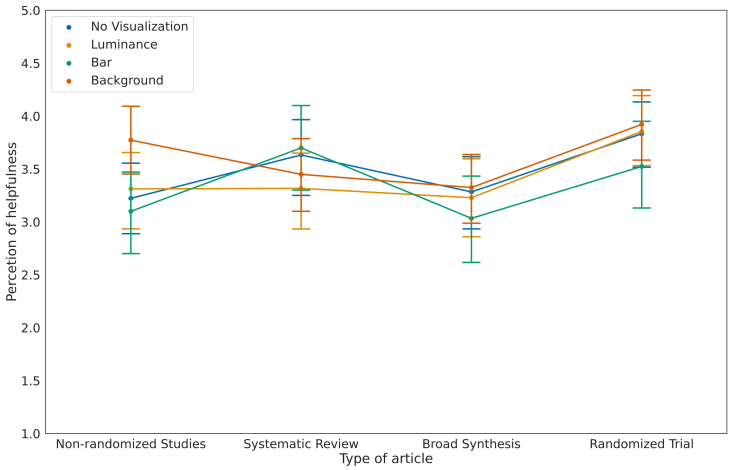

Fig. 4. 双因素方差分析用于分析文章类型与注意力可视化在模型预测概率感知有用性上的交互作用。结果表明，无论被分类的文档类型如何或者注意力是如何可视化的，显示模型的概率始终被认为是有帮助的。

表 2 显示了文本中不同注意力可视化方式的平均 NASA-TLX 评分。背景颜色条件与最低的心理需求（37.2）、体力需求（24.4）、时间需求（36.5）和努力（42.4）相关联，表明相比其他可视化方法而言总体认知负荷较低。尽管无可视化条件下的感知性能最高（61.6）且挫败感最低（27.8），但与背景颜色相比，也表现出更高的心理需求（46.1）和时间需求（44.2）。相比之下，柱状长度和单词亮度在所有工作负荷维度上得分更高，其中柱状长度显示出最高的挫败感（49.4）和努力（56.5）。

这些结果表明，背景颜色在可解释性和认知努力之间提供了一个良好的平衡。在无可视化条件下观察到的高性能和低挫折感可能反映了用户对传统 Epistemonikos 界面的熟悉程度，而更复杂或不熟悉的格式似乎增加了认知负荷。

7 结论

本研究评估了基于注意力的解释和预测概率是否支持医学专家分类生物医学文献。注意力权重通常被认为不是很有帮助，其有用性因可视化方式和被分类的文档类型而异。简单的可视化方式，如背景颜色，优于更精确但认知要求较高的方式，比如条形长度。相反，预测概率在所有可视化设置和文档类型中始终被认为有帮助。

NASA-TLX 的结果加强了这些发现：背景颜色与较低的认知负荷相关，而条形长度和亮度则增加了努力和挫折。在无可视化条件下的高性能和低挫折可能反映了用户对基于证据的医学标注标准界面的熟悉程度。

局限性包括仅关注单一领域、样本量小以及仅使用一种解释机制。

未来的工作将探讨交互式解释，并将研究扩展到更广泛的用户群体和医疗任务。

8

利益披露 作者声明他们没有竞争利益。

9

致谢 此项工作得到了 ANID Basal 基金、国家人工智能中心 CENIA FB210017、千禧科学计划 ICN2021_004（iHealth）、博士后 FONDECYT 资助 3240001 以及 FONDECYT 常规资助 1231724 的支持。

Bibliography

Vladimir Araujo, Andrés Carvallo, Carlos Aspillaga, Camilo Thorne, and Denis Parra. 2021. Stress test evaluation of biomedical word embeddings. *arXiv preprint arXiv:2107.11652*.

Vladimir Araujo, Andrés Carvallo, and Denis Parra. 2020a. Adversarial evaluation of bert for biomedical named entity recognition. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 79–82.

Vladimir Araujo, Andrés Carvallo, and Denis Parra. 2020b. Adversarial evaluation of bert for biomedical named entity recognition. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 79–82.

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. *arXiv preprint arXiv:2002.06261*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 258–262.

Santiago Canchila, Carlos Meneses-Eraso, Javier Casanoves-Boix, Pascual Cortés-Pellicer, and Fernando Castelló-Sirvent. 2024. Natural language processing: An overview of models, transformers and applied practices. *Computer Science and Information Systems*, (00):31–31.

Andres Carvallo and Denis Parra. 2019. Comparing word embeddings for document screening based on active learning. In *BIRNDL@ SIGIR*, pages 100–107.

Andrés Carvallo, Denis Parra, Hans Lobel, and Gabriel Rada. 2023. A comparative dataset: Bridging covid-19 and other diseases through epistemonikos and cord-19 evidence. *Data in Brief*, 51:109720.

Andres Carvallo, Denis Parra, Hans Lobel, and Alvaro Soto. 2020a. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125(3):3047–3084.

Andres Carvallo, Denis Parra, Gabriel Rada, Daniel Perez, Juan Ignacio Vasquez, and Camilo Vergara. 2020b. Neural language models for text classification in evidence-based medicine. *arXiv preprint arXiv:2012.00584*.

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*.

Izegbua E. Ihongbe, Shereen Fouad, Taha F. Mahmoud, Arvind Rajasekaran, and Bahadar Bhatia. 2024. Evaluating explainable artificial intelligence (xai) techniques in chest radiology imaging through a human-centered lens. *Plos one*, 19(10):e0308758.

Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 96–106.

Julian H Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian PT Higgins, Chris Mavergames, and Russell L Gruen. 2014. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2):e1001603.

Mor Geva, Tal Schuster, and Jonathan Berant. 2022. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

S Gopalakrishnan and P Ganeshkumar. 2013. Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *Journal of family medicine and primary care*, 2(1):9.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Doha Kim, Yeosol Song, Songyie Kim, Sewang Lee, Yanqin Wu, Jungwoo Shin, and Daeho Lee. 2023. How should the results of artificial intelligence be explained to users?-research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change*, 188:122343.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Stephen R Midway. 2020. Principles of effective data visualization. *Patterns*, 1(9).

D Parra, H Valdivieso, A Carvallo, G Rada, K Verbert, and T Schreck. 2019. Analyzing the design space for visualizing neural attention in text classification. In *Proc. ieee vis workshop on vis x ai: 2nd workshop on visualization for ai explainability (visxai)*.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will llms replace the encoder-only models in temporal relation classification? *arXiv preprint arXiv:2410.10476*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jesse Vig. 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *ICLR Workshop: Debugging Machine Learning Models*.

Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. 2024. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. 2023. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):262–272.