Transformer Explainer: Learning LLM Transformers with Interactive Visual Explanation and Experimentation

Aeree Cho
Georgia Tech
Atlanta, USA
aeree@gatech.edu

Grace C. Kim
Georgia Tech
Atlanta, USA
gracekim3@gatech.edu

Alexander Karpekov
Georgia Tech
Atlanta, USA
alex.karpekov@gatech.edu

Seongmin Lee
Georgia Tech
Atlanta, USA
seongmin@gatech.edu

Alec Helbling
Georgia Tech
Atlanta, USA
alechelbling@gatech.edu

Benjamin Hoover
IBM Research AI
Cambridge, USA
Georgia Tech
Atlanta, USA
bhoov@gatech.edu

Zijie J. Wang
Georgia Tech
Atlanta, USA
jayw@gatech.edu

Minsuk Kahng*
Yonsei University
Seoul, Republic of Korea
minsuk@yonsei.ac.kr

Duen Horng (Polo) Chau*
Georgia Tech
Atlanta, USA
polo@gatech.edu

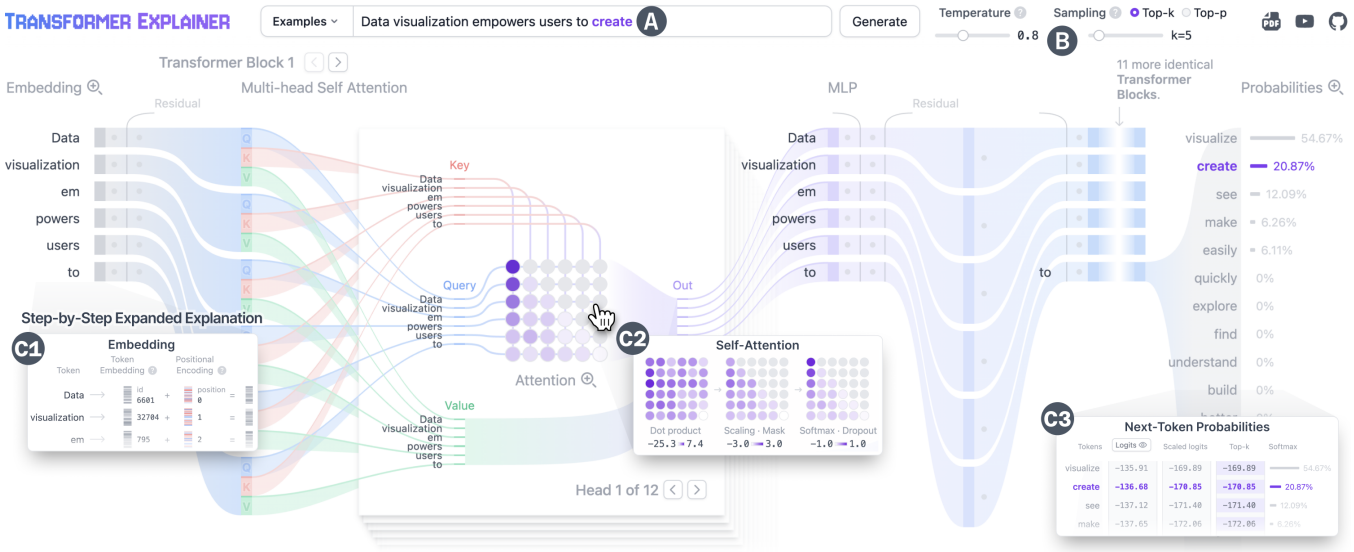


Figure 1: Transformer Explainer helps users (A) visually explore how a Transformer text-generation model (GPT-2) processes input text into a prediction for the next token, (B) interactively manipulate often-confused hyperparameters, such as temperature and sampling strategies, to understand their effects on prediction determinism; and (C) seamlessly transition between abstraction levels to visualize the interplay between high-level model structures and low-level mathematical operations for (C1) embedding, (C2) self-attention, and (C3) next-token probabilities.

Abstract

The Transformer architecture underpins modern large language models powering state-of-the-art text generation and AI applications. However, its complexity makes it difficult for non-experts to learn. Existing resources often lack interactivity, rely on static descriptions of simplified architectures, or fail to reflect models' behavior with real data. To address this gap, we introduce Transformer Explainer, an interactive visualization tool for non-experts to learn Transformers. The tool integrates an overview illustrating the Transformer's data flow with on-demand explanations that gradually reveal mathematical details. Smooth transitions across abstraction levels highlight the interplay between high-level structures and low-level operations. Running a live GPT-2 instance directly in the browser, Transformer Explainer empowers learners to experiment with custom input and hyperparameters without setup, observing next-token predictions in real time. A 90-participant user

*Corresponding author

This work is licensed under a Creative Commons Attribution 4.0 International License.

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791725

study showed that our tool offered significant advantages in improving user understanding and engagement. Transformer Explainer has attracted over 490,000 users.

CCS Concepts

• **Human-centered computing → Information visualization**.

Keywords

Deep Learning, Transformers, Visual Explanations, Interactive Experimentation, Visual Analytics, AI Education

ACM Reference Format:

Aeree Cho, Grace C. Kim, Alexander Karpekov, Seongmin Lee, Alec Helbling, Benjamin Hoover, Zijie J. Wang, Minsuk Kahng, and Duen Horng (Polo) Chau. 2026. Transformer Explainer: Learning LLM Transformers with Interactive Visual Explanation and Experimentation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3791725

1 Introduction

The Transformer [81] has become the state-of-the-art neural network architecture across diverse domains, including natural language processing and computer vision, and forms the backbone of large language models (LLMs) such as ChatGPT, DeepSeek, and Gemini. However, its complex internal structure poses significant learning challenges for non-experts, hindering their understanding and engagement [66]. Existing resources typically rely on static or non-interactive explanations that do not support experiential learning [1, 3] and may not fully reflect the model's behavior with real data [12]. Research from our HCI community has demonstrated the significant benefits of interactive visual explanations in empowering learners to more easily engage with and learn complex concepts [25, 41, 45, 73, 85]. For non-experts wishing to take their first steps toward understanding this technology, engaging and interactive explanations may be crucial for supporting such learning.

Key challenges in designing learning tools for Transformers. An increasing body of research is leveraging interactive visualizations to explain deep learning concepts. However, Transformers introduce unique challenges that require novel visual designs. For example, unlike earlier models such as Convolutional Neural Networks (CNNs), where operations can be made easier to understand through visualizations (e.g., filters sliding over input images [85]), Transformers operate on high-dimensional numerical representations of text that may be harder to understand. The mathematical operations performed on these representations add further complexity: Transformers consist of *many repeating blocks*, each containing many interacting operations (Fig. 2)—requiring learners to develop a mental model of not only how individual components work but also how they interact [23, 51]. Notably, the *multi-head self-attention* mechanism, unique to Transformers and present in every block [81], is significantly more complex than other model operations like CNN convolutions [66], as it involves multiple matrix operations that enable every input text *token* (typically a word, subword, or character; § 2) to simultaneously interact with *every other token*. Furthermore, Transformer prediction is *autoregressive* [54], which introduces additional challenges—each output token depends on all previously generated tokens, with key sampling hyperparameters influencing every step of the generation process. Therefore, the difficulty in learning about Transformers originates not only in the complexity of each component, but also in understanding the dynamic, high-dimensional, iterative, and probabilistic interplay between the large number of components. Existing learning resources have not adequately addressed such crucial learning needs, as they often rely on static descriptions of simplified architectures, lack real-time interactivity, or may not fully reflect the models' behavior with real data. We aim to bridge this critical gap.

In this work, we contribute:

(1) **Transformer Explainer, an interactive tool for non-experts to learn and experiment with Transformers for text generation** (Fig. 1). Transformer Explainer overcomes unique challenges in developing interactive learning tools for Transformers (§ 4), distilled through a literature review of visual learning resources and analytics tools for Transformers, and more broadly, deep learning (§ 3). Through iterative design and consultation with machine learning instructors (§ 7), Transformer Explainer offers statistically significant advantages over conventional learning resources (blog and video) in improving non-experts' understanding of Transformer concepts (§ 9).

(2) **Novel techniques and interactive system designs** to improve non-expert users' understanding of complex Transformer concepts.
- Transformer Explainer adopts a token-centric flow-based visual design, which is supported by recent studies that highlight the importance of tracing information flow through a Transformer model to understand its behavior [23, 51]. The flow-based design offers an overview that visually communicates the *token embedding data flow* across the model components, illustrating how *inputs* are processed and transformed across model components to reach the final output: the *next token* (§ 6.1).
- Our tool enables users to interactively expand the model components through animated transitions that preserve context and data flow, to explore *step-by-step* explanations of mathematical details (Fig. 1: C2). By smoothly transitioning between abstraction levels, users can see the interplay between low-level, detailed mathematical operations and high-level model structures, gaining a comprehensive understanding of Transformers (§ 6.2).
- Our tool allows users to input their own text (Fig. 1A) and directly manipulate sampling hyperparameters (Fig. 1B), enabling them to observe model behavior under different conditions in real time. Unlike existing educational resources—which often overlook how the generated probability distribution determines next-token predictions and how hyperparameters shape output variability [3, 12]—our interface visualizes these effects (Fig. 1: C3). For instance, while temperature is frequently anthropomorphized as a "creativity" control, users can experiment to see how it actually modifies the probability distribution and randomness of next-token predictions (§ 6.3).
- We introduce a guided learning feature—an interactive, step-by-step text explanation card embedded within the tool (§ 6.6).

Users can progressively learn Transformer concepts while linking them to visual components and interactive actions (e.g., adjusting hyperparameters and highlighting the resulting changes). Unlike conventional onboarding tutorials [75], guided learning supports both conceptual understanding and tool usage, serving as an in-situ explanation that can be accessed whenever additional clarification is needed.

(3) **Design lessons derived from a user study** on interactive visual explanation and experimentation. To evaluate Transformer Explainer's usability and effectiveness, we conducted a 90-participant between-subjects user study, which identified key advantages of the tool and confirmed its effectiveness in helping non-experts better understand complex model architectures and underlying operations (§ 8). Our findings highlight important lessons for future AI education tools, showing how interactivity enhances understanding, flow-based visualization clarifies complex architectures, and guided learning reduces entry barriers.

(4) **Open-sourced, web-based implementation powered by a live model** that broadens public access to our tool. Unlike many existing tools that require specialized software setups or lack inference functionality [10], Transformer Explainer hosts a live, fully-functional Transformer model that runs directly in the user's web browser. We selected the GPT-2 model for its widespread recognition, fast inference speed, and architectural similarities to more advanced models such as GPT-4 or later, making it suitable for educational use. Anyone can access Transformer Explainer directly in their browser without the need for installation or specialized hardware, allowing a large number of users to explore and learn from the tool simultaneously on their own devices. Our tool is open-source; a demo video, source code, and a live demo are included as supplementary material. Since its launch, Transformer Explainer has reached over 490,000 users across more than 200 countries and continues to contribute to the democratization of modern generative AI education.

2 Background for Transformers

In this section, we provide a high-level overview of a Transformer model architecture [81] (Fig. 2) in the context of text generation, which will help ground our discussion in the paper. Transformers like GPT-2 that are used to purely generate text are known as *decoder-only* Transformers, and they contain all the core components of the original Transformer [81]. Because Transformer Explainer focuses on explaining decoder-only Transformers, we use the term *Transformer* to specifically refer to decoder-only variants throughout this work.

To generate text, a Transformer performs the following processes: First, the input text is split into *tokens*: units of text that are typically words, subwords, or characters (e.g., "`Data visualization empowers`" → `["Data", " visualization", " em", "powers"]`). Each token is converted to a *token embedding* (i.e., its numerical vector representation), and positional encoding is added to preserve the order of tokens. This embedding passes through multiple Transformer blocks, each containing a *self-attention* mechanism with causal masking to prevent tokens from attending to future tokens in the sequence. Self-attention transforms each token embedding in a sequence of embeddings into a "query", "key", and "value", and the relationship between these three representations is best understood through an intuitive analogy: to predict the next token, the current token's *query* identifies which of the preceding tokens hold the most relevant information—specifically, the *values* associated with the most similar *keys*. The degree of similarity between a query and a key is quantified as the *attention score*. A single self-attention mechanism is known as an attention *head*. Stacking individual heads in parallel forms *multi-head self-attention*,[1] which enables the model to form a richer contextual understanding of the input, where each head can focus on different aspects of the input.

After the self-attention mechanism, the transformed embedding passes through an *MLP*, or multi-layer perceptron, which further increases the representational capacity of the Transformer. Across the multiple Transformer blocks, earlier blocks tend to capture low-level features, while later blocks represent more abstract semantic features. Finally, the model projects the transformed embedding into a probability distribution over all possible tokens, determining the likelihood of each token being the next in the sequence. During inference, sampling hyperparameters such as *temperature* control the sharpness or smoothness of the probability distribution, while sampling strategies such as *top-k* or *top-p* are used to select the next-token candidates. This process continues iteratively until the model fully generates a sequence of text.

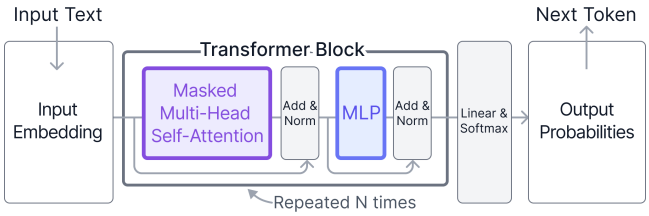


Figure 2: Diagram illustrating the data flow of the Transformer architecture (decoder-only), which consists of many components: input text is converted to input embeddings that pass through repeated identical Transformer blocks, each containing masked multi-head self-attention mechanisms, followed by multi-layer perceptron (MLP) layers with residual connections (Add) and normalization (Norm). The final linear and softmax layers output probabilities for the next-token prediction.

3 Related Work

3.1 Traditional Visual Learning Resources for Transformers

Most introductory resources for Transformer-based models have been offered as blog posts or video tutorials. A well-known example is The Illustrated Transformer by Jay Alammar [3], which provides static illustrations with explanatory text. Video tutorials are also popular, with some focusing on theoretical concepts [1] and others on code implementation [44]. While these resources reach a large audience, their static and linear formats often struggle to convey the input-dependent dynamic processes in Transformers, making it difficult for learners to understand them in detail and stay engaged

[1] Original architecture [81] proposed 6 Transformer blocks, and 8 heads.

in the learning process [25]. This challenge suggests a need for learning materials that support interactivity and real-time feedback.

3.2 Interactive Articles for Deep Learning Education

Beyond static blogs, interactive "explorable explanations" [82] emerged as a new medium for deep learning education. Following Chris Olah's interactive blog posts in 2014 [60], interactive articles with visualizations [13, 14, 62, 90] began gaining traction [85]. Distill, a scientific journal for machine learning, was also established to publish articles featuring interactive graphics and explorable explanations [29, 31]. While such articles provide more engaging learning experiences than static blogs, their interactivity remains limited: users typically follow predefined storylines and cannot freely manipulate inputs or explore alternative scenarios. Learners often need to scroll back and forth between overviews and detailed explanations, which disrupts continuity and makes it difficult to connect concepts across levels of detail.

3.3 Interactive Visual Tools for Deep Learning Education

More recently, to address the limitations of the traditional educational mediums mentioned above, fully interactive visualization tools have been developed for deep learning education. Early systems like ConvNetJS MNIST Demo and TensorFlow Playground offered intuitive parameter tweaking at the level of small models [42, 73]. Later, tools such as GAN Lab and CNN Explainer advanced the genre by introducing interactive graphics on real or synthetic data, bridging low-level mathematical details with high-level conceptual explanations [41, 85].

Recent work has proposed diverse "explainer" systems, but these do not converge on a single, settled design pattern. Instead, different model families and application contexts demand distinct visualization and interaction techniques. For example, to explain diffusion models, Diffusion Explainer lets users change various model parameters and observe their effects through timestep-based animations [47] and Patch Explorer visualizes internal representations as an interactive 3D heatmap [30]. GNN101 overlays a graph neural network as stacked visual layers and connects nodes across layers to depict layer-wise message passing [52]. Raise Playground uses a block-based programming interface to let learners practice and learn AI concepts hands-on [56]. PromptAid helps non-experts and practitioners explore, perturb, and test prompts for large language models with dashboard-style collection of visualizations [55].

To our knowledge, the only existing interactive visualization tools designed for learning about Transformers are LLM Visualization [12] and TransforLearn [28]. LLM Visualization offers a step-by-step guide through the Transformer architecture visualized in 3D, but it lacks support for custom user inputs and focuses on presenting mathematical details without abstraction levels, which may overwhelm beginners. TransforLearn supports interactive model exploration for machine translation but has key limitations that affect its educational impact. Its high-level overview displays heatmaps of embedding vectors, but otherwise relies on static text and diagrams that do not adapt to input changes, hindering user understanding [25]. Users can click on individual components to view separate visualizations, but the absence of continuous data flow makes it difficult to track how data transforms across the architecture. Moreover, the emphasis on embeddings in the overview may detract from learning self-attention, the core Transformer component [58, 81]. Finally, TransforLearn requires a server for machine translation tasks and is not deployed online. Transformer Explainer overcomes all the above limitations as the only online interactive visualization tool for learning Transformers, offering real-time inference with custom user input.

3.4 Visual Analytics Tools for Transformer Interpretability

While educational tools are designed to help non-experts learn about Transformer concepts, another large body of work focuses on visual analytics systems designed for researchers and practitioners to interpret and analyze the internal behaviors and computations of Transformer models. These systems typically target expert users, offering fine-grained inspection of attention patterns, hidden states, or neuron activations. A prominent line of work investigates attention mechanisms. AttentionViz provides global overviews of attention patterns across layers and heads in both language and vision Transformers [89]. Attention Flows introduces mechanisms to query, trace, and compare attention shifts during pretraining and fine-tuning [17]. Other systems extend this approach to specific domains, such as head-level analysis for Vision Transformers (ViTs) [49], cross-modal interactions in vision-language Transformers [2], and examines reasoning in Transformer-based Visual Question Answering (VQA) models [40].

Beyond attention, attribution-based systems offer complementary interpretability. VEQA analyzes open-domain QA with BERT by visualizing retrieval–reader decision flows [69]. Recent neuron-level methods further attribute knowledge in LLMs by identifying value and query neurons [91]. The family of Logit Lens approaches [6, 37, 59, 61] performs layer-wise decoding to show how next-token predictions evolve across intermediate layers, providing an alternative way to interpret the information encoded internally by Transformers. Additionally, work from Anthropic has advanced interpretability through interactive, explorable articles that enable users to analyze model components [4, 21, 50] and with tools like Circuit Tracer [32], which enables users to trace and visualize circuits formed by interactions between model components that contribute to specific model behaviors. These tools provide powerful interpretability for expert analysis but are not intended for non-experts, as they often require deep ML background knowledge to use effectively. Our goal, by contrast, is to help non-experts acquire this foundational background by introducing the core architectural concepts and data-flow processes that underlie Transformer models.

3.5 How Our Work Fills Unique Research Gaps

In summary, prior work for Deep Learning education can be grouped into (1) traditional static resources, such as blogs and videos; (2) interactive articles that guide readers through predefined narratives; (3) interactive educational visualization tools that support non-experts but do not address Transformers, while those that do are limited (§ 3.3); and (4) visual analytics systems for experts that provide in-depth interpretability but are not designed for beginners.

Transformer Explainer fills a unique gap between these categories. Unlike traditional or narrative-driven media, it provides real-time, interactive feedback on custom user input. Unlike existing educational tools for Transformers, it supports fully interactive, real-time exploration and complete data flow visualization, using levels of abstraction to introduce the entire model architecture while gradually revealing details to avoid overwhelming users. And unlike expert interpretability systems, it is designed to support non-expert learners, aiming to make complex architectures comprehensible. By combining these strengths, Transformer Explainer advances the landscape of educational resources for deep learning, offering a novel, fully online interactive system for learning Transformers at scale.

4 Design Challenges

Transformer-based LLMs introduce fundamentally different challenges across structural, visual, and computational dimensions that existing explainer systems designed for vision models or classification models [41, 42, 47, 73, 85] did not address. To design Transformer Explainer, we identified four main design challenges (C1-C4) related to Transformer models:

C1. Understanding How Input Text Is Processed Across Complex Model Structures. Transformers are complex models composed of many components, such as multi-head self-attention and multi-layer perceptron (MLP), repeated across multiple layers [27, 81] (Fig. 2). While existing resources have attempted to provide an overview of the model [3], they either present all details at once [12], which may overwhelm beginners, or display model components in disconnected views, often using disjoint visual encodings that increase users' cognitive load [79]. However, research shows that presenting all layer operations and their connections in a unified view has the potential to help users better follow how input data (i.e., token embedding) is transformed into final predictions. Such attempts typically use diagram-style layouts that place each layer or component as a boxed node and connect them with a line, which works reasonably well for vision models because each node can be visualized as an image [41, 42, 47, 85]. In contrast, Transformers operate on token representations and do not follow the relatively simple feedforward structure of classification models [42, 73, 85]. As token representations are interactively transformed as they go through attention, visualizing their transformations introduces unique challenges. There have been attempts to apply similar diagram-style approaches to Transformers, using component-level boxes connected by simple lines [28], but, they fail to foreground tokens as the primary and coherent visual element and cannot convey an end-to-end transformation path of how each token representation evolves throughout the model. Hence, an innovative visualization is needed to create a visual summary of the Transformer that maintains a connected view and preserves data flow. Adopting an information-flow perspective has potential to help non-experts conceptually link inputs, intermediate computations, and outputs into a coherent process [23].

C2. Mathematical Complexity in Multi-Head Self-Attention. Non-experts often struggle to understand the underlying operations in deep learning models [73]. In models like CNNs for images, operations can be more easily understood via visual metaphors—such as filters sliding over input images [85]. In contrast, Transformers operate on high-dimensional numerical representations of text, which are less interpretable on their own [58]. The complexity deepens with operations like self-attention, where every token interacts with every other token, and with attention-specific operations such as projecting embeddings into Q, K, and V vectors and partitioning them into multiple heads [5, 81]. This multi-head attention is foundational to how the model selects, gathers, and combines relevant information from different perspectives [81]. Understanding multi-head attention therefore helps learners understand the concrete mechanism in which words are used to predict the next token, demystifying how the model uses context. Existing resources typically explain these operations separately and in detail, but do not effectively help non-experts visually connect how the components work together to transform the token representations [3, 28]. Moreover, detailed mathematical explanations are often foregrounded, which can overwhelm non-expert learners with limited mathematical background and discourage further engagement. This gap highlights the need for more accessible explanation, motivating novel visualization and interaction techniques that can unpack and clarify these operations.

C3. Understanding Hyperparameters' Impact on Prediction Variability. Transformer models generate a probability distribution over possible tokens, from which the next token is sampled during autoregressive generation. Yet many educational resources either fix the prompt and the output, or they omit how the distribution is constructed and how sampling hyperparameters (e.g., temperature, top-k, top-p) reshape both the candidate set and the final choice [12, 28]. As a result, many learners remain unaware of these underlying mechanisms, often viewing Transformers as magical or even anthropomorphizing them [7]. The key challenge is therefore not only to explain the components that produce the pre-sampling distribution, but to make the entire sampling pipeline observable and learnable as it happens: how logits are formed, how they are transformed by sampling hyperparameters, and how a next token is selected from the resulting distribution. This demands a seamless integration of live inference, probabilistic visualization, and hyperparameter manipulation, which are absent in existing Transformer tutorials or earlier explainer tools.

C4. Deployment for Scalable Iterative Learning. Most educational resources for deep learning tend to rely on static content or provide limited interactivity (e.g., [3] for Transformer). This limitation stems largely from the technical challenge of hosting a live Transformer model in-browser—these models are large and computationally intensive, making it difficult to achieve the low latency required for real-time interaction [68]. As a result, existing tools tend to rely on pre-selected examples [12, 47] or offer restricted outputs [85], limiting educational opportunities for interactive hands-on learning, ultimately

hindering beginners from gaining deeper understanding and engagement.

5 Design Goals

Based on the design challenges, we distill four design goals (G1-G4) for Transformer Explainer, an interactive visualization tool to help non-experts learn and experiment with Transformer models.

G1. Model Overview Prioritizing Token-Centric Data Flow. We aim to create a visual summary of the Transformer architecture as a single, input token centered flow (**C1**). We visualize each token embedding as a one-dimensional heatmap that serves as the primary, consistently used visual element throughout the model. We animate this embedding as it travels along a single continuous path through the architecture, branching and merging in attention, expanding in the MLP, and passing through repeated blocks. Band width is proportional to embedding dimensions, directly revealing how the input evolves. This innovative design, the first for explaining Transformers, draws inspiration from the Sankey diagram [65], which is designed to communicate flow (visually encoded via its edges) between components (via nodes); we adapt it to emphasize how information is transformed within the model, building on recent studies that view Transformers as dynamic systems [20]. Our Sankey diagram-inspired design helps users see how input information "flows" through the various components and repeated blocks of the Transformer, undergoing successive processing and transformations before reaching the final output (§ 6.1).

G2. Visual Disambiguation of Multi-Head Self-Attention with Step-by-Step Visual Explanations. Given that the multi-head self-attention mechanism is considered the most important yet complex component of the Transformer [81] (§ 2), our goal is to visually unpack how the input embedding is projected into Q, K, and V, branches across heads, participates in attention, and subsequently rejoins through the path within a token-centered animated flow (**C2**), allowing non-experts to form an intuitive visual mental model. At the same time, we adopt a progressive disclosure technique [77] to support learners who want deeper detail. We first present the overall model structure, while leaving detailed mathematical operations to be revealed dynamically through user interaction (§ 6.2). In the detail view, we employ step-by-step animated visualizations to explain the underlying mathematical operations, presenting intermediate, successive steps converging toward a final output, inspired by prior research on algorithm visualizations showing that such steps help learners build mental models [24]. For example, in the self-attention, we aim to gradually reveal its intermediate steps through animations and visualizations, allowing users to hover over each matrix element to view its value interactively. For mathematical operations such as matrix multiplication, we aim to animate how matrices interact, enabling users to trace the computation of each output element (§ 6.2).

G3. Dynamic Experimentation Through User-Provided Text and Hyperparameter Manipulation. To help users understand how the next token is selected from the probability distribution generated by the Transformer (**C3**), we aim to run a live model in browser to support an interactive interface that allows real-time adjustment of sampling hyperparameters (§ 6.4) and to visualize the entire sampling pipeline with real-time intermediate values, so their influence on the next token selection is explicit and understandable. The user-provided text input is directly applied in the model's generation process, with their chosen hyperparameters used to predict the next token (§ 6.3). By experimenting with these settings, users can observe how the output becomes more predictable or more random. Predicted tokens are then appended to the input, allowing users to continue generating subsequent tokens and see how early sampling choices propagate through later predictions, while flow-preserving animations maintain token-level data flow across updates. This tight integration of live inference, probabilistic visualization, and hyperparameter manipulation helps users understand that the model is not magic, but rather follows a well-defined sequence of operations.

G4. Web-based Tool Powered by Live Model for Interactive Learning. To broaden the public's education access to our tool (**C4**), we aim to build a web-based application that hosts a live, fully-functional Transformer model. Users would directly interact with the model in their browsers, eliminating the need for installations or specialized hardware. Additionally, to encourage future research and educational use, we open-source our code. (§ 6.7)

6 Visualization Interface of Transformer Explainer

Transformer Explainer visualizes how a trained Transformer model transforms input text into probabilities for the next-token prediction. Users can explore the model at different levels of abstraction through *Overview* (§ 6.1) and *Step-by-Step Expanded Explanations* (§ 6.2). A live GPT-2 Transformer model running in the user's

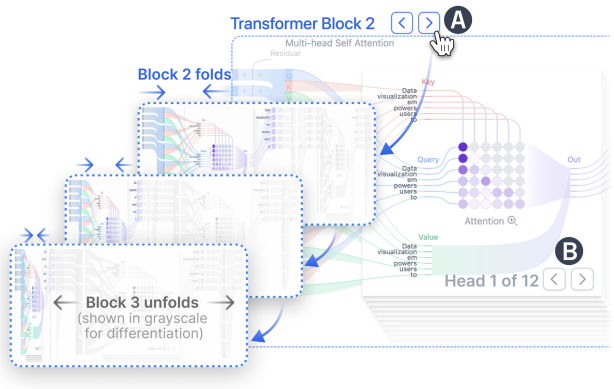


Figure 3: (A) Navigating between Transformer blocks triggers an animation that folds the current block's flow while unfolding the next. (B) Attention heads are depicted as a stack of cards, with transitions cycling smoothly through the stack in a looping animation. Buttons and annotations enlarged for clarity.

browser allows real-time experimentation with custom text inputs (§ 6.3) and sampling hyperparameters (§ 6.4), enabling users to immediately observe how these modifications influence the next-token prediction. Our system is targeted towards non-experts, visually guiding them through the mathematical operations underlying a Transformer model during text generation.

6.1 Overview

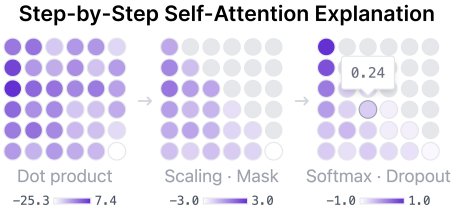
Transformer Explainer's visual design draws inspiration from the Sankey diagram, effective for visualizing data flow [65], to communicate a high-level overview of how input data flows through the Transformer model (**G1**; Fig. 1). Gradient-colored paths illustrate transformations of token embedding vectors across model components, allowing users to understand the model's structure from a data-flow perspective. Vectors are visualized as vertical bars scaled to their actual dimensionality, with a 1D heatmap revealed on hover. These visuals act as illustrative symbols that convey vector shape while achieving a balance between conceptual understanding and visual complexity. The color mappings are intentionally designed to reflect the role of each component. Token embeddings are rendered in grayscale to emphasize that they are the untransformed initial representations. Q, K, and V use distinct RGB-based colors to clearly differentiate their roles; for example, Q is blue and K is red, and the resulting attention scores appear in purple, visually reflecting the combination of the two. The MLP expands representations, so we preserve a consistent blue gradient to highlight continuity rather than introduce new semantic colors.

The dense Transformer architecture—consisting of many blocks, each containing multiple attention heads—is visually simplified by displaying only one selected block and one attention head at a time, reducing information overload [12]. Users can navigate between different blocks and attention heads using pagination buttons. To visually convey the presence of additional blocks and heads beyond the currently displayed one, we use visual metaphors and animated transitions. Repeated Transformer blocks are represented with gradually fading data flows between them, subtly suggesting continuity. When navigating between blocks, an animated transition smoothly folds the current block's flow while simultaneously unfolding the next block's flow into view (Fig. 3A). Attention heads within each block are represented as a stack of cards (Fig. 3B). Transitioning between heads triggers a looping animation, in which cards cycle through the stack. These animated transitions help keep viewers oriented and facilitate the perception of changes [80].

6.2 Step-by-Step Expanded Explanations

We design *Step-by-Step Expanded Explanation Views* to visualize a model component's internal computations, which involve multiple steps with intermediate results. For brevity, we call them *Expanded Views*. To avoid overwhelming users with details presented all at once (**G2**), our tool enables users to interactively open these views by clicking a component that has a magnifying glass icon next to its title. This triggers a smooth animated expansion of the component that preserves the high-level model structure while gradually fading out surrounding areas. This transition enables the details of the selected component to be presented, while maintaining the high-level context of data flow.

Self-Attention. The *Expanded Self-Attention View* animates the computation of attention scores in three sequential steps to help users understand how each mathematical operation transforms values into final attention scores, and to enable side-by-side comparison of intermediate results, which reduces cognitive load when tracking changes across steps [39]. When the user clicks Attention (Fig. 1: C2), the flows through the **key** and **query** come together to perform the dot product, producing an intermediate matrix. Next, we duplicate this intermediate matrix and present it next to the original one, and show the animation of how it is scaled and masked.



The same visual process is used to explain softmax and dropout, resulting in the final attention scores matrix. All matrix values are visualized using a heatmap with a purple color scale (e.g., −1.0 1.0). For any matrix output by the three steps, users can hover over individual elements to inspect their numerical values via tooltips.

Probabilities. The *Expanded Probabilities View* (Fig. 4A) incrementally displays each step in probability computation for next-token prediction, from left to right, following how values are transformed as the operation proceeds. When a user clicks Probabilities, which initially shows the final probabilities of next-token candidates (Fig. 1: C3), the expanded view opens to show each operation

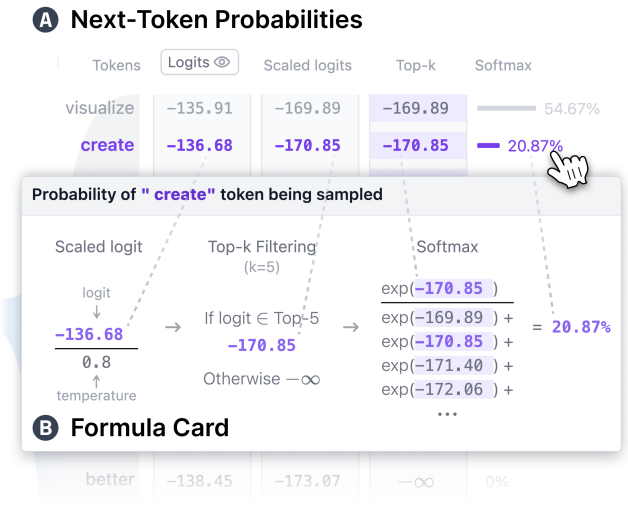


Figure 4: (A) A step-by-step expanded explanation view for next-token probabilities, showing the incremental steps in the probability computation. When a user hovers over a token, (B) the formula card updates to show the detailed equations used to compute the probability.

and its intermediate values: first, the logits (raw prediction scores) are computed from the final embedding; next, these logits are scaled by the user-selected temperature (Fig. 7A); then, the resulting values are filtered based on the user's selected sampling strategy (top-k or top-p) (Fig. 7B), retaining only the most probable candidate tokens—such hyperparameters are important for users to experiment with, as they often cause confusion (§ 6.4); and finally, probabilities for these tokens are computed using softmax. Tokens are listed by probability, with the most likely candidate at the top. Hovering over any intermediate value for a token reveals a formula card (Fig. 4B) showing the detailed equations used to compute the probability for that specific token at each step.

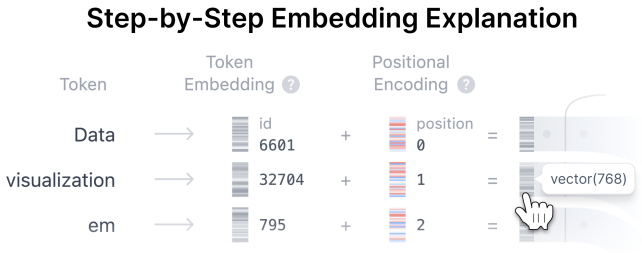


Figure 5: A step-by-step expanded explanation view for embedding visualizes how an input token is converted into its numerical embedding vector.

Embedding. The *Expanded Embedding View* (Fig. 5) illustrates how each token from the input text is converted into its numerical embedding vector. When a user clicks Embedding, the interface displays each token's predefined token ID and position in the sequence, along with its token embedding vector and positional encoding vector, which are summed to produce the final embedding vector that flows into the next component of the model. Users can inspect all three vectors (i.e., token embedding, positional encoding, summed embedding) as 1-dimensional heatmaps and hover over them to view their dimensions.

Animated Matrix Multiplications. Throughout the Transformer architecture, embeddings repeatedly undergo matrix multiplications with pretrained model weights, changing their dimensions, reflected in the Sankey diagram paths. Our tool provides a consistent visualization of these frequently occurring embedding-weight multiplications through popovers (Fig. 6), helping users recognize that these operations follow the same computational pattern. Clicking a data flow path between embeddings opens a view that displays the corresponding matrix calculations, animating how input embedding vectors are multiplied by weights to form new embeddings in transformed dimensions. Due to the high dimensionality, embeddings and weights are visualized as condensed heatmaps, with exact dimensions displayed below each matrix. These heatmaps serve as illustrative symbols of vector and matrix computations, maintaining consistent orientations and aspect ratios to convey tensor shape while keeping visual complexity manageable. Hovering over a specific element in the resulting matrix highlights the contributing inputs (Fig. 6D), clarifying the calculation details.

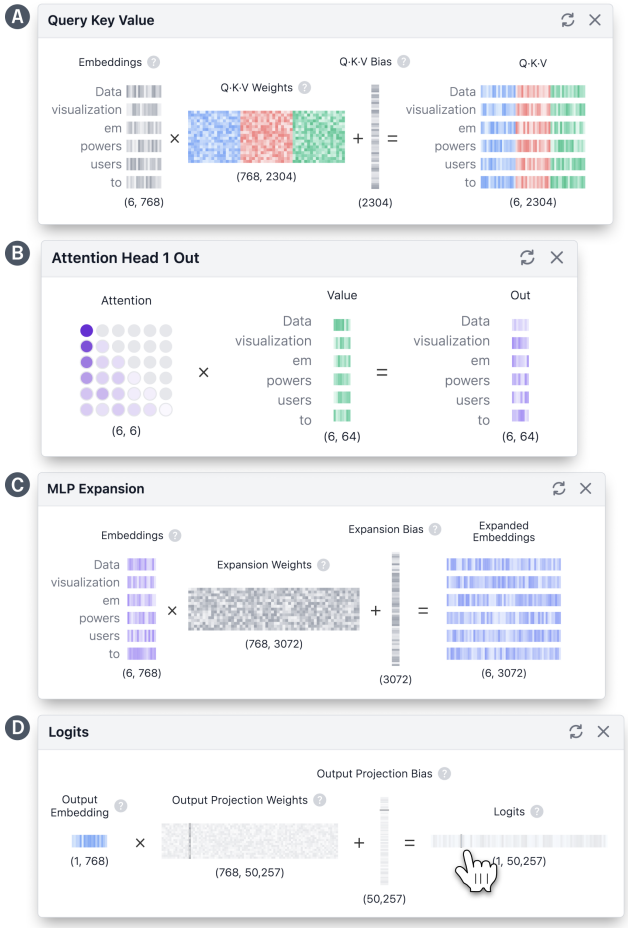


Figure 6: *Animated Matrix Multiplication* popovers provide an interactive visualization of embedding-weight multiplication operations within the Transformer architecture. (A) shows how token embeddings are linearly projected into query, key, and value vectors for attention computation. (B) visualizes the multiplication of the attention matrix with value vectors to produce the attention output. (C) shows the dimensional expansion of the embedding through a multilayer perceptron (MLP). (D) illustrates how the final embedding from the last Transformer block is multiplied by the output weight matrix to produce logits (raw prediction scores) for next-token prediction. Hovering over an element in the output vector highlights its contributing input elements, helping users understand how specific values are formed.

6.3 Real-time Inference for Next-Token Prediction

Users can enter custom text into the input bar and click Generate to observe in real-time how the next token is predicted. The input text is immediately broken into tokens and updated in the visualization; then, a smooth animation visualizes the updated data flowing through the model. The animation provides continuity between

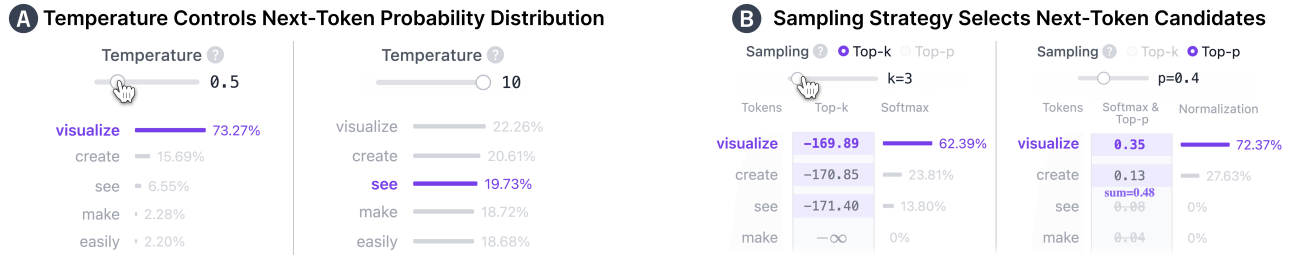


Figure 7: Transformer Explainer enables users to adjust inference hyperparameters and observe in real time how they affect next-token prediction. (A) The temperature slider lets users experiment with how temperature shapes the next-token probability distribution. A: left: Lower temperatures sharpen the distribution, making the output more deterministic. A: right: Higher temperatures flatten the distribution, increasing randomness and resulting in less predictable outputs. (B) The sampling strategy selector lets users choose between *top-k* and *top-p*, and adjust the corresponding k or p value using the slider below. B: left: *top-k* sampling with $k = 3$, where the model samples from the top 3 most likely tokens. B: right: *top-p* sampling with $p = 0.4$, where the model samples from the smallest possible set of tokens whose cumulative probability exceeds 0.4.

embeddings, clearly showing changes in position, size, shape, and color, helping users track data updates [35]. While this animation plays, a loading indicator appears in the input bar, and the predicted next token is appended once the data flow reaches the output.

Users can repeatedly click the Generate button to continue generating tokens, visually understanding how a Transformer model builds sentences one token at a time (**G3**). To support installation-free access, when the user visits our tool, the GPT-2 model (including its weights) is downloaded to the user's browser and runs entirely in it (§ 6.7). To maintain usability in high-latency or low-bandwidth environments where model loading may take time, five example prompts with pre-computed intermediate data extracted from models are provided, allowing users to explore the interface instantly.

6.4 Adjustable Sampling Hyperparameters Influencing Next-Token Prediction

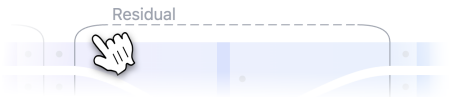
Transformer Explainer enables users to adjust inference hyperparameters and observe in real-time how these settings influence next-token prediction (**G3**). Users can modify temperature, sampling strategies, and their associated hyperparameters using the interactive controls located next to the Generate button (Fig. 1B). By testing their hypotheses and observing immediate feedback, users can see that Transformers select the next token based on probabilistic algorithms—not randomly or through "magic."

Temperature (Fig. 7A). The temperature hyperparameter shapes the generated probability distribution for the next-token prediction, making it sharper (lower temperature) or smoother (higher temperature). Users can adjust the temperature using a slider and test how it affects prediction determinism, understanding that temperature determines whether the output becomes more deterministic or random.

Sampling Strategies (Fig. 7B). We provide two widely-used sampling strategies: *top-k* and *top-p*. Users can select among sampling strategies using radio buttons and adjust the corresponding hyperparameters through sliders, observing how these hyperparameters influence which tokens are considered for the next prediction and the likelihood of each token being selected. Adjustments are reflected instantly, with the *Expanded Probabilities View* displaying how probabilities are computed based on the selected inference hyperparameters.

6.5 Auxiliary Architectural Features

We treat layer normalization, residual connections, activation functions, and dropout as supporting or conditioning mechanisms: they modulate and stabilize the primary computations (attention mixing and MLP transformations) and help preserve signal across depth, but are not the central conceptual steps we target for explaining how next-token probabilities are produced and sampled. This categorization was informed by consultations with machine learning instructors (§ 7), who noted that these mechanisms introduce additional mathematical detail that can overwhelm beginners without an ML background. To balance complexity, our tool visualizes these auxiliary features using visual scent [86]: dots represent layer normalization, dropout, and activations while lines indicate residual connections.



On hovering, residual connections are represented as dashed flowing lines, with animations illustrating the flow direction from their origin to their destination. Hovering over a symbol displays a brief explanatory tooltip, and users can click the *Read More* button to access a supplementary article located below the tool.

6.6 Guided Learning

Guided learning is an interactive text card (Fig. 8) that introduces Transformer concepts by following the flow of data, starting from the principle of autoregression and progressively covering the overall architecture and its main components—embedding, attention, MLP, and output probability.

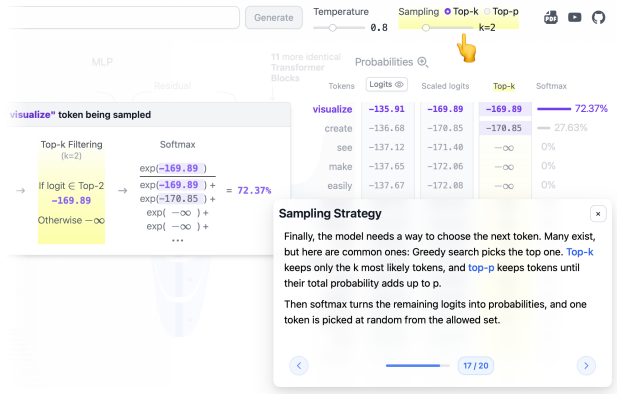


Figure 8: Guided learning provides an interactive, step-by-step text explanation that introduces Transformer concepts and connects relevant visual components. For example, on the *sampling strategy* page, it guides users to manipulate hyperparameters while simultaneously highlighting the elements that change in response on the screen.

At each step, users can follow the yellow finger icon to click dynamic elements to expand the view into detailed mathematical operations (§ 6.2), manipulate input text or hyperparameters (§ 6.3, § 6.4), and navigate across blocks and heads (§ 6.1). The visual elements directly related to the current guided learning page are highlighted in yellow, helping users focus their attention and connect concepts. Through this process, users not only learn the core concepts of the Transformer progressively and contextually but also naturally become familiar with the tool's interactive features.

Guided learning can be accessed at any time via the floating button located at the bottom right of the screen, and users can move directly to any page using the navigation buttons or the page dropdown 1 / 20 at the bottom of each card. In addition, guided learning also functions as a form of in-situ text explanation: when users hover over visual elements related to Transformer concepts, a help cursor appears, and clicking it opens the corresponding guided learning page. This design prevents the context switching where users would otherwise need to scroll down to the article at the bottom to view an explanation while freely exploring the tool.

The guided learning feature was introduced after a preliminary usability assessment conducted in the second phase of the tool's design iteration (§ 7.2), reflecting participants' feedback that an onboarding tutorial and in-situ text explanations could help lower the initial learning curve.

6.7 Web-Based, Open-Source Implementation

Transformer Explainer is a web-based, open-source visualization tool designed to help non-experts understand how Transformers work (**G4**). Users can access our tool using only a web browser, with no installation or specialized hardware required. We use a HuggingFace Transformers' [87] GPT-2 Small model from NanoGPT [43] to extract model data used in calculations during inference. To run the model in the browser, we converted a PyTorch model into ONNX format and used the ONNX Runtime Web API [19]. To minimize loading time, we split the model files into smaller chunks for parallel downloads and cache the model data in the browser using IndexedDB. As a result, the model only needs to be downloaded once, upon the user's first visit. The frontend is built with Svelte [33] and D3.js [8].

7 Informed Design Through Iterations

The current design of our tool is the result of over a year of iterative investigation and development, shaped by feedback gathered across three major phases.

- **Phase 1: Initial Prototype Feedback.** We built an early prototype and collected in-the-wild usage signals, complemented by informal feedback from instructors who regularly teach Transformer-related topics.
- **Phase 2: Enhanced Model Exploration.** Building on the initial feedback, we expanded the tool to support deeper exploration of model internals, enabling users to flexibly navigate the architecture (e.g., across Transformer blocks and attention heads (§ 6.1); and experiment with the output generation process through adjustable sampling hyperparameters (§ 6.4). We then conducted a preliminary usability assessment to gather early feedback on the tool's effectiveness as a learning resource.
- **Phase 3: Guided Learning Support.** Drawing on the preliminary feedback, we refined the design into its final form by adding guided learning scaffolds (§ 6.6). We then validated the resulting tool through a summative evaluation (§ 8).

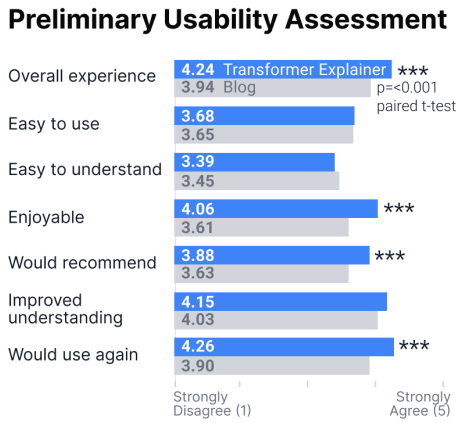
7.1 Phase 1: Initial Prototype Feedback

The first release of Transformer Explainer implemented our initial design goals (§ 5): a flow–based visualization (**G1**), step-by-step visual explanation (**G2**), and dynamic experimentation through user input and hyperparameter manipulation (**G3**). To manage early complexity, it rendered only the first block and first head—visually signaling repetition but not enabling traversal across blocks or heads. Following the release, we collected informal feedback from early users and three instructors who have taught graduate and undergraduate courses in machine learning and NLP topics (two instructors are co-authors). They noted that the multi-head and multi-block concepts remained abstract without the ability to systematically explore individual blocks and heads, and compare attention patterns; they also requested a more granular account of the output probabilities generation process that produces next-token distributions. These observations motivated the next iteration, emphasizing richer navigation across blocks and heads and an explicit, step-wise visualization of the final output probabilities layer.

7.2 Phase 2: Enhanced Model Exploration

We extended the tool to support navigation across attention heads and Transformer blocks (§ 6.1), and introduced a step-by-step output probabilities formula card (§ 6.2), including controls for sampling hyperparameters (§ 6.4). With these features in place, our tool was ready for a preliminary usability assessment to collect feedback on how its interactivity would benefit non-expert learners. We used

a within-subjects design to compare against a blog post baseline, which was static and non-interactive.



A large majority of the 145 study participants (74%) preferred Transformer Explainer over the blog, and rated our tool significantly higher on most usability items (as shown in the right figure), including *overall experience*, *enjoyment*, and *likelihood to recommend* or *use again*. These findings provided early evidence that interactive visual exploration and experimentation can better support non-expert learners than static materials.

7.3 Phase 3: Guided Learning Support

Through the preliminary usability assessment, we also identified opportunities to improve the tool. Transformer Explainer offers rich interactivity and dynamic visualizations; however, these same features sometimes created usability challenges. Some participants noted that the initial learning curve could be eased with a more guided onboarding experience. These observations align with usability survey responses, which indicated room for improvement on *"easy to use"* and *"easy to understand."*

Therefore, we introduced the guided learning (Fig. 8, § 6.6), an interactive, step-by-step text card that explains Transformer concepts while linking them to visual components and interactive actions. With these updates in place, we conducted a summative between-subjects evaluation to assess the tool's usability and usefulness, and to examine whether Transformer Explainer improves non-expert learners' understanding of Transformer concepts compared to blogs and videos, which are popular means of learning.

8 Evaluation: User Study

We conducted a user study to evaluate how effectively Transformer Explainer meets the design goals identified in § 5 and supports our high-level research contributions (§ 1). Specifically, we address three research questions (RQ1-3):

RQ1. How does Transformer Explainer improve non-expert learners' understanding of Transformer concepts compared to blog posts and videos, which are popular means of learning?

RQ2. How is learners' personal experience enhanced when learning Transformer concepts?

RQ3. How do the features of Transformer Explainer support effective learning?

In addition, we examined differences in usability, engagement, and perceived learning experience across educational resources.

8.1 Study Design

We conducted a controlled between-subject experiment to evaluate the learning effectiveness of Transformer Explainer in comparison to existing educational resources. Participants were randomly assigned to one of three conditions: Transformer Explainer, a blog post, or an educational video. A between-subjects design was chosen to minimize knowledge transfer effects across conditions that participants would otherwise experience if they were to go through the conditions in sequence. This study was approved by our institution's IRB.

The blog and video baselines were selected because they represent two popular modes of learning resources [16, 57, 70, 76]. Blogs are static narrative media, composed of text and figures that learners interpret at their own pace; whereas videos are multimodal media that integrate narration, animations, and other elements in a fixed sequence. Comparing our tool against these two formats allowed us to examine the added contribution of interactivity and dynamic visualization beyond static or multimodal presentation. For a fair comparison, we ensured that all three learning modes convey the same information. Specifically, we selected the blog post on decoder-only Transformer (GPT-2),[2] which is part of the blog returned as the top-ranked Google search result.[3] Similarly, we selected the Transformers video series by 3Blue1Brown[4] which has been viewed over 7.5 million times.

As the blog post and video contained content irrelevant to Transformer learning in our context (e.g., image generation model rather than text-based model), we consulted with the three instructors (§ 7.1) to identify six key learning objectives (Table 1) important for resources helping non-experts gain a conceptual understanding of Transformers. After removing irrelevant content, the learning resources could be fully explored in about 30 minutes by a learner.

8.2 Participants

We recruited participants from Prolific[5], an online user study platform, for a 1-hour study. The average completion time for the study was 43 minutes, and each participant received compensation of $12. To ensure the study targeted non-expert learners, we asked prospective participants to self-rate their familiarity with generative AI on a 5-point scale and to indicate whether they were interested in learning how text-based generative AI works. Only individuals who both expressed interest and reported low familiarity levels were eligible to participate (i.e., either 1: *Don't know what it is*; 2: *Heard of it only*; or 3: *Aware but don't understand*). 90 of them completed the study, balanced across the three experimental conditions. (§ 8.3.1 describes the approach for checking participant engagement.)

Overall, the participants spanned a wide range of educational backgrounds, disciplines and industries. In detail, their educational backgrounds included bachelor's degree (35.6%), master's degree (18.9%), high school (16.7%), and some college (13.3%). Participants

[2] https://jalammar.github.io/illustrated-gpt2/
[3] For search terms such as "Transformer explained" and variations
[4] https://youtu.be/wjZofJX0v4M
[5] https://www.prolific.com

Learning Objective (LO)	Quiz Question
LO1: How GPT-2 generates text one token at a time	Q1: How does a Transformer generate text?
LO2: The overall Transformer architecture	Q2: Which structure matches a text-generation Transformer?
LO3: Text to Embedding transformation	Q3: Which is the best description of tokens and embeddings?
LO4: Multi-head Self-Attention mechanism	Q4-1: In self-attention, what do Query, Key, and Value do? Q4-2: Why use multiple heads in attention?
LO5: MLP (feed-forward network)	Q5: Why add a MLP (feed-forward) layer after attention?
LO6: Final probabilities and sampling parameters	Q6: How do "top-k" or "temperature" affect text generation?

Table 1: Learning objectives (LO) and their corresponding multiple-choice quiz questions.

also spanned diverse disciplines, including computer science and information technology (30.0%) and business/management (23.3%), as well as social sciences, arts and humanities, and health sciences. Industry backgrounds were varied, with notable representation from health care (14.4%), information services (13.3%), technical services (7.8%), and finance (7.8%).

8.3 Procedure

After providing informed consent, participants first completed a **demographic and background survey**, which included self-ratings of mathematical proficiency and generative AI familiarity. They were then randomly assigned to one of three experimental conditions: interactive tool, blog, or video (referred to as Transformer Explainer, Blog, and Video, respectively). Participants studied their assigned resource freely for up to 45 minutes, enough time to fully explore the material (§ 8.1) and pace their learning while keeping exposure timing comparable across conditions.

Next, participants completed the post-study evaluation, which included both objective and subjective measures.

- For objective assessment, participants took a **closed-book multiple-choice quiz** with 7 questions (Table 1) closely aligned with the learning objectives identified by instructors (§ 7.1), who also helped review the questions. The quiz questions focused only on concepts shared across all three resources, and the quiz items and answer choices were identical for all conditions.
- For subjective assessment, participants (1) reported their **self-perceived understanding** of Transformer concepts on a 5-point Likert scale and **rated the learning experience** of their assigned material; and (2) answered **three open-ended reflection questions** ("most helpful aspects," "most confusing aspects," and "suggested improvements") to capture qualitative feedback. Participants in the Transformer Explainer condition also provided tool-specific **evaluations of individual features**.

8.3.1 Checking for participant engagement. To safeguard the integrity of the study results, we employed a two-pronged verification to ensure participants had used the tools and engaged both meaningfully and honestly. First, participants were required to answer three simple resource-specific questions that anyone who had studied the resource could easily answer (e.g., "What is the temperature range that can be adjusted with the slider in the tool?"). Only those with at least two correct responses were included, serving as both an engagement check and a verification of minimal attention. Second, we excluded participants showing signs of inattention or dishonesty, such as unrealistically short quiz times ($\leq$ 7.1 seconds per question, below the 5th percentile) or insufficient effective tool-use time ($\leq$ 5 minutes of focused engagement, discounting tab-outs or early exits). Using this two-pronged process, we excluded 38 of 128 initial participants, resulting in a final sample of 90 motivated non-experts.

8.4 Data Analysis

We employed a mixed-methods approach to analyze the study data, combining quantitative and qualitative techniques to address our research questions. The quantitative analyses (§ 8.4.1) examined participants' objective quiz performance and subjective ratings for achieving the learning objectives (Table 1) to evaluate the effectiveness of each learning resource in supporting Transformer concept understanding (**RQ1**). We also asked participants to rate their personal experience using each material (**RQ2**). In addition, Transformer Explainer participants provided feature-usefulness ratings, allowing us to investigate how they engaged with the tool and to uncover usage patterns associated with successful learning (**RQ3**). To complement these measures, we conducted a qualitative thematic analysis (§ 8.4.2) of participants' open-ended responses to capture deeper insights into their experiences, sources of confusion, and suggestions for improvement.

8.4.1 Quantitative Analysis.

Quiz Accuracy. Quiz accuracy was measured with a 7-question multiple-choice quiz aligned with six learning objectives (Table 1). We analyzed quiz accuracy using generalized linear mixed models (GLMMs) [11] with a binomial logit link. Condition (Transformer Explainer, Blog, Video) was entered as a fixed effect, and participant was modeled as a random intercept to account for repeated responses per participant. Participants' self-reported math proficiency and generative AI familiarity were included as covariates to control for prior knowledge, but neither showed reliable effects ($p > 0.20$). Quiz question was initially included as an additional fixed effect, but since its effect was not significant ($p > 0.30$), we report models focusing on condition-level differences. For significant condition effects, we conducted planned one-sided contrasts

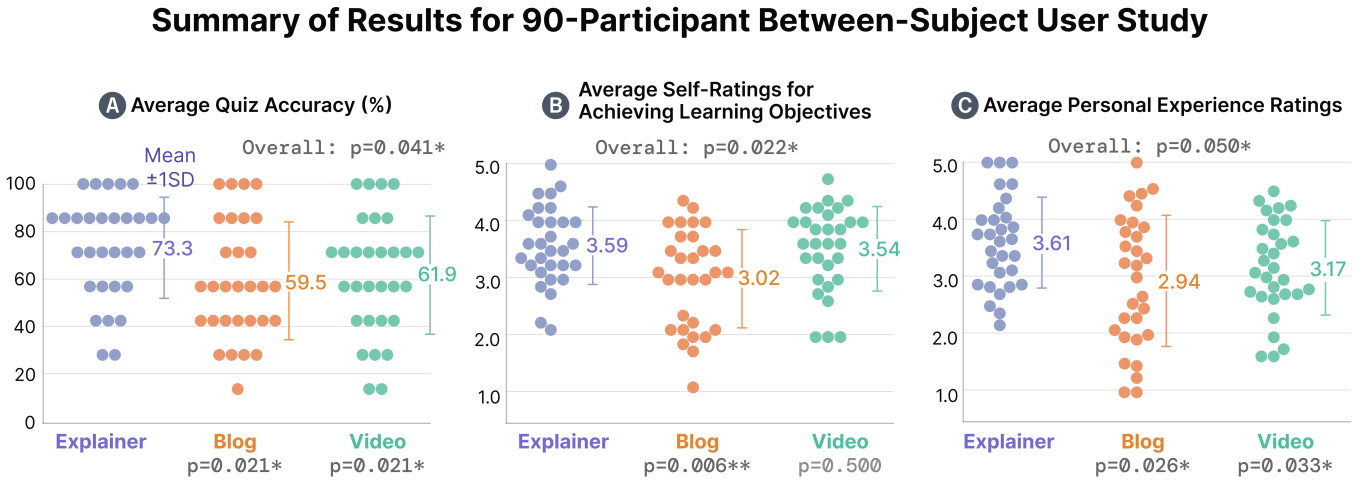


Figure 9: Transformer Explainer was significantly more effective than both Blog and Video in understanding Transformer concepts and learning experience. (A) Transformer Explainer participants achieved higher quiz accuracy than Blog ($p = 0.021$) and Video ($p = 0.021$), as confirmed by pairwise tests from a GLMM controlling for participants' prior math and AI knowledge. (B) They also reported higher achievement of learning objectives than Blog ($p = 0.006$), and (C) higher learning experience ratings than Blog ($p = 0.026$) and Video ($p = 0.033$), on follow-up pairwise comparisons after Kruskal–Wallis tests.

(Transformer Explainer > Blog, Transformer Explainer > Video), applying Holm correction across the two comparisons [38].

Subjective Ratings. Participants reported their self-perceived understanding of each learning objective (Table 1) on a 5-point Likert scale, as well as overall learning experience including usability, engagement, clarity, and self-efficacy. Cognitive load was measured on a 0–10 scale. Because the data were ordinal and Shapiro-Wilk tests [71] indicated violations of normality, we used non-parametric Kruskal-Wallis tests [46] to examine group differences. Planned one-sided contrasts (Transformer Explainer > Blog, Transformer Explainer > Video) were tested with Holm correction across the two comparisons [38]. **Usability** was measured with the UMUX-Lite scale for *ease of understanding*, and *usefulness* [48]; **engagement** with measures adapted from the Intrinsic Motivation Inventory for *enjoyment*, *interestingness*, *attention-holding* [48]; and **mental demand** with a measure from NASA-TLX [34]. We also measured **clarity** (*"I could follow the explanations without much confusion"*) and **self-efficacy** (*"I feel confident that I can explain the basics of a Transformer"*). Internal consistency for the two *usability* measures and the three *engagement* measures was high, with Cronbach's $\alpha = 0.82$ and $\alpha = 0.88$ respectively [15], supporting the use of averaged scores across measures.

8.4.2 Qualitative Analysis. Participants also responded to three open-ended questions: (1) what aspects of the material were most helpful, (2) what aspects were confusing, and (3) what improvements they would suggest. We analyzed these responses using thematic analysis following Braun and Clarke's six-phase approach [9]. We conducted open coding to identify meaningful units of text, then iteratively grouped codes into candidate themes. Through refinement, we developed a final codebook that captured recurring patterns across responses. Representative quotes for each theme were selected to illustrate the findings.

9 Findings and Reflections

Here, we present the findings for our three research questions: § 9.1 discusses how our tool improves non-expert learners' understanding (**RQ1**); § 9.2 describes how learners' personal experience is enhanced (**RQ2**); and § 9.3 discusses how our tool's specific features support effective learning (**RQ3**). Finally, § 9.4 reflected on our lessons learned on user needs, explanation effectiveness, and learning outcomes.

9.1 How does Transformer Explainer improve non-expert learners' understanding of Transformer concepts compared to blog posts and videos, which are popular means of learning? (RQ1)

9.1.1 Summary. Our analyses present converging evidence that our tool improves non-experts' understanding of Transformer concepts more effectively than Blog and Video (Fig. 9), showing **statistically significant** advantages in both *objective* quiz accuracy and *subjective* ratings in achieving learning objectives. Fig. 9A shows Transformer Explainer participants answered statistically significantly more quiz questions correctly (73.3% correct on average across 7 questions) than those in Blog ($p = 0.021$) and Video ($p = 0.021$). They also self-rated their understanding as significantly higher than Blog ($p = 0.006$), and comparable to Video (Fig. 9B).

9.1.2 Question and learning objective level analysis. Transformer Explainer's benefits are further highlighted by analyses at the quiz question and learning objective level. Fig. 10A shows that participants using Transformer Explainer achieved significantly higher quiz accuracy than Blog on the *output probability* (LO6), and marginally higher than both Blog and Video on the *attention mechanism* (LO4)—concepts often regarded as the main reasons

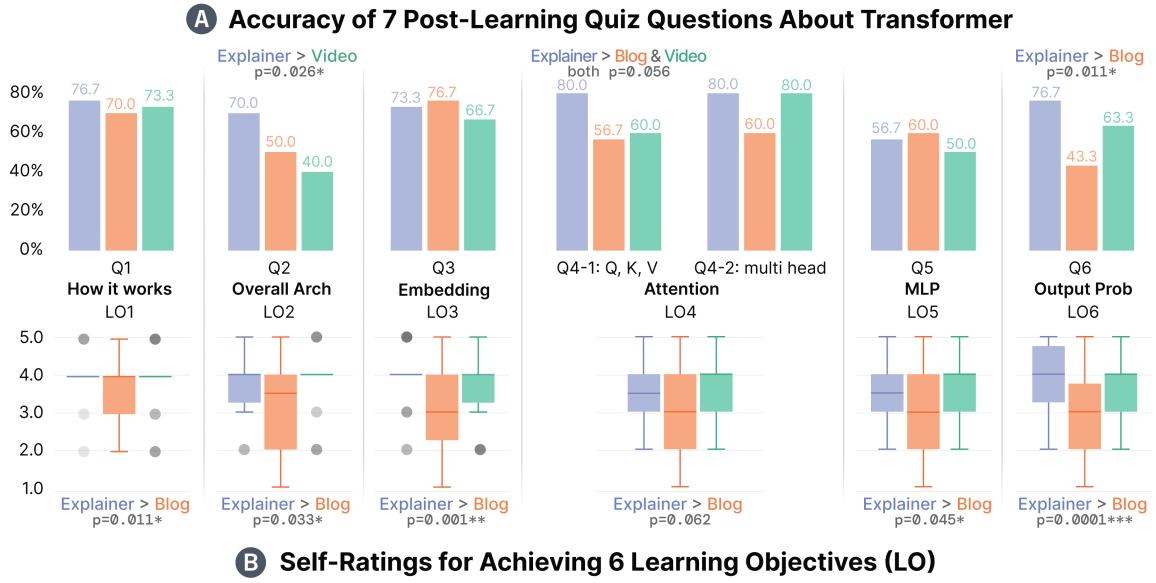


Figure 10: Across all learning objectives, Transformer Explainer led to significantly higher understanding scores than both Blog and Video. (A) Across the seven quiz questions aligned with six learning objectives, Transformer Explainer participants scored higher on *architecture overview* (LO2) than Video ($p = 0.026$); marginally higher on *attention mechanism* (LO4) than both Blog and Video (both $p = 0.056$); and significantly higher on *output probability* (LO6) than Blog ($p = 0.011$). (B) For self-reported achievement of the learning objectives, Transformer Explainer participants rated themselves significantly higher than those in Blog across all items.

of difficulty when learning Transformers. They also outperformed Video on the *architecture overview* (LO2), a topic that requires a clear understanding of the model's overall structure.

Transformer Explainer participants self-rated significantly higher in achieving all learning objectives (Fig. 10B) compared to Blog, except for the *attention mechanism* (LO4) where the difference was marginally significant. While participants in the Video condition reported understanding levels comparable to Transformer Explainer, their overall quiz accuracy, averaged across questions, was as low as Blog's, and significantly lower than Transformer Explainer's (Fig. 9A). This pattern may reflect an "illusory understanding" effect [67], where prior research suggests that watching a seemingly coherent narrative (e.g., a video) can create the impression of understanding without supporting recall [18]. Indeed, one Video participant rated their understanding 4 out of 5, but admitted: *"When I got to the quiz, I didn't really remember the details. I think maybe short quizzes along the way [would help]"* Another noted, *"it was not confusing, but hard to retain. I am a visual person and remembering without the video is difficult for me."* Several others ($n = 3$) explicitly suggested adding quizzes or interactive activities during video viewing to improve retention.

Interestingly, Transformer Explainer participants commented on how the tools provide the type of interactive elements that are "missing" from Video. One participant explained, *"I like how it is interactive and allowed me to change certain settings. I find interactive learning helps me better understand and retain compared to only reading."* Another noted, *"I really found the interactive visuals useful in aiding comprehension while I was doing the quiz."*

9.2 How is learners' personalized experience enhanced when learning Transformer concepts? (RQ2)

Fig. 9C shows that Transformer Explainer participants rated their overall personal experience statistically significantly higher than both Blog ($p = 0.026$) and Video ($p = 0.033$). Breaking this down by measures (Fig. 11), Transformer Explainer consistently outperformed the baselines across multiple measures. Specifically, on **usability** (*ease of understanding*, *usefulness*) and **engagement** (*enjoyment*, *interestingness*, *attention-holding*), our tool scored significantly higher than both Blog and Video. **Self-efficacy** ratings were also significantly higher than Video ($p = 0.047$), suggesting that learners felt more confident in explaining Transformer basics after using Transformer Explainer. As one participant noted: *"I feel that I could map out the process [...] the equivalent of a children's drawing of the solar system in crayon. I understand that everything moves about one another."* On **mental demand**, Transformer Explainer was rated statistically significantly lower than Blog ($p = 0.044$), while being comparable to Video. This result is noteworthy because interactive tools are sometimes associated with higher cognitive load due to user actions [72, 74], whereas videos are typically regarded as cognitively lightweight since learners only need to follow the narrative [22]. We attribute this outcome to two design features: a flow-based visualization with visual consistency (**G1**), and step-by-step guided explanations (**G2**) that scaffolded learners through abstraction levels appropriate for non-experts (§ 9.3).

For clarity, all three tools' scores were modest, reflecting the inherent challenge of the Transformer topic for non-experts with little

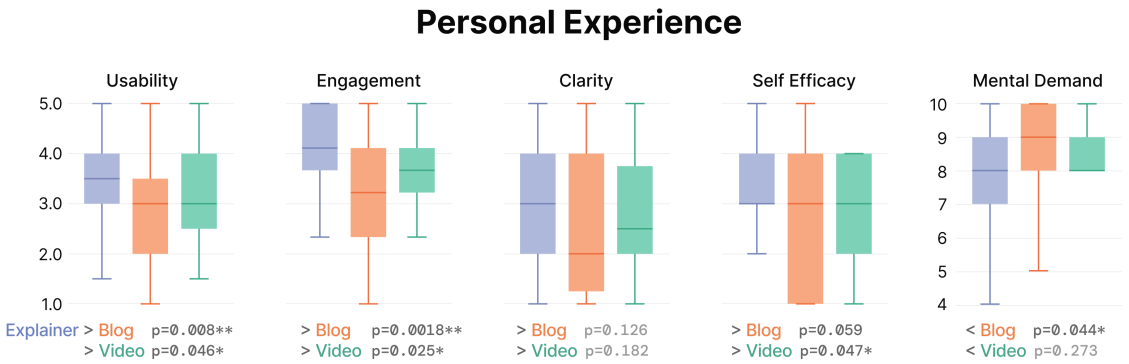


Figure 11: For participants' personal experience using the three tools, Transformer Explainer outperformed both Blog and Video on all measures. Usability and engagement were significantly higher than both baselines, self-efficacy was significantly higher than Video ($p = 0.047$), and mental demand was substantially lower than Blog ($p = 0.044$).

prior background knowledge. Nonetheless, Transformer Explainer participants experienced the highest average clarity (at 3.07), while reported the best ratings across all other measures. Taken together, these findings demonstrate the viability of interactive visualization approaches in enhancing participants' personal experience when learning Transformer concepts.

9.3 How do the features of Transformer Explainer support effective learning? (RQ3)

Our results suggest that the effectiveness of Transformer Explainer may be attributed to its success in achieving three design goals that we identified (§ 5): enabling dynamic experimentation with live models through user input and hyperparameter manipulation (**G3**), maintaining coherence with flow-based visualizations (**G1**), and providing step-by-step guided explanation (**G2**). Each of these features contributed to higher engagement and lower mental demand, and together they led Transformer Explainer to achieve stronger overall usability evaluations (Fig. 9C).

9.3.1 Interactivity Enhances Understanding of Transformer Concepts. Unlike the Blog and Video, only Transformer Explainer allowed participants to directly manipulate inputs and model hyperparameters while observing immediate system feedback (**G3**, § 6.3, § 6.4). In addition, the step-by-step expanded view enabled learners to transition between overview and detail at their own learning pace (**G2**, § 6.2). These features not only contributed to an enhanced personal experience (Fig. 11) but were also strongly associated with improved participants' understanding. In open-ended responses, 26 of 30 participants **explicitly** identified interactivity as **one of the most helpful features** for achieving the learning objectives. Many specifically highlighted entering their own text and adjusting sampling parameters as particularly impactful. One participant noted, *"The most helpful part to learn is the interactiveness with the generate button, being able to add my own words to play around, as well as the probabilities of the words coming next being able to play with the sample parameters."* Another remarked, *"The material that helped me the most was the interaction piece, looking at how temperature, p and k change with the different words and how much creativity I want it to have."*

9.3.2 Token-centric Flow-based Visualization Clarifies Complex Model Architecture. Although all three learning materials relied heavily on visual explanations, the flow-based visual design of Transformer Explainer offered participants a clearer and more coherent experience. As shown in Fig. 12, flow visualization was rated as the second most helpful feature, and in open-ended responses, participants praised its clarity in mapping data trajectories onto the actual model structure ($n = 12$). One noted, *"I really liked seeing the flow visuals of the transformer. I finally have a clear mental image for how it all works."* and another, *"The graphic was very efficient, and made me understand more clearly what was happening at each step."* Such comments highlight how this design translated abstract operations into tractable narratives, helping users build a clearer mental model of the Transformer's internal architecture.

In contrast, Blog and Video participants described visual overload. One participant commented, *"The visuals were somewhat helpful, but they could be confusing as well. Too much information at once, including technical terms and graphics."* while another noted, *"Have it be less columns and charts. My eyes just glaze over and attention drifts after the first ten to fifteen minutes of that."* This suggests that Transformer Explainer reduced participants' mental demand by maintaining a consistent Sankey-style flow across components and integrating overview and detail within a single framing (**G1**, § 6.1), thereby avoiding the introduction of new charts or diagrams for each subtopic as in the Blog and Video—as reflected in the lower mental demand scores compared to the baselines (Fig. 11).

These findings demonstrate the importance of a consistent visual narrative that spans the entire learning process. Rather than offering many disparate visual elements, representing complex concepts through a unified visual language reduces learners' cognitive burden and facilitates deeper understanding.

9.3.3 Guided Learning Reduces Entry Barriers and Improves Clarity. As shown in Fig. 12, the guided learning feature of Transformer Explainer (§ 6.6) was rated the highest among all features, as "extremely" or "very" helpful by most participants. Several participants emphasized how the structured walkthrough supported their learning. One noted, *"I relied a lot on the guided tour so that I could track the progress from start to finish. The interactive design was pretty*

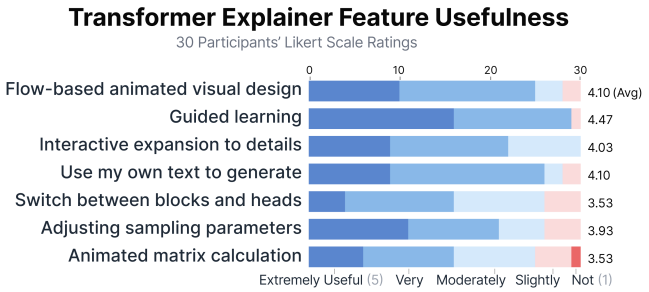


Figure 12: Participants rated most features of Transformer Explainer as useful for learning. The highest ratings were given to *guided learning*, *flow-based animated visual design*, and *use my own text*, with a majority of participants rating them as "extremely useful" or "very useful." Other features also received generally positive responses.

effective with keeping things moving left→right, and the 1–20 stepped process was essential for tracking over the loops." and another, *"The part that helped most was the box on the bottom right corner. It broke down each concept one by one and had helpful navigation on the site."* These scaffolds directly addressed issues observed in the preliminary usability assessment (§ 7.2), where some participants reported not knowing "what to look at first." Now, such comments have virtually disappeared.

9.4 Reflections on User Needs, Explanation Effectiveness, and Learning Outcomes

9.4.1 [User Needs] Transformer Explainer promotes active learning by combining visual narrative with interactive exploration. Reflecting on our design process and study results, we identify several needs of non-expert learners trying to understand Transformers: the need to work with personally meaningful inputs, to experiment with model behavior in a low-stakes way, and to receive immediate feedback that makes abstract mechanisms concrete. Active learning research suggests that when learners work with inputs they have personally chosen, they develop a stronger sense of stake in the outcome, leading the brain to treat the information as more valuable and thus increasing their readiness to learn and retain [18, 26]. Transformer Explainer directly supports these needs by allowing participants to (1) choose their own prompts and settings and immediately observe the resulting attention patterns; (2) interactively modify temperature and sampling hyperparameters while seeing how such changes alter the next-token distribution. This tight experiment–feedback loop turns parameters that are often discussed only in the abstract into manipulable objects, in ways that passive media such as Video or Blog cannot easily provide.

Transformer Explainer's higher quiz accuracy (§ 9.1) on the more challenging topics such as *attention mechanism* (LO4) and *output probability* (LO6), both of which supported by rich interaction, suggests that addressing these needs for hands-on experimentation and immediate feedback can translate into better understanding of model internals. At the same time, participants in the Video condition frequently reported difficulty retaining material and expressed a desire for additional mid-lesson activities to reinforce key concepts, underscoring that learners may need more than continuous exposition as they learn. Transformer Explainer addresses this need in part through interactive exploration and guided learning, and future work could explore adding more structured practice activities. Overall, our findings highlight the value of tools that blend visual explanations with opportunities for active engagement and experimentation, and point to user needs (e.g., personal relevance, controllable pacing, and concrete feedback) that future AI education tools should explicitly support.

9.4.2 [Explanation Effectiveness] Transformer Explainer eases non-experts' interpretation of complex Transformer mechanisms. Participants using our tool rated **usability** (ease of understanding and perceived usefulness) significantly higher than Blog and Video, and reported higher **self-efficacy**, indicating greater confidence in their ability to explain core Transformer mechanisms after interacting with the system. At the same time, **mental demand** was rated lower than in both baselines, counter to the common assumption that interactivity necessarily increases cognitive load. This finding suggests that our interactive explanations helped participants make sense of the model's internal processes without feeling overwhelmed, rather than burdening participants with interface complexity. Taken together, these three self-reported measures indicate that participants experienced the explanations as both *accessible* and *empowering*. This subjective picture is echoed by the objective quantitative results (§ 9.1). Specifically, quiz accuracy averaged 73.3% across seven learning-objective-aligned questions (Fig. 9), a statistically significant improvement over Blog and Video. In other words, participants did not only feel that the explanations were clear; they were also better able to answer mechanism-focused questions correctly. In the next section, we examine these learning outcomes in more detail and discuss how they may be attributed to specific design choices in our tool's explanatory features.

9.4.3 [Learning Outcomes] From Design Features to Learning Gains. Our evaluation suggests that Transformer Explainer not only improves overall quiz performance and self-rated understanding, but does so in a way that varies across learning objectives in informative ways. Across the seven quiz questions aligned with six learning objectives, participants using Transformer Explainer achieved statistically significant gains on *architecture overview* (LO2, $p = 0.026$), and *output probabilities* (LO6, $p = 0.011$), and a marginally higher performance on *attention mechanism* (LO4, $p = 0.056$), outperforming Blog and often Video on these challenging concepts. These patterns indicate that our tool is particularly effective at helping learners construct a coherent mental model of the model's global structure and the probabilistic nature of text generation, areas where non-experts often default to a "black box" view of Transformers.

These outcome differences closely track where our design offers the richest interactive support. Architecture overview (LO2) is reinforced by the flow-based visualization and block navigation (§ 6.1), which let learners move between overview and detail on demand; attention mechanism (LO4) is supported through head-level navigation (§ 6.1) and step-by-step animated explanations (§ 6.2); and output probabilities (LO6) are made tangible through direct experimentation with temperature, top-k, and top-p (§ 6.4).

In contrast, MLP (LO5)—the only objective without dedicated interactive elements—shows comparatively lower quiz accuracy and self-rated understanding (Fig. 10). This divergence suggests our interactive visual scaffolds are not merely engaging add-ons but are tightly coupled to conceptual learning, and it highlights MLP as a concrete target for future work (e.g., by adding analogous visual and experimental affordances).

Finally, comparing subjective and objective outcomes reveals how different resource formats shape learners' sense of understanding. As noted in § 9.1, participants in the Video condition reported understanding comparable to Transformer Explainer, despite substantially lower quiz accuracy. Prior work characterizes this kind of divergence as possible "illusory understanding" effect, where coherent narrative can foster confidence without supporting recall [18]. Building on this, our results suggest that Transformer Explainer's interactive manipulation of inputs and parameters may help learners calibrate their understanding more accurately: by actively testing how changes in inputs, attention, or sampling parameters affect model behavior, participants receive immediate feedback about what they do and do not yet understand. For a learning-outcomes perspective, this calibration is itself valuable, even though our current measures focus on short-term conceptual gains rather than long-term retention or transfer.

10 Discussion, Limitations, and Future Work

Our user studies provide promising evidence of the effectiveness of interactive visualizations for helping non-expert learners understand Transformer models. At the same time, our evaluations surfaced important challenges and opportunities for further refinement. In this section, we discuss current limitations of our tool and outline future directions for broadening and deepening interactive visual explanations for AI education.

10.1 Extending Support for Other Transformer Modalities

While Transformer Explainer currently focuses on text-based Transformers, we observed from our user study that participants showed significant interest in seeing the tool expanded to domains beyond language (e.g, *"I would like to learn about how the model behaves when its used for generating images. How does it differ from this one?"*). Many participants expressed curiosity about how the same architectural principles manifest in other domains, particularly multimodal and non-linguistic applications. Extending support to models such as Vision Transformers (ViT) for vision [92], Whisper for speech [64], or vision-and-language models like CLIP and SigLIP [63, 78] would highlight that Transformers are not limited to text, but operate as a general-purpose architecture across diverse modalities [88]. Our flow-based visual design (§ 6.1) and step-by-step explanations (§ 6.2) could naturally generalize to these architectures, though they would also introduce new challenges. For instance, tasks such as image generation may require specialized visualization techniques from prior work on diffusion models and visual interpretability [47, 53], including dimensionality-reduced visual summaries of generation trajectories or saliency-based overlays for visual attention. Exploring how these methods might be integrated into our framework presents an exciting direction for future development.

At the same time, expanding the conceptual framing beyond the current *word = token* representation offers another avenue for broadening the tool's reach. Our current walkthrough primarily employs text-generative Transformers—the domain most familiar to general audiences—to explain tokenization and embedding using user-provided sentences (§ 6.3). However, Transformers fundamentally operate on discrete token representations, which need not correspond to linguistic words. Protein sequences, gameplay logs, and other forms of structured data can all be represented as tokens and processed in the same way as text. Making this abstraction explicit, and offering interactive examples that illustrate how non-linguistic token streams are embedded, transformed, and attended to, could help learners more deeply appreciate the universality of the architecture.

Taken together, these directions suggest that Transformer Explainer could evolve into a cross-domain explainer of Transformer architectures. By combining modality-specific visualization techniques with a token-centric framing, the tool could transcend its role as a text-model explainer and become a platform that communicates the generality of Transformers across a wide range of applications. Such an expansion would not only benefit researchers and practitioners in various domains, but also help non-experts recognize that systems capable of generating text, interpreting images, recognizing speech, or even modeling biological sequences all rely on the same architectural building blocks. We see this as an opportunity to move from "explaining GPT-2" toward "explaining the Transformer paradigm" more holistically.

10.2 Deepening and Scaling Up Transformer Explanations

In our study, participants expressed interest not only in exploring larger models and extended contexts (e.g., GPT-4), but also in obtaining deeper insights into core mechanisms like attention. Supporting these interests requires tackling complementary challenges: enriching interpretive depth while simultaneously scaling visualization strategies for larger and more complex models.

10.2.1 Deepening Attention Interpretation and Visualization. Non-expert learners in particular expressed strong interest in understanding the meaning and role of attention, indicating that deeper interpretation could support more comprehensive conceptual understanding. As one participant noted, *"I'd love to see how multiple heads collaborate in making predictions. Do some heads agree on key words, or do they specialize and work independently?"* Designing visualizations that make the functions of multi-head attention clearer—how different heads capture varying complexities or attend to distinct aspects of the input—could enhance learning outcomes. A fruitful direction is to reflect that attention has no privileged basis in query–key (or value–projection) space [4] by finding a rotation that best groups high-activation neurons, making visual patterns easier to interpret.

While existing attention visualization tools [49, 83, 84, 89] have largely been designed for researchers, future works could carefully adapt such ideas for non-experts, enabling users to interactively

explore head relationships and block-level behaviors while preserving ease of use. Beyond attention, additional opportunities lie in visualizing advanced behaviors such as in-context learning, showing how prepending a few examples changes attention patterns and alters predictions on a new task. These expansions could help learners move beyond introductory explanations and explore both the fundamental mechanisms and higher-level capabilities of modern Transformers.

10.2.2 Supporting Larger Models and Extended Contexts. Recent Transformer models increasingly handle longer inputs and richer contextual information. Our user study echoed this interest in educational tools for exploring large-scale models, such as GPT-4. However, a long input prompt with many tokens could quickly increase visual complexity, even at the overview abstraction level. Additionally, running such large models directly in web browsers remains technically challenging [47]. Future research may therefore explore specialized visualization strategies alongside efficient browser-based implementations (e.g., WebAssembly optimization [68], model compression), making large-scale Transformer models accessible to broader audiences.

10.3 Limitations of Study Design

While our study highlights the potential of interactive visualization tools for teaching complex AI concepts, it also carries several limitations. First, we conducted a one-hour user study and capped participants' tool usage at 45 minutes (§ 8.1). This controlled environment allowed fair comparisons across conditions, but may not fully reflect the most natural learning contexts for all participants. In practice, learners might spend more than an hour exploring, revisit concepts after a break, or adopt other study rhythms. Such interaction patterns may only emerge over extended periods of use.

Second, our evaluation measured learning outcomes immediately after the session using quizzes and surveys (§ 8.3). While this design captures short-term comprehension, it does not assess long-term retention or transfer. In this study, we intentionally focused on whether the tool improves non-expert learners' understanding of high-level Transformer concepts, and therefore excluded detailed assessment of mathematical mechanisms. However, such fine-grained understanding may only be measurable after learners have had time to consolidate and re-apply the knowledge. Future research may investigate longitudinal effects, evaluating retention and application weeks or even months later.

Third, our participant pool primarily consisted of non-expert learners with limited prior exposure to Transformers (§ 8.2). This choice was appropriate for studying how the tool supports beginners who are interested in AI but lack formal ML expertise. Nonetheless, perspectives from advanced learners, instructors, and practitioners remain important for future work. Experts and educators might use the tool differently—for example, as a teaching aid, a debugging resource, or a way to illustrate abstract concepts in the classroom. Their feedback would provide valuable insights into how the tool can extend to broader educational contexts. Moreover, studies with ML students could compare the tool against lecture materials or textbooks to evaluate whether interactive visualization also aids in understanding finer-grained mathematical mechanisms.

Finally, we compared Transformer Explainer against some of the most popular baseline resources—a widely-read blog post and a highly-viewed YouTube video (§ 8.1). Even against these well-known materials, our tool demonstrated clear advantages. Still, the variety of existing blogs and videos is much broader. Expanding the range of baselines in future comparisons would help further validate and contextualize our findings.

10.4 Positioning Our Work in the AI Education Tool Landscape

We are the first to visualize all blocks and components of a Transformer as a token-centric data flow (**G1**), showing that multi-head self-attention, often viewed as the steepest learning hurdle, can be visually unpacked within a unified graphical language (**G2**). We further address a core Transformer characteristic that many prior educational tools overlook: its autoregressive and probabilistic behavior. To do this, we run a live Transformer model directly in the browser, extract intermediate attention values and output probabilities in real time, and integrate them seamlessly into the visualization. Coupled with animations that reveal the model's internal computation flow, our system enables an immediate experiment–feedback loop that responds to user edits of the input and sampling hyperparameters (**G3**). We validate the effectiveness of this approach through a user study, showing meaningful learning benefits for non-experts without an ML background (§ 8).

Beyond the formal study, we observe substantial educational impact in real-world use. Since releasing the initial prototype, the tool has been used by over 490,000 users across over 200 countries and adopted as course material in undergraduate and graduate ML-related courses at leading universities worldwide. Some instructors report appreciating the ability to use the tool with large classes without server constraints, and both instructors and students value being able to run models live and explore how different hyperparameters shape model behavior. In addition, we open-sourced the system to broaden accessibility and extensibility. As a result, the community has released versions of Transformer Explainer in multiple languages and begun adapting it to additional models.

Taken together, our work contributes a novel design that combines token-centric flow visualization with interactive, in-browser model experimentation, and demonstrating both its educational effectiveness and its potential for broad adoption and extension.

10.5 The Role of Interactive Visualization in Advancing AI Education

Our study demonstrates that interactive visualization tools can provide significant benefits for AI education (§ 9). By allowing learners to directly engage with model components and observe the effects of their interactions, such tools transform learning from passive knowledge acquisition into active, exploratory engagement. In designing our tool, we drew inspiration from earlier examples [41, 42, 73, 85], which pioneered visual experimentation with machine learning concepts. At the same time, we acknowledge that developing high-quality educational visualizations can require substantial time and effort. For example, our iterative design process (§ 7) extended over a year of refinement and evaluation.

To mitigate these challenges, recent efforts have explored modularizing visualization components to reduce development effort and accelerate the creation of educational tools for emerging AI architectures (e.g., ManimML [36]). We view this approach as a promising direction: by lowering barriers to tool development, the community can expand the availability of interactive learning resources and better align educational tools with the rapid pace of advances in AI models. We are excited to see how the HCI, visualization, and AI education communities will continue contributing to this effort, building on our work to make AI concepts more accessible and understandable to diverse learners.

11 Conclusion

We presented Transformer Explainer, an interactive visualization tool aimed at helping non-experts understand a text-generative Transformer model. Our tool provides a seamless transition between a data flow-based overview visualization and detailed step-by-step explanations. Users can directly interact with a live GPT-2 model in the browser, experimenting with custom input text and hyperparameters. Results from a user study indicate improved understanding in Transformers and user engagement.

Acknowledgments

This work was supported in part by NSF awards 2403297 and 2502793, the IITP (MSIT, Korea) grant RS-2024-00353131, and gifts from Google, Amazon, Meta, NVIDIA, Avast, Fiddler Labs, Bosch. Alec Helbling is supported by NSF GRFP.

References

[1] 3Blue1Brown. 2024. But what is a GPT? Visual intro to transformers. https://youtu.be/wjZofJX0v4M.

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 21406–21415.

[3] Jay Alammar. 2018. The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer/.

[4] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread* (2025). https://transformer-circuits.pub/2025/attribution-graphs/methods.html

[5] Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and M. de Rijke. 2019. Understanding Multi-Head Attention in Abstractive Summarization. *ArXiv* abs/1911.03898 (2019). https://api.semanticscholar.org/CorpusID:207853291

[6] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112* (2023).

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 610–623. doi:10.1145/3442188.3445922

[8] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[10] Adrian M. P. Braşoveanu and Răzvan Andonie. 2020. Visualizing Transformers for NLP: A Brief Survey. In *24th International Conference Information Visualisation (IV)*. 270–279.

[11] N. E. Breslow and D. G. Clayton. 1993. Approximate Inference in Generalized Linear Mixed Models. *J. Amer. Statist. Assoc.* 88, 421 (1993), 9–25. http://www.jstor.org/stable/2290687

[12] Brendan Bycroft. [n. d.]. LLM Visualization. https://bbycroft.net/llm.

[13] Chen Chen, Jinbin Huang, Ethan Remsberg, and Zhicheng Liu. 2024. A Visual Tour to Empirical Neural Network Robustness. https://cchen-vis.github.io/Narrative-Viz-for-Neural-Network-Robustness/.

[14] Matthew Conlen and Fred Hohman. 2018. The Beginner's Guide to Dimensionality Reduction. https://visxai-dimensionality-reduction-1dbad0a67a092b007c526a45.vercel.app/. In *1st Workshop on Visualization for AI Explainability (VISxAI)*.

[15] Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (1951), 297–334. doi:10.1007/BF02310555

[16] Carol Azumah Dennis. 2015. Blogging as public pedagogy: Creating alternative educational futures. *International journal of lifelong education* 34, 3 (2015), 284–299.

[17] Joseph F DeRose, Jiayao Wang, and Matthew Berger. 2020. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1160–1170.

[18] Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* 116, 39 (2019), 19251–19257.

[19] ONNX Runtime developers. 2021. ONNX Runtime. https://onnxruntime.ai/. Version: x.y.z.

[20] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. 2021. Redesigning the Transformer Architecture with Insights from Multi-particle Dynamical Systems. In *NeurIPS*. https://proceedings.neurips.cc/paper/2021/file/2bd388f731f26312bfc0fe30da009595-Paper.pdf

[21] Nelson Elhage et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021).

[22] Enqi Fan, Matt Bower, and Jens Siemon. 2024. Video Tutorials in the Traditional Classroom: The Effects on Different Types of Cognitive Load. *Technology, Knowledge and Learning* 29, 4 (Dec. 2024), 2017–2036. doi:10.1007/s10758-024-09754-1

[23] Javier Ferrando and Elena Voita. 2024. Information Flow Routes: Automatically Interpreting Language Models at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17432–17445. doi:10.18653/v1/2024.emnlp-main.965

[24] Eric Fouh, Monika Akbar, and Clifford A. Shaffer and. 2012. The Role of Visualization in Computer Science Education. *Computers in the Schools* 29, 1-2 (2012), 95–117. arXiv:https://doi.org/10.1080/07380569.2012.651422 doi:10.1080/07380569.2012.651422

[25] Eric Fouh, Monika Akbar, and Clifford A Shaffer. 2012. The role of visualization in computer science education. *Computers in the Schools* 29, 1-2 (2012), 95–117.

[26] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences* 111, 23 (2014), 8410–8415.

[27] Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics* 9 (2021), 1061–1080.

[28] Lin Gao, Zekai Shao, Ziqin Luo, Haibo Hu, Cagatay Turkay, and Siming Chen. 2023. Transforlearn: Interactive visual tutorial for the transformer model. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 891–901.

[29] Gabriel Goh. 2017. Why Momentum Really Works. *Distill* (2017). doi:10.23915/distill.00006

[30] Imke Grabe, Jaden Fiotto Kaufman, Rohit Gandikota, and David Bau. 2025. Patch Explorer: Interpreting Diffusion Models through Interaction. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*.

[31] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. 2019. A Visual Exploration of Gaussian Processes. *Distill* (2019). doi:10.23915/distill.00017

[32] Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. circuit-tracer. https://github.com/safety-research/circuit-tracer. The first two authors contributed equally and are listed alphabetically..

[33] Rich Harris and Svelte Contributors. 2016. Svelte: Cybernetically enhanced web apps. https://svelte.dev/

[34] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[35] Jeffrey Heer and George Robertson. 2007. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1240–1247.

[36] Alec Helbling and Duen Horng Chau. 2023. ManimML: Communicating Machine Learning Architectures with Animation. *arXiv preprint arXiv:2306.17108* (2023).

[37] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation

decoding in transformer language models. *arXiv preprint arXiv:2308.09124* (2023).
[38] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 2 (1979), 65–70.
[39] Jooyoung Jang, Christian D Schunn, and Timothy J Nokes. 2011. Spatially distributed instructions improve learning outcomes and efficiency. *Journal of educational psychology* 103, 1 (2011), 60.
[40] Theo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Visqa: X-raying vision and language reasoning in transformers. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 976–986.
[41] Minsuk Kahng, Nikhil Thorat, Duen Horng (Polo) Chau, Fernanda B. Viégas, and Martin Wattenberg. 2019. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics* (2019).
[42] Andrej Karpathy. 2016. ConvNetJS MNIST Demo. https://cs.stanford.edu/people/karpathy/convnetjs/demo/mnist.html.
[43] Andrej Karpathy. 2023. nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs. https://github.com/karpathy/nanoGPT.
[44] Andrej Karpathy. 2024. Let's build GPT: from scratch, in code, spelled out. https://youtu.be/kCc8FmEb1nY.
[45] Colleen Kehoe, John Stasko, and Ashley Taylor. 2001. Rethinking the evaluation of algorithm animations as learning aids: an observational study. *International Journal of Human-Computer Studies* 54, 2 (2001), 265–284. doi:10.1006/ijhc.2000.0409
[46] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.
[47] Seongmin Lee, Benjamin Hoover, Hendrik Strobelt, Zijie J. Wang, ShengYun Peng, Austin Wright, Kevin Li, Haekyu Park, Haoyang Yang, and Duen Horng Polo Chau. 2024. Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion. In *2024 IEEE Visualization and Visual Analytics (VIS)*.
[48] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. doi:10.1145/2470654.2481287
[49] Yiran Li, Junpeng Wang, Xin Dai, Liang Wang, Chin-Chia Michael Yeh, Yan Zheng, Wei Zhang, and Kwan-Liu Ma. 2023. How does attention work in vision transformers? A visual analytics attempt. *IEEE Transactions on Visualization and Computer Graphics* 29, 6 (2023), 2888–2900.
[50] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread* (2025). https://transformer-circuits.pub/2025/attribution-graphs/biology.html
[51] Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. 2021. Influence patterns for explaining information flow in BERT. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 341, 14 pages.
[52] Yilin Lu, Chongwei Chen, Yuxin Chen, Kexin Huang, Marinka Zitnik, and Qianwen Wang. 2024. GNN 101: Visual Learning of Graph Neural Networks in Your Web Browser. *arXiv preprint arXiv:2411.17849* (2024).
[53] Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. 2023. Visualizing and understanding patch interactions in vision transformer. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
[54] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* 1 (2020), 3.
[55] Aditi Mishra, Bretho Danzy, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2025. PromptAid: Visual prompt exploration, perturbation, testing and iteration for large language models. *IEEE Transactions on Visualization and Computer Graphics* (2025).
[56] MIT RAISE Initiative and Personal Robots Group, MIT Media Lab. 2025. RAISE Playground. https://playground.raise.mit.edu/
[57] Evelyn Navarrete, Andreas Nehring, Sascha Schanze, Ralph Ewerth, and Anett Hoppe. 2025. A closer look into recent video-based learning research: A comprehensive review of video characteristics, tools, technologies, and learning effectiveness. *International Journal of Artificial Intelligence in Education* (2025), 1–64.
[58] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62. doi:10.1016/j.neucom.2021.03.091
[59] nostalgebraist. 2020. Interpreting GPT: The Logit Lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.
[60] Chris Olah. 2014. Neural Networks, Manifolds, and Topology. https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/.
[61] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897* (2023).
[62] R2D3. [n. d.]. A Visual Introduction to Machine Learning. http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.
[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
[64] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
[65] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. 2005. Interactive sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 233–240.
[66] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics* 8 (2020), 842–866. doi:10.1162/tacl_a_00349
[67] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.
[68] Charlie F Ruan, Yucheng Qin, Xun Zhou, Ruihang Lai, Hongyi Jin, Yixin Dong, Bohan Hou, Meng-Shiun Yu, Yiyan Zhai, Sudeep Agarwal, et al. 2024. WebLLM: A High-Performance In-Browser LLM Inference Engine. *arXiv preprint arXiv:2412.15803* (2024).
[69] Zekai Shao, Shuran Sun, Yuheng Zhao, Siyuan Wang, Zhongyu Wei, Tao Gui, Cagatay Turkay, and Siming Chen. 2023. Visual explanation for open-domain question answering with bert. *IEEE Transactions on Visualization and Computer Graphics* 30, 7 (2023), 3779–3797.
[70] Wang Shaohui and Ma Lihua. 2008. The application of blog in modern education. In *2008 International Conference on Computer Science and Software Engineering*, Vol. 4. IEEE, 1083–1085.
[71] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4 (1965), 591–611.
[72] Alexander Skulmowski and M. Xu. 2022. Understanding Cognitive Load in Digital and Online Learning: a New Perspective on Extraneous Cognitive Load. *Educational Psychology Review* 34, 1 (March 2022), 171–196. doi:10.1007/s10648-021-09624-7
[73] Daniel Smilkov, Shan Carter, D. Sculley, Fernanda B. Viégas, and Martin Wattenberg. 2017. Direct-Manipulation Visualization of Deep Networks. *CoRR* abs/1708.03788 (2017). arXiv:1708.03788 http://arxiv.org/abs/1708.03788
[74] Hyuksoon S. Song, Martin Pusic, Michael W. Nick, Umut Sarpel, Jan L. Plass, and Adina L. Kalet. 2014. The cognitive impact of interactive design features for learning complex materials in medical education. *Comput. Educ.* 71 (Feb. 2014), 198–205. doi:10.1016/j.compedu.2013.09.017
[75] Christina Stoiber, Markus Wagner, Florian Grassinger, Margit Pohl, Holger Stitz, Marc Streit, Benjamin Potzmann, and Wolfgang Aigner. 2023. Visualization onboarding grounded in educational theories. In *Visualization psychology*. Springer, 139–164.
[76] Petra Ten Hove and Hans van der Meij. 2015. Like it or not. What characterizes YouTube's more popular instructional videos? *Technical communication* 62, 1 (2015), 48–62.
[77] Jenifer Tidwell. 2010. *Designing interfaces: Patterns for effective interaction design*. " O'Reilly Media, Inc.".
[78] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).
[79] Edward R Tufte and Peter R Graves-Morris. 1983. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT.
[80] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International journal of human-computer studies* 57, 4 (2002), 247–262.
[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need *(NIPS'17)*. Curran Associates Inc., 6000–6010.
[82] Bret Victor. 2011. Explorable Explanations. https://worrydream.com/ExplorableExplanations/.
[83] Jesse Vig. 2019. BertViz: A tool for visualizing multihead self-attention in the BERT model. In *ICLR workshop: Debugging machine learning models*, Vol. 3.
[84] Zijie J Wang, Robert Turko, and Duen Horng Chau. 2021. Dodrio: Exploring transformer models with interactive visualization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*.
[85] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. 2021. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* (2021). doi:10.1109/TVCG.2020.3030418

[86] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2007. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1129–1136.

[87] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[88] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. 2024. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092* (2024).

[89] Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. 2024. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics* (2024). doi:10.1109/TVCG.2023.3327163

[90] Yuzhe You, Jarvis Tse, and Jian Zhao. 2025. Panda or not Panda? Understanding Adversarial Attacks with Interactive Visualization. *ACM Transactions on Interactive Intelligent Systems* (2025). arXiv:2311.13656 [cs.HC] doi:10.1145/3725739

[91] Zeping Yu and Sophia Ananiadou. 2023. Neuron-level knowledge attribution in large language models. *arXiv preprint arXiv:2312.12141* (2023).

[92] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*. 558–567.