DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Evaluating and Benchmarking Attention Visualisation Tools for Interpretability Analysis of Open-Source Large Language Models

Akash Agarwal, Prof. Mohamed Abdelrazek

Background and problem statement

Understanding how Large Language Models (LLMs) make decisions is critical for building trustworthy AI systems. Attention mechanisms are the core computational component of transformer-based models, yet there's limited systematic research comparing different visualisation tools and their effectiveness at providing human-interpretable insights.

While tools like BertViz, Inspectus, and mattneary/attention exist for visualising attention patterns, there is no comprehensive benchmark that:

1. Compares these tools' capabilities and limitations across different model architectures
2. Evaluates which attention patterns correlate with specific model behaviours (reasoning, factual recall, hallucinations)
3. Translates raw attention matrices into actionable insights for model developers and researchers.

This project addresses the gap between technical attention analysis and practical interpretability by systematically evaluating visualisation tools across diverse task types and open-source models, ultimately establishing best practices for attention-based model interpretation.

Aim

To conduct a comprehensive comparative analysis of attention visualisation tools by benchmarking them against diverse open-source LLMs and task categories, producing human-interpretable insights that advance model interpretability research and establish evaluation standards for attention visualisation methods.

Objectives

- Survey and catalog existing attention visualization tools (BertViz, Inspectus, mattneary/attention, Transformer Explainer, etc.) with analysis of their features, supported models, and technical requirements.
- Establish a standardized benchmark framework for evaluating attention visualization tools across dimensions including: ease of integration, computational efficiency, visualization quality, and insight generation capability.

- • Curate a diverse evaluation dataset spanning multiple task categories:
    - Factual recall (simple facts, multi-hop reasoning, temporal queries)
    - Mathematical and logical reasoning (arithmetic, deduction, chain-of-thought)
    - Creative generation (storytelling, code generation, constrained writing)
    - Instruction following (task completion, constraint adherence)
- Test tools with multiple open-source LLMs of varying scales (1B-13B parameters) including Llama, Mistral, Gemma, Phi, and Qwen model families to assess cross-model compatibility.
- Collect and analyze attention patterns across 200-500 test cases, creating a structured dataset of attention visualizations with metadata (model, prompt category, attention statistics).
- Identify distinguishing attention signatures for different models and task types using quantitative metrics (entropy, locality, concentration ratios, token-specific behaviors).
- Translate technical findings into human insights by documenting patterns like:
    - How models handle ambiguity and context
    - Attention behaviors correlated with correct vs. incorrect outputs
    - Differences in reasoning strategies across model families
- Develop practical guidelines for researchers and practitioners on selecting appropriate visualization tools and interpreting attention patterns for model debugging and improvement.
- Document implementation challenges including memory requirements, processing time, integration complexity, and model compatibility issues.
- Produce open-source deliverables including benchmark code, curated dataset, analysis scripts, and comprehensive documentation for community use.

Expected outcomes

- Comprehensive tool comparison report documenting 5-8 attention visualization tools with feature matrices, performance benchmarks, and use-case recommendations.
- Open-source benchmark suite with standardized scripts for evaluating attention visualization tools across models and tasks, enabling reproducible research.
- Curated attention pattern dataset containing 200-500 annotated examples across diverse task categories with attention matrices, visualizations, and metadata in a structured, accessible format.
- Model attention signature profiles quantifying distinguishing characteristics of major open-source LLM families (Llama, Mistral, Gemma, Phi) across attention metrics.
- Human-interpretable insight documentation translating attention patterns into actionable findings such as:
    - Which models exhibit stronger locality vs. global attention
    - How attention concentration correlates with task performance
    - Attention patterns associated with hallucinations or reasoning failures

- Best practices guide for attention-based model interpretation including tool selection criteria, common pitfalls, and interpretation frameworks.
- Technical implementation guide covering integration steps, optimization techniques, and troubleshooting for each evaluated tool.
- Research publication or technical report suitable for submission to ML interpretability workshops (e.g., ICML Workshop on Interpretable ML) or publication on arXiv.
- Public GitHub repository with all code, data, documentation, and visualization examples to benefit the open-source AI community.
- Presentation materials including slides and demonstration notebooks showcasing key findings and practical applications.

How will this opportunity benefit you?

- Machine Learning Model Evaluation – Systematic assessment of model behavior through attention analysis, benchmarking methodologies, and comparative evaluation across architectures and scales.
- LLM Architecture & Interpretability – Deep understanding of transformer attention mechanisms, multi-head attention, layer-wise processing, and how architectural differences manifest in attention patterns.
- Data Science & Statistical Analysis – Collection, cleaning, and analysis of attention pattern datasets; statistical comparison of models; quantitative metrics design (entropy, concentration ratios, discriminative scores).
- Python Programming & ML Frameworks – Advanced use of PyTorch, Transformers library, NumPy for attention matrix manipulation; experience with visualization libraries (Matplotlib, Seaborn, Plotly).
- Research Methodology – Literature review, experimental design, hypothesis testing, reproducible research practices, and technical writing for ML research.
- Open-Source Tool Integration – Practical experience integrating multiple third-party libraries, handling compatibility issues, and working with diverse model formats and APIs.
- Data Visualization & Communication – Creating effective visualizations for complex technical data; translating technical findings into insights accessible to non-experts; documentation skills.
- Critical Thinking & Model Analysis – Developing intuition for model behavior, identifying patterns and anomalies, formulating and testing hypotheses about attention mechanisms.
- Software Engineering Best Practices – Code organization, version control with Git, documentation, testing, and creating reusable benchmark frameworks.
- Scientific Communication – Presenting technical findings through reports, documentation, presentations, and potentially academic papers; explaining complex concepts clearly.

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Other details

- Located on-campus at Burwood
- 14 weeks in duration (Unpaid)
- Undertaken as 3 days per week for a total of ~320 hours
- To be commenced within the 12-week trimester period