Mechanistic Interpretability of Analogical Reasoning in Gemma-2-2B

A Sparse Autoencoder Attribution Graph Analysis

Olalekan Alagbe · Joseph Lawrence · Anish Maheshwar · Konstantinos Krampis

March 2026

Abstract

We present a mechanistic analysis of analogical reasoning in Gemma-2-2B using Neuronpedia attribution graphs and Sparse Autoencoder (SAE) features. By generating and comparing five attribution graphs across structurally distinct analogical prompts — covering geographic analogies (*Paris - France → Berlin - ?*, *Rome - ?*, *Tokyo - ?*) and semantic role analogies (*Doctor - hospital → teacher - ?*, *Fish - water → bird - ?*) — we identify a shared **analogical reasoning circuit** comprising 180 features active across all five prompts and 510 features active across at least three. Each feature is identified by a stable *(layer, feature index)* pair, identifying circuits as lists of recurring internal model feature activation patterns that retain similar structure across analogical prompts.

We discover dedicated analogy-encoding features at layers 5, 8, 9, and 13, including a feature at layer 5 labeled literally as **"analogies"** and a layer 8 feature encoding **"analogies or comparisons"** appearing across all graphs with high influence. Early layers (0–4) contain circuit templates tracking the "X is to Y as Z is to" pattern, while mid-to-late layers (5–13) house increasingly semantic representations of the relational structure. The circuit spans all 26 transformer layers and exhibits cross-domain generalization, with the same core features activating for both geographic and semantic role analogies. Causal validation via 159 feature steering experiments confirms that the identified backbone features are collectively necessary for the model's predictions, and that Phase 2 features collectively implement the relational transfer operation that is the computational core of analogical reasoning.

1. Introduction

Analogical reasoning — the ability to recognize and complete structural relationships between concepts — is a foundational cognitive ability underlying scientific discovery, language understanding, and abstract problem solving. The classic analogy task, *"Paris is to France as Berlin is to ____,"* tests whether a model can identify the capital-city relationship abstractly and apply it to a new country. Large language models (LLMs) exhibit striking competence on such tasks [1], yet the internal computational mechanisms remain poorly understood.

Mechanistic interpretability research has made significant progress in understanding factual recall circuits [2], indirect object identification [3], and syntactic processing [4]. Sparse autoencoders (SAEs) have emerged as a central tool in this effort, learning sparse, interpretable decompositions of model activations [5, 6] that can be applied at scale across all layers and sublayers of large models [7]. The Neuronpedia platform [8] operationalizes this infrastructure, providing public APIs for attribution graph generation and feature steering that democratize circuit-level analysis beyond institutions with direct model access.

However, analogical reasoning presents a distinct challenge beyond prior circuit analyses: it requires not merely retrieving a stored fact, but recognizing a **relational structure** and applying it compositionally to novel inputs. The relation type is never named in the prompt — the model must infer *capital-of* from the example alone, hold it as a variable, and transfer it to a new argument pair. Prior work has documented that LLMs exhibit apparently emergent analogical reasoning [1] and identified internal attention-head mechanisms supporting abstract reasoning [9], yet a feature-level, causally-validated circuit account has been absent.

We address this gap using attribution graphs generated from the `gemmascope-transcoder-16k` SAE suite [7], which provides cross-layer transcoder features for every layer of Gemma-2-2B. Our analysis identifies a three-phase circuit with explicitly labeled analogy-concept features, provides causal validation through 159 steering experiments, and constitutes — to our knowledge — an SAE-level mechanistic account of analogical reasoning in a large language model.

1.1 Research Questions

1. Does Gemma-2-2B employ a **shared circuit** for analogical reasoning, or does it use different mechanisms for different analogy types?
2. Which SAE features — identified by stable *(layer, feature index)* pairs — are most **consistently activated** across diverse analogical prompts?
3. Are there interpretable, semantically meaningful features that encode the **abstract relational structure** of analogies, and how are they discovered?
4. How is the analogical computation **distributed across transformer layers**, and can phase boundaries be causally validated?

2. Methodology

2.1 Prompt Selection

We selected five prompts spanning two structural analogy types to ensure cross-domain coverage:

ID	Prompt	Expected	Type
`analog_berlin`	*"Paris is to France as Berlin is to"*	Germany	Capital
`analog_rome`	*"Paris is to France as Rome is to"*	Italy	Capital

ID	Prompt	Expected	Type
`analog_tokyo`	*"Paris is to France as Tokyo is to"*	Japan	Capital
`analog_teacher`	*"Doctor is to hospital as teacher is to"*	school	Semantic role
`analog_bird`	*"Fish is to water as bird is to"*	air / sky	Semantic role

2.2 Attribution Graph Generation

Attribution graphs were generated using the Neuronpedia API [8] (`/api/graph/generate`) with Gemma-2-2B and the `gemmascope-transcoder-16k` SAE [7] — a 26-layer cross-layer transcoder with 16,384 features per layer. Each graph request returns a JSON object containing nodes (SAE feature activations with layer, index, influence score, and activation magnitude) and directed edges (attribution scores). Graphs were downloaded and loaded into NetworkX `DiGraph` objects for analysis.

Parameter	Value
Model	`gemma-2-2b`
SAE	`gemmascope-transcoder-16k`
Max feature nodes	3,000
Desired logit probability	0.95
Node threshold	0.80
Edge threshold	0.85

Key technical finding: The correct API endpoint for `gemmascope-transcoder-16k` requires a **layer-prefixed SAE identifier** (e.g., `4-gemmascope-transcoder-16k` for layer 4) rather than the global SAE name.

2.3 Feature Identification and Cross-Graph Analysis

2.3.1 Feature Identity via Stable (Layer, Feature Index) Pairs Each feature in the attribution graphs is identified by a stable *(layer, feature index)* pair — for example, *(5, 5793)* uniquely and persistently identifies a feature within the `gemmascope-transcoder-16k` SAE [7]. These identifiers are fixed properties of the trained SAE and do not vary across prompts, sessions, or API calls.

Cross-graph feature overlap was computed by finding which *(layer, feature index)* pairs appear as nodes across multiple independently generated graphs. Formally, let G_i denote the set of feature IDs active in graph i. The shared circuit at threshold k is:

$$\mathcal{C}_k = \left\{ f \;\middle|\; \sum_{i=1}^{5} \mathbf{1}[f \in G_i] \geq k \right\}$$

Three thresholds were analyzed: $k \in \{3, 4, 5\}$. The 180-feature core circuit ($k = 5$) is therefore a concrete, enumerable list of *(layer, feature index)* identifiers that recur across all five independently generated graphs regardless of whether the prompt is geographic or semantic in nature. Feature labels were retrieved using the Neuronpedia feature explanation API [8].

2.4 Three-Phase Architecture: How Phase Boundaries Were Identified

The three-phase architecture was identified through two converging lines of evidence, neither of which required the authors to impose phase boundaries a priori.

Semantic label analysis. After retrieving Neuronpedia automated labels for the top recurring features, a consistent gradient emerged across layer depth. Features in layers 0–4 carry purely syntactic labels: "the word 'to' ", " 'to' followed by a verb", "the phrase 'it is to' ". Features in layers 5–9 carry explicitly relational-semantic labels: "analogies", "analogies or comparisons", "comparison between two things". Features in layers 10–13 carry integrative labels: "comparisons between disciplines and relationships between concepts". The phase boundaries emerge from the content of the labels rather than an arbitrary partition of layers. This three-stage organization mirrors the emergent symbolic architecture documented by Webb et al. [9] for abstract reasoning more broadly, where early layers abstract tokens into relational variables, intermediate layers perform induction over those variables, and later layers retrieve the answer.

Activation magnitude progression. Average activation magnitudes of core features increase monotonically through the phases:

Phase	Layer Range	Role	Activation Magnitude
Circuit template	L0	Token and syntax parsing	1.5 – 6.4
Analogy recognition hub	L5	Analogy concept activation	7.4 – 11.1
Comparison detectors	L8–L9	Relational detection	~13.4
Relational integration	L10–L13	Domain + relation integration	9.1 – 16.3

Causal validation. Phase boundaries were then tested causally via collective suppression experiments (§3.4), which confirmed that each phase is collectively necessary and that earlier phases are prerequisites for later ones.

2.5 Discovery of Analogy-Concept Features

The key features — L5 SAE#5793 ("analogies") and L8 SAE#13766 ("analogies or comparisons") — were not specifically sought. They emerged from the cross-graph overlap analysis described in §2.3. Once the intersection feature set was computed,

each feature's automated Neuronpedia explanation [8] was retrieved. L5 SAE#5793 returned the label "analogies"; L8 SAE#13766 returned "analogies or comparisons."

The significance of these labels is their **domain-agnosticism**. Both features appear in attribution graphs for Berlin, Rome, and Tokyo (geographic capital analogies) and for teacher and bird (semantic role analogies). This is consistent with the broader finding in the analogical reasoning literature that LLMs encode relational information in a domain-general manner [10, 11], and extends that behavioral finding to a specific, causally-validated internal feature. L8 SAE#13766 was additionally notable for having 21 appearances across the five graphs and an influence score of 0.533, placing it among the highest-influence recurring features.

2.6 Phase 2 Definition

Phase 2 is defined by two jointly applied criteria: **layer position (5–9)** and **feature label content**. Features in this layer range whose Neuronpedia labels explicitly reference analogies, comparisons, or relational structure constitute Phase 2. The four members are:

Feature	Label
L5 SAE#5793	"analogies"
L5 SAE#2141	"comparisons of people or figures using well-known public figures"
L8 SAE#13766	"analogies or comparisons"
L9 SAE#13344	"phrases suggesting uncertainty or comparison between two things"

This grouping is validated causally: suppressing all four Phase 2 features simultaneously collapses all five circuits, with capital analogies producing "France" — the source-pair answer — rather than the target country. An arbitrary phase definition would not produce such a consistent and semantically meaningful failure mode.

2.7 Circuit Definition and Causal Validation via Feature Steering

A **circuit** for a given prompt is defined as the backbone of causally important features in that prompt's attribution graph. The extraction procedure is identical for all prompts: trace the top-5 causal paths backward from the logit node (greedy walk following highest-weight incoming edges) and the top-5 paths forward from the highest-influence embedding nodes. Any transcoder feature appearing on at least one of these 10 paths is a backbone member. This procedure is analogous to the automated circuit discovery approach of Conmy et al. [4], applied here at the SAE feature level rather than the attention head level.

Causal validation was performed using the Neuronpedia `/api/steer` endpoint [8] with `modelId: "gemma-2-2b"` and `strength_multiplier: 4`. Four experimental paradigms were applied:

1. **Necessity (individual suppression):** Suppress a single backbone feature at strength −20.
2. **Necessity (full backbone suppression):** Suppress all backbone features simultaneously.
3. **Sufficiency (hub boost):** Boost a backbone hub feature at strength +20 on altered prompts.
4. **Specificity (non-backbone suppression):** Suppress high-activation features not on any causal path.

Two additional circuits from an expanded 30-prompt analysis (Cairo→Kenya, Puppy→cat) were included for cross-validation, for a total of **159 individual steering API calls across 7 circuits**.

3. Results

3.1 Graph Structure

All five attribution graphs exhibited a consistent structural pattern, with features activated across all 26 transformer layers (0–25) plus the embedding layer (E):

Graph	Prompt	Nodes	Edges	Max Influence
`analog_berlin`	Paris - France → Berlin - ?	930	25,915	0.8001
`analog_rome`	Paris - France → Rome - ?	963	27,608	0.8002
`analog_tokyo`	Paris - France → Tokyo - ?	905	22,414	0.8001
`analog_teacher`	Doctor - hospital → teacher - ?	1,040	35,481	0.8001
`analog_bird`	Fish - water → bird - ?	1,071	38,741	0.8000

The semantic role analogies (*teacher*, *bird*) have notably larger graphs (1,040–1,071 nodes, 35k–38k edges) compared to the capital analogies (905–963 nodes, 22k–27k edges). We interpret this as reflecting greater ambiguity in the expected completion domain: the *capital-of* relation maps to a discrete, well-encoded fact [2], whereas professional and ecological roles require broader world-knowledge access.

3.2 The Core Analogical Reasoning Circuit

Cross-graph feature overlap analysis over the stable *(layer, feature index)* identifier space revealed a substantial shared circuit:

Threshold	Features Found
Active in ≥3/5 graphs	**510 features**

Threshold	Features Found
Active in ≥4/5 graphs	**277 features**
Active in all 5 graphs	**180 features**

Core features by layer group (5/5 graphs):

Layer Group	Core Features
L0	12
L1–L4	19
L5–L6	12
L8–L13	7

early layers (L0–L3) account for the plurality of core features, consistent with circuit template processing occurring first. The mid-range layers (L5–L6) show elevated feature counts relative to neighbors — these are the **analogy recognition hub** layers. Isolated high-influence features appear at L8, L9, L11, and L13.

3.3 The Three-Phase Analogical Reasoning Circuit

We provide evidence that Gemma-2-2B performs **genuine multi-step analogical reasoning internally**. The attribution graph reveals a three-phase computational process that activates for both geographic and semantic role analogies — evidence of a domain-agnostic relational reasoning mechanism. This three-stage organization parallels the symbolic architecture identified by Webb et al. [9] through causal mediation analysis and the internal representation findings of Lee et al. [10].

Phase 1 · layers 0–4 · Circuit Template Parsing

Feature	Label
L0 SAE#11651	*"the word 'to' "*
L1 SAE#11356	*"the word 'to' followed by a verb"*
L2 SAE#11475	*"the word 'refers' and related words"*
L4 SAE#10752	*"uses of the verb 'to be' preceded by 'to' "*
L5 SAE#9672	*"the phrase 'it is to' "*

These features encode the syntactic skeleton of the analogy prompt. Their progression from individual tokens to multi-word patterns reflects hierarchical parsing of the relational connective. These are *structural* features — they fire on any text with this grammatical form, not specifically on analogical content.

Phase 2 · layers 5–9 · Analogy Recognition Hub

Feature	Label
L5 SAE#5793	*"analogies"* ← dedicated analogy concept feature
L5 SAE#2141	*"comparisons of people or figures using well-known public figures"*
L8 SAE#13766	*"analogies or comparisons"* (21 activations across 5 graphs, influence 0.533)
L9 SAE#13344	*"phrases suggesting uncertainty or comparison between two things"*

This is where circuit template processing gives way to semantic recognition of the *relational concept itself*. The presence of L5 SAE#5793, labeled "analogies" by Neuronpedia's automated SAE feature explanation system [8], is particularly significant: it activates consistently for both capital-city and semantic role analogies. It is not a geographic feature — it fires equally for "Doctor - hospital → teacher - ?". This is direct evidence of the kind of abstract relational representation that prior behavioral work [1, 11] has hypothesized but not directly observed inside a model.

Phase 3 · layers 10–13 · Relational Integration

Feature	Label
L11 SAE#15947	*"references to historical or social change"*
L13 SAE#10969	*"comparisons between disciplines and relationships between concepts"*

L13 SAE#10969 serves an integrative role, combining the recognized relational structure from Phase 2 with domain-specific knowledge to produce the final completion. layers 14–25 then handle domain-specific knowledge retrieval and output token formatting, analogous to the factual recall circuits identified by Meng et al. [2].

> **Note:** This diagram simplifies the true mechanisms considerably. The attribution graph for any single prompt contains hundreds of features; the circuit shown represents the semantically interpretable core.

3.4 Top Recurring Features

Directly analogical features (Neuronpedia labels explicitly reference analogical reasoning or comparison):

Feature	Appearances	Avg Influence	Label
L5 #5793	11/5	0.590	"analogies"
L8 #13766	21/5	0.533	"analogies or comparisons"
L9 #13344	14/5	0.681	"comparison between two things"

Feature	Appearances	Avg Influence	Label
L5 #2141	12/5	0.647	“comparisons of public figures”
L13 #10969	11/5	0.676	“comparisons between disciplines”

Circuit templates (encode the “X is to Y as Z is to” scaffold):

Feature	Appearances	Avg Influence	Label
L0 #11651	10/5	0.633	“the word ‘to’ ”
L1 #11356	10/5	0.609	“ ‘to’ followed by a verb”
L2 #11475	10/5	0.638	“the word ‘refers’ ”
L4 #10752	10/5	0.626	“ ‘to be’ preceded by ‘to’ ”
L5 #9672	12/5	0.579	“the phrase ‘it is to’ ”

High-recurrence formal text features (labels unrelated to analogical reasoning):

Feature	Appearances	Avg Influence	Label
L4 #14857	22/5	0.681	“code snippets and license agreements”
L6 #2267	20/5	0.724	“words in programming code, legal jargon, or scientific texts”
L3 #3205	20/5	0.670	“code snippets and documentation references”

These formal-text features have higher raw appearance counts than the explicitly analogical features. Causal steering (§3.7.6) confirms they are inert for all high-confidence circuits, consistent with their role as detectors of syntactic formality rather than relational semantics. The polysemanticity of neurons in large models [6] is precisely why SAE-based feature decomposition [5, 6, 7] is necessary to distinguish these classes of activation.

3.5 Cross-Domain Generalization

The consistent activation of L5 SAE#5793 (“analogies”) and L8 SAE#13766 (“analogies or comparisons”) across both capital-city and semantic role analogy types provides the most direct evidence for a **domain-general analogical reasoning mechanism**. The 180 features active in all five graphs form the stable intersection of the two analogy type families, and this intersection includes the core analogy-concept features at L5 and L8.

The slightly larger graphs for semantic role analogies (teacher, bird: 1,040–1,071 nodes) relative to capital analogies (Berlin, Rome, Tokyo: 905–963 nodes) may reflect that semantic role completions require broader world-knowledge access — knowing that teachers work in schools, or that birds inhabit air — rather than purely relational computation over a discrete, well-encoded geographic fact [2].

3.6 Activation Magnitudes Build Through Layers

Average activation magnitudes of core circuit features increase substantially with layer depth:

Layer Range	Role	Typical Activation
L0 (structural)	Token and syntax parsing	1.5 – 6.4
L5 (analogy hub)	Analogy concept activation	7.4 – 11.1
L8–L9 (detectors)	Comparison detection	~13.4
L10–L13 (integration)	Relational + domain integration	9.1 – 16.3

This monotonically increasing pattern is consistent with an accumulating signal as the relational structure is assembled.

3.7 Causal Validation via Feature Steering

The attribution graph analysis identifies recurring features and causal path structures but does not by itself establish whether these features are causally necessary for the model's predictions. To distinguish load-bearing circuit components from high-activation but functionally inert nodes, we performed systematic causal steering experiments via the Neuronpedia API [8].

3.7.1 Late-Layer Backbone Necessity (Individual Suppression) `analog_berlin` — "Paris is to France as Berlin is to" → Germany (p=0.973)

Feature	Layer	Index	Role	Steered Token	Necessary?
Science hub	21	4827	Strongest path entry (edge +198.0)	Germany	no
Relay	22	15670	Path 1 relay	Germany	no
Output driver A	25	4717	Final amplifier (shared across circuits)	Germany	no

Feature	Layer	Index	Role	Steered Token	Necessary?
Location encoder	16	6491	Location/direction feature, path 2 entry	Germany	no
Relay	17	14546	Mid-cascade relay	Germany	no
Relay	19	5773	Late relay	Germany	no
Integrator	21	7482	Integration hub (paths 2–4)	Germany	no
Output driver B	25	2725	Secondary output driver (edge −2.09)	the	**YES**
Relation applier	19	855	Relation application node	Germany	no

1/9 necessary. Only L25/2725 is a single point of failure. The highest-weight feature (L21/4827, edge +198.0) is not individually necessary, demonstrating that attribution weight alone does not predict causal necessity — a key methodological lesson consistent with prior circuit analysis work [3, 4].

`analog_rome` — "Paris is to France as Rome is to" → Italy (p=0.974)

Feature	Layer	Index	Role	Steered Token	Necessary?
Relay	20	15360	Backward path from logit	Italy	no
Late gate	24	16122	Backward path, L24 suppression gate	the	**YES**
Output driver	25	286	Backward path, output driver	the	**YES**
Final amplifier	25	4717	Shared final amplifier (act=265.2)	the	**YES**
Output driver C	25	10521	Tertiary output driver	the	**YES**

Feature	Layer	Index	Role	Steered Token	Necessary?
Relay	17	14546	Mid-cascade relay (shared with Berlin)	Italy	no
Relay A	22	12202	Late relay	Italy	no
Relay B	22	14727	Late relay	Italy	no
Relay	23	5917	Late relay	Italy	no
Secondary gate	24	13277	Late gate	Italy	no

4/10 necessary. The Rome circuit has more single points of failure than Berlin despite near-identical confidence (p=0.974 vs 0.973), indicating path redundancy varies even among structurally similar geographic analogies.

`analog_tokyo` — "Paris is to France as Tokyo is to" → Japan (p=0.990)

Feature	Layer	Index	Role	Steered Token	Necessary?
Relay	20	15360	Backward path from logit	Japan	no
Output driver	25	286	Backward path, output driver	the	**YES**
Output driver B	25	12223	Backward path, secondary output	Japan	no
Relay	17	14546	Mid-cascade relay (shared)	Japan	no
Late relay A	23	850	Late relay	the	**YES**
Late relay B	23	13914	Late relay (also necessary in Cairo circuit)	the	**YES**
Gate	24	13277	Late gate (shared with Rome)	Japan	no
Output driver C	25	10152	Tertiary output	Japan	no

Feature	Layer	Index	Role	Steered Token	Necessary?
Hub	20	6648	L20 convergence hub	Japan	no
Integration	21	7764	Late integration	Japan	no

3/10 necessary. L23/13914 is necessary in both Tokyo and Cairo circuits — a shared bottleneck consistent with a late-layer "geographic entity selector" role. L25/286 recurs as necessary in Rome, Tokyo, and Cairo, making it the single most critical output driver across geographic analogies.

`analog_teacher` — "Doctor is to hospital as teacher is to" → school (p=0.486)

Feature	Layer	Index	Role	Steered Token	Necessary?
Embedding	0	17	Backward path, embedding-level	the	**YES**
Gateway	18	6532	Backward path, mid-late gateway	school	no
Hub	20	6179	Backward path, convergence hub	school	no
Output driver	25	4975	Backward path, output driver	…	**YES**
Final amplifier	25	4717	Shared final amplifier (act=135.6)	a	**YES**
Relay	22	15670	Late relay (shared)	school	no
Relay B	18	11952	Mid-late relay	school	no
Legal docs	18	13586	Legal docs feature	school	no
Convergence	21	2655	Late convergence hub	school	no
Gate	24	15259	Late suppression gate	school	no

3/10 necessary. The teacher circuit is the only one where an **L0 embedding-level feature** (L0/17) is individually necessary — suggesting the semantic role analogy relies on an early feature not redundantly encoded by later layers, unlike the capital analogies.

`analog_bird` — "Fish is to water as bird is to" → air (p=0.117)

Feature	Layer	Index	Role	Steered Token	Necessary?
Backward A	22	4252	Backward path from logit	(space)	**YES**
Backward B	24	8106	Backward path, late gate	____	**YES**
Final amplifier	25	4717	Shared final amplifier (act=122.3)	the	**YES**
Output driver	25	11801	Output driver	?	**YES**
Relay A	22	15670	Late relay (shared)	_____	**YES**
Relay B	22	14727	Late relay	(space)	**YES**
Relay C	22	13619	Late relay	(space)	**YES**
Gate A	24	4383	Suppression gate	air	no
Gate B	24	12559	Suppression gate	the	**YES**
Hub	20	3094	Integration hub	air	no

8/10 necessary. This is the most fragile circuit in the dataset. Three L22 relay features (15670, 14727, 13619) are all independently necessary despite occupying the same layer, indicating they carry non-redundant information through parallel channels. This fragility is consistent with the circuit's very low prediction confidence (p=0.117).

Cross-circuit pattern. Necessity inversely correlates with prediction confidence: Berlin (p=0.973): 1/9; Rome (p=0.974): 4/10; Tokyo (p=0.990): 3/10; Teacher (p=0.486): 3/10; Bird (p=0.117): 8/10. Three features recur as necessary across multiple circuits: **L25/#286** (Rome, Tokyo, Cairo), **L25/#4717** (Rome, Teacher, Bird), and **L23/#13914** (Tokyo, Cairo).

3.7.2 Full Backbone Suppression

Circuit	p	N feat.	Default Output	Steered Output	Disrupted?
analog_berlin	0.973	9	Germany. It is the	of of of of of	YES
analog_rome	0.974	10	Italy. It is the	pleafure pleafure plea	YES
analog_tokyo	0.990	10	Japan. It is the	country country count	YES
analog_teacher	0.486	10	school. The doc	1111	YES
analog_bird	0.117	10	air. The fish	(newline) The the the	YES
Cairo→Kenya	0.963	9	Kenya. It is the	(whitespace)	YES
Puppy→cat	0.756	4	cat. I’	cat. I think	no

6/7 circuits fully disrupted. Failure modes are qualitatively informative: capital analogies degenerate to repetitive or archaic text, indicating the backbone is required for entity selection while the prompt structure alone partially activates a “country” category. Teacher collapses to “1111”; bird falls through to generic continuation. Puppy→cat is the sole exception, apparently carried by direct embedding-to-logit connections outside the multi-hop backbone.

3.7.3 Phase 1 and Phase 2 Feature Necessity Individual suppression of the 9 key phase features across all five prompts (45 tests):

Feature	Phase	Label	Berlin	Rome	Tokyo	Teacher	Bird
L0/11651	1	“the word ‘to’ ”	Berlin	Rome	Tokyo	school	water
L1/11356	1	“ ‘to’ followed by a verb”	—	—	—	—	—
L4/10752	1	“ ‘to be’ pre-ceded by ‘to’ ”	—	—	—	classroom	sky
L5/9672	1	“the phrase ‘it is to’ ”	—	—	—	—	sky
L5/5793	**2**	**“analogies”**	—	—	—	—	—

Feature	Phase	Label	Berlin	Rome	Tokyo	Teacher	Bird
L5/2141	2	“comparisons of public figures”	—	—	—	—	—
L8/13766	2	“analogies or comparisons”	—	—	—	—	fish
L9/13344	2	“comparison between two things”	—	—	—	—	sky
L13/10969	3	“comparisons between disciplines”	—	—	—	—	—

Cells show the steered first token when suppressed at strength −20. “—” = prediction unchanged.

Phase 1 results. L0/11651 is necessary in 4/5 circuits. Suppressing it causes capital analogies to predict the *city name itself* (Berlin, Rome, Tokyo), indicating the model reverts to the most recently mentioned entity rather than completing the analogy. For the bird circuit, suppression produces “water” (source-pair element).

Phase 2 results. L5/5793 (“analogies”) is **never individually necessary** in any circuit — it is individually redundant for high-confidence circuits. For the fragile bird circuit (p=0.117), however, Phase 2 features become individually necessary: L8/13766 changes “air” to “fish” (source-domain animal); L9/13344 changes “air” to “sky.”

Phase 3 results. L13/10969 is not individually necessary for any circuit.

3.7.4 Collective Phase Suppression

Experiment	Features Suppressed	Berlin	Rome	Tokyo	Teacher	Bird
All Phase 2 (4 feat.)	L5/5793, L5/2141, L8/13766, L9/13344	**France**	**France**	**France**	be	fish
All Phase 1 (5 feat.)	L0/11651, L1/11356, L4/10752, L5/9672, L2/11475	(empty)	(empty)	(empty)	to	to

Experiment	Features Suppressed	Berlin	Rome	Tokyo	Teacher	Bird
Phase 1+2 (9 feat.)	All Phase 1 + Phase 2	:	:	:	:	:
Phase 1+2+3 (10 feat.)	All Phase 1 + Phase 2 + L13/10969	:	:	:	be	:

All 20/20 cells disrupted.

Phase 2 collective suppression is the most informative experiment in this paper. All three capital analogies output **"France"** — retaining the factual association "Paris is to France" but losing the relational transfer "as Berlin is to ___." This is direct causal evidence that Phase 2 features collectively implement the relational transfer operation. The failure mode is precisely what one would predict from the internal representation findings of Lee et al. [10], where reasoning failures reflect missing relational information in mid-upper layers.

Phase 1 collective suppression produces empty outputs for capital analogies and "to" for semantic role analogies — a more severe failure, consistent with Phase 1 being a prerequisite for Phase 2.

Combined Phase 1+2 suppression produces ":" for 4/5 circuits, consistent with the model defaulting to list-formatting punctuation when both template parsing and analogy recognition are disabled.

These results establish a **causal hierarchy**: Phase 1 → Phase 2 → Phase 3 + late layers. Each phase is collectively necessary, and earlier phases are prerequisites for later ones.

3.7.5 Sufficiency (Hub Boost on Altered Prompts)

Circuit	Hub Boosted	Altered Prompt	Induced?
`analog_berlin`	L21/4827	"Cairo is to Egypt as Nairobi is to"	no
`analog_berlin`	L21/4827	"Madrid is to Spain as Berlin is to"	**YES → Germany**
`analog_rome`	L20/15360	"Paris is to France as Tokyo is to"	no
`analog_rome`	L20/15360	"Madrid is to Spain as Rome is to"	**YES → Italy**
`analog_tokyo`	L20/15360	"Paris is to France as Rome is to"	no

Circuit	Hub Boosted	Altered Prompt	Induced?
`analog_tokyo`	L20/15360	"Beijing is to China as Tokyo is to"	**YES → Japan**
`analog_teacher`	L0/17	"Nurse is to hospital as teacher is to"	no
`analog_teacher`	L0/17	"Doctor is to hospital as chef is to"	no
`analog_bird`	L22/4252	"Cat is to land as bird is to"	no
`analog_bird`	L22/4252	"Fish is to water as eagle is to"	**YES → air**
Cairo→Kenya	L15/15954	"Lagos is to Nigeria as Nairobi is to"	**YES → Kenya**

5/11 tests succeed. Sufficiency holds when the altered prompt retains the target entity or a semantically close substitute, and fails when it crosses domain boundaries. The capital hubs encode domain-specific geographic associations rather than general-purpose "answer slot" activators.

3.7.6 Specificity (Non-Backbone Feature Suppression)

Circuit	Feature	Label	Steered Token	Disrupted?
`analog_berlin`	L6/3335	"difficulty/challenges"	Germany	no
`analog_berlin`	L13/4435	"opera-related terms"	Germany	no
`analog_rome`	L6/2267	"formal text/code"	Italy	no
`analog_rome`	L4/14857	"code snippets"	Italy	no
`analog_tokyo`	L6/2267	"formal text/code"	Japan	no
`analog_tokyo`	L3/10018	early structural feature	Japan	no
`analog_teacher`	L4/14857	"code snippets"	school	no
`analog_teacher`	L8/13766	"analogies or comparisons"	school	no
`analog_bird`	L6/2267	"formal text/code"	**sky**	YES
`analog_bird`	L5/5793	"analogies"	air	no

Circuit	Feature	Label	Steered Token	Disrupted?
Cairo→Kenya	L5/5500	“profanity and comparisons”	Kenya	no
Puppy→cat	L9/2909	“formulas/ratios”	cat	no

12/13 pass specificity. The sole exception — L6/2267 tipping bird from “air” to “sky” — occurs at the margin of unresolved token competition (p=0.117) and is confirmed inert for all high-confidence circuits. L5/5793 (“analogies”) passes specificity for the bird circuit, consistent with it being individually dispensable but collectively necessary.

3.7.7 Summary of Causal Validation

Circuit	p	Type	Individual Necessity	Full Suppress	Phase 2 Collective	Sufficiency	Specificity
analog_berlin	0.197	Capital	1/9	DISRUPTED	~~Đ~~ France	1/2	PASS
analog_rome	0.974	Capital	4/10	DISRUPTED	~~Đ~~ France	1/2	PASS
analog_tokyo	0.990	Capital	3/10	DISRUPTED	~~Đ~~ France	1/2	PASS
analog_teacher	0.486	Sem. role	3/10	DISRUPTED	~~Đ~~ be	0/2	PASS
analog_bird	0.117	Sem. role	8/10	DISRUPTED	~~Đ~~ fish	1/2	1/2
Cairo→Kenya	0.963	Capital	2/9	DISRUPTED	~~Đ~~	1/2	PASS
Puppy→cat	0.756	Sem. role	0/4	intact	—	—	PASS

Across 159 steering experiments, five principal findings emerge:

1. **The late-layer backbone is collectively necessary.** Full backbone suppression disrupts 6/7 circuits.
2. **Individual necessity scales inversely with prediction confidence.** High-confidence circuits have 1–4 necessary features; the lowest-confidence has 8/10.
3. **Phase 2 is collectively necessary but individually redundant.** Simultaneous suppression collapses every circuit; capital analogies revert to the source-pair answer (“France”).
4. **Phase 1 circuit template features are individually necessary.** L0/11651 alone disrupts 4/5 circuits.
5. **Formal-text features are causally inert.** L4/#14857 and L6/#2267 do not affect any high-confidence prediction.

4. Discussion

4.1 The Analogical Reasoning Circuit in Gemma-2-2B

Our analysis reveals that Gemma-2-2B implements analogical reasoning through a distributed circuit spanning all 26 transformer layers, with specific functional specialization at each phase. The most significant finding is the existence of **explicitly semantic analogy features** at layers 5, 8, 9, and 13 — features whose automated explanations use the words "analogies," "comparisons," and "relationships between concepts." This suggests that the model has internalized analogical structure as a discrete, reusable computational primitive.

This is qualitatively distinct from multi-hop factual reasoning. Analogical reasoning requires extracting an unnamed relation type, holding it as a variable, and applying it to a new argument pair. The Phase 2 collective suppression experiment demonstrates that this extraction and transfer are implemented by identifiable, causally load-bearing internal components whose removal causes the model to echo the source-pair answer rather than transfer the relation — consistent with the "missing relational information" failure mode documented by Lee et al. [10] at the behavioral level. Our work provides a feature-level causal account of this phenomenon.

Prior behavioral evidence [1] established that LLMs can match human performance on analogical tasks; Webb et al. [9] identified emergent symbolic mechanisms supporting abstract reasoning through causal mediation of attention heads. The present work extends these findings to the SAE feature level: the relational reasoning primitive is not just a pattern of attention head behavior but a specifically labeled, causally load-bearing feature in the SAE's learned decomposition of residual stream activations.

4.2 The Role of Formal Text Features

The high-recurrence "code and legal text" features present an interpretive puzzle best understood through the lens of polysemanticity and superposition [6]. Two complementary explanations:

Functional hypothesis: These features detect formal, template-driven text patterns generally. The analogy syntax "X is to Y as Z is to" is highly structured, resembling legal definitions, code comments, and mathematical notation. The model reuses a general "formal syntax" detector.

Training data hypothesis: The analogy format appears frequently in SAT preparation and educational materials — which also contain code examples and legal definitions — creating a statistical association between formal-text features and analogy-completion contexts.

Both are compatible with the causal steering data. The formal features process the syntactic surface of the template while the analogy features process the relational semantics; only the latter are collectively necessary for relational transfer. The SAE-based decomposition [5, 6] is what makes this functional distinction visible — raw neuron activations would not cleanly separate these roles.

4.3 Comparison with the Capital City Recall Circuit

Comparison with the capital city *factual recall* circuit (prompt: "The capital of X is") reveals:

- **Overlap:** Formal-text features (L4/#14857, L6/#2267) appear with high frequency in both circuits, activated by the formal definitional structure of both prompt types. This is analogous to the shared MLP modules Meng et al. [2] identified across different factual recall tasks.
- **Divergence:** The L5 "analogies" feature and L8 "analogies or comparisons" feature appear to be specific to the analogical task — they were not among the top recurring features in the factual recall circuit — supporting the interpretation that these features are selectively activated by relational structure recognition.

4.4 Relation to Anthropic's Attribution Graph Methodology

The present work is in direct methodological continuity with Anthropic's *On the Biology of a Large Language Model* [12], which applied attribution graphs to Claude 3.5 Haiku using cross-layer transcoders. Both papers find that models implement multi-step, staged computation rather than direct input-to-output pattern matching, and both validate circuit hypotheses through feature steering. Anthropic's paper groups related features into manually curated "supernodes" to present a cleaner narrative; the present work uses automated cross-graph intersection, which is more scalable and less susceptible to confirmation bias but produces a less narratively refined picture of any single circuit. The two approaches are complementary.

4.5 Redundancy as a Property of Well-Learned Computation

The inverse relationship between prediction confidence and circuit fragility — ranging from 1/9 individually necessary features (Berlin, p=0.973) to 8/10 (bird, p=0.117) — suggests a general principle: well-learned associations are protected by redundant parallel causal paths, while barely-resolved predictions rely on non-redundant chains. This principle aligns with the circuit redundancy findings in [12] and may reflect a general property of how transformers allocate computational resources across tasks of varying difficulty.

5. Limitations

1. **SAE-feature-level intervention only.** Steering operates at the SAE feature level, not the attention head or residual stream level. The causal role of non-SAE circuit components is not assessed.
2. **SAE coverage.** The `gemmascope-transcoder-16k` SAE [7] covers only cross-layer transcoder features. Attention head contributions and residual stream features are not captured.
3. **Threshold sensitivity.** Results are sensitive to node and edge thresholds (0.80/0.85). Lower thresholds would reveal more features; higher thresholds would produce sparser, more focused circuits.
4. **Label quality.** Neuronpedia [8] automated feature explanations are LLM-generated and may not perfectly capture feature semantics.

5. **Prompt set size.** Five prompts are sufficient for initial circuit identification but too few to claim statistical robustness. A larger prompt set covering arithmetic, cross-lingual, and abstract relational analogies [13] would strengthen conclusions.

Future work: Direct causal validation with TransformerLens activation patching at the attention head and residual stream level; expanded prompt sets; analysis across model scales (Gemma-2-9B, 27B); comparison with factual recall and multi-hop reasoning circuits; and testing whether the Phase 2 features generalize to the cross-lingual analogical settings studied in [14].

6. Conclusions

We have identified and characterized the **analogical reasoning circuit in Gemma-2-2B** using SAE attribution graphs from the Neuronpedia platform [8]. The key conclusions are:

1. **A stable shared circuit exists, identified by common feature IDs.** 180 features — identified by stable *(layer, feature index)* pairs — appear in all five independently generated attribution graphs.
2. **Dedicated analogy features exist at layers 5, 8, 9, and 13.** These features have Neuronpedia explanations explicitly referencing analogies, comparisons, and relational concepts — providing direct SAE-level evidence of interpretable analogy-concept features in a large language model.
3. **The circuit exhibits a three-phase architecture, identified by label semantics and validated causally.** Circuit template parsing (L0–L4), analogy recognition (L5–L9), and relational integration (L10–L13), with activation magnitude increasing through the sequence.
4. **Cross-domain generalization is confirmed.** The same core features, including L5 SAE#5793 ("analogies"), activate for both geographic and semantic role analogies — a domain-agnostic relational reasoning primitive consistent with behavioral findings [1, 10, 11].
5. **Phase 2 implements relational transfer, collectively but not individually.** Simultaneous suppression collapses every circuit; capital analogies revert to the source-pair answer.
6. **Circuit fragility tracks prediction confidence.** High-confidence predictions route through redundant parallel causal paths (1–4 necessary features); low-confidence predictions rely on fragile non-redundant chains (up to 8/10 necessary).

Attribution Graphs

The five Neuronpedia attribution graphs generated for this study are publicly available for interactive exploration. Full graph descriptions, inference prompts, and the agent pipeline methodology are documented in the Supplementary Material.

Prompt	Neuronpedia Graph
Paris is to France as Berlin is to	analog_berlin
Paris is to France as Rome is to	analog_rome
Paris is to France as Tokyo is to	analog_tokyo
Doctor is to hospital as teacher is to	analog_teacher
Fish is to water as bird is to	analog_bird

References

[1] Webb, T., Holyoak, K.J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7, 1526–1541. arXiv: https://arxiv.org/abs/2212.09196

[2] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *NeurIPS 2022*. https://arxiv.org/abs/2202.05262

[3] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *ICLR 2023*. https://arxiv.org/abs/2211.00593

[4] Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *NeurIPS 2023*. https://arxiv.org/abs/2304.14997

[5] Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *ICLR 2024*. https://arxiv.org/abs/2309.08600

[6] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features

[7] Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., & Nanda, N. (2024). Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. https://arxiv.org/abs/2408.05147

[8] Lin, J., & Bloom, J. (2023). Neuronpedia: Interactive platform for sparse autoencoder research and feature steering. https://www.neuronpedia.org

[9] Webb, T.W., Frankland, S.M., Altabaa, A., Segert, S., Krishnamurthy, K., Campbell, D., Russin, J., Giallanza, T., O'Reilly, R., Lafferty, J., & Cohen, J.D. (2025). Emergent symbolic mechanisms support abstract reasoning in large language models. https://arxiv.org/abs/2502.20332

[10] Lee, T., et al. (2025). The curious case of analogies: Investigating analogical reasoning in large language models. https://arxiv.org/abs/2511.20344

[11] Wijesiriwardene, T., et al. (2025). Analogical reasoning inside large language models: Concept vectors and the limits of abstraction. https://arxiv.org/abs/2503.03666

[12] Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N.L., et al. (2025). On the biology of a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[13] Turney, P.D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3), 379–416. [Foundational work on relational similarity benchmarks underlying analogy tasks.]

[14] Allen, C., & Hospedales, T. (2019). Analogies explained: Towards understanding word embeddings. *ICML 2019*. https://arxiv.org/abs/1901.09813

[15] Marks, S., Rager, C., Michaud, E.J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. https://arxiv.org/abs/2403.19647

Supplementary Materials

Interactive Presentation: 20-slide reveal.js presentation with circuit flow diagrams, feature tables, and layer-by-layer analysis.
https://kkrampis.github.io/autocircuit/presentation.html

Live Attribution Graphs:

- `analog_berlin` — Paris - France → Berlin - ?
- `analog_rome` — Paris - France → Rome - ?
- `analog_tokyo` — Paris - France → Tokyo - ?
- `analog_teacher` — Doctor - hospital → teacher - ?
- `analog_bird` — Fish - water → bird - ?

Code: https://github.com/kkrampis/autocircuit