A Tiny Interpretable Codec for Analogy Structure in Qwen Hidden States

Larry Richards

April 27, 2026

Abstract

This report describes a small experiment in extracting a human-readable semantic coordinate system from the hidden states of a frozen language model. We take hidden activations from Qwen 2.5 for four words—*king*, *queen*, *man*, and *woman*—and train a linear encoder that maps those high-dimensional activations into a two-dimensional latent space. In the learned space, the classic analogy relation

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

holds with cosine similarity around 0.9996 under the full objective. A classification-only ablation still recovers an analogy cosine of 0.9840, suggesting that the relation is not created solely by the explicit analogy term. The four word classes are also perfectly separated by a nearest-mean classifier in the two-dimensional space, and the learned gender and royalty directions are nearly orthogonal. The point of the experiment is not to claim a universal semantic basis inside Qwen. It is to show, in a controlled setting, that useful semantic structure can be pulled out of model activations and compressed into a very small, inspectable interface. Held-out vocabulary tests show partial generalization: royal words such as *prince* and *duke* land nearest to *king*, while *princess* and *duchess* land nearest to *queen*. A pseudoinverse hidden-state intervention gives small but directional evidence for write access: increasing the 2D gender shift raises log-probabilities for *queen*, *woman*, and *princess*, while slightly lowering *king*.

1 Introduction

Large language models represent words and contexts as high-dimensional internal states. These states are powerful, but they are not naturally easy to inspect. The broad question behind this experiment is simple: can we build a small interface, or *codec*, that exposes a meaningful slice of this internal structure in a form a human can look at?

This experiment focuses on one deliberately tiny semantic domain: the familiar analogy

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}.$$

[2]. Rather than testing this relation in an embedding table, we extract hidden states from a frozen Qwen 2.5 instruction model [12] while the target words appear in simple sentence contexts. We then train a two-dimensional linear projection that organizes those hidden states into an interpretable plane.

The result is a compact proof of concept. A two-dimensional learned space is enough to preserve the analogy, separate the four classes, and expose two nearly orthogonal semantic directions corresponding roughly to gender and royalty. This is a small result, but it is useful because it gives the larger "latent bus" idea something concrete to stand on: model internals can be wrapped with small, testable semantic interfaces.

2 Experiment Overview

The experiment has three stages.

1. Extract hidden-state representations for the target words from a frozen Qwen model.
2. Train a linear codec from the model's hidden dimension into a two-dimensional latent space.
3. Validate the learned space using analogy, classification, and axis geometry checks.

The target vocabulary is intentionally small:

$$\mathcal{W} = \{\text{king}, \text{queen}, \text{man}, \text{woman}\}.$$

For each word, the extraction script generates 100 short sentence contexts from simple templates such as:

I saw the king today.
The king was there.
Look at the king.
Where is the king?

The target word is located using tokenizer offset mappings, and the hidden state for the final token of the word is extracted from a selected model layer. In the original run, the selected hidden layer used an offset of 10 from the final layer. This gives 400 total hidden-state examples: 100 for each of the four words.

3 Codec Model

Let $x \in \mathbb{R}^d$ be a Qwen hidden-state vector for a target word in context. The codec learns a linear encoder

$$z = xW_e + b_e,$$

where $z \in \mathbb{R}^2$. The purpose of the encoder is not merely to reduce dimension. It is trained to produce a two-dimensional space where the analogy geometry and class separation are both explicit. In this respect it is close in spirit to linear probing, but with an explicit geometric target rather than only a diagnostic classifier [3, 5].

The training objective combines several terms:

$$\mathcal{L} = \mathcal{L}_{cls} + 3.0\mathcal{L}_{analogy} + \mathcal{L}_{perp} + 0.1\mathcal{L}_{orth} + 0.5\mathcal{L}_{sep}.$$

Here, $\mathcal{L}_{cls}$ is a classification loss in the two-dimensional space. The analogy loss encourages

$$\mu_{\text{king}} - \mu_{\text{man}} + \mu_{\text{woman}} \approx \mu_{\text{queen}},$$

where μ_w is the class mean for word w after projection. The perpendicularity term encourages the gender and royalty directions to be close to orthogonal. A light encoder regularizer keeps the projection from becoming pathological, and a separation term discourages the four class means from collapsing together.

This is deliberately simple: no nonlinear encoder, no complicated probing architecture, and no model fine-tuning. The language model is frozen; only the small projection and a lightweight classifier head are trained.

4 Validation

The experiment uses three main validation gates.

4.1 Analogy Gate

The first gate checks whether the analogy holds in the learned two-dimensional space:

$$\cos(\mu_{\text{king}} - \mu_{\text{man}} + \mu_{\text{woman}}, \mu_{\text{queen}}) \geq 0.95.$$

In the saved run, this value is approximately 0.9996. This is the strongest single sign that the codec is organizing the four concepts into the intended geometry.

4.2 Classification Gate

The second gate asks whether individual projected examples can be classified correctly in the two-dimensional space. The validation code uses a simple nearest-mean classifier. Each held-out projected point is assigned to whichever class mean is closest in Euclidean distance:

$$\hat{y}(z) = \arg\min_{w \in \mathcal{W}} \|z - \mu_w\|_2.$$

The data points used here are not synthetic perturbations in the two-dimensional space. They are real hidden-state activations extracted from different sentence contexts, then projected through the learned codec. The original validation uses a 70/30 split per class for the classification report, producing 30 held-out examples per word and 120 held-out examples total.

4.3 Axis Geometry Gate

The third gate checks whether the two main semantic directions are nearly orthogonal. The gender direction is defined as

$$g = \mu_{\text{woman}} - \mu_{\text{man}},$$

and the royalty direction is defined as

$$r = \mu_{\text{king}} - \mu_{\text{man}}.$$

The angle between these two directions is approximately 89.7°, which is very close to the intended right angle.

5 Results

Table 1 summarizes the main validation results.

Gate	Metric	Result	Target
Analogy	Cosine similarity	0.9996	≥ 0.95
Classification	Held-out accuracy	100.0%	$\geq 98\%$
Axis geometry	Gender/royalty angle	89.7°	near 90°

Table 1: Main validation gates for the trained two-dimensional codec.

True class	king	queen	man	woman
king	30	0	0	0
queen	0	30	0	0
man	0	0	30	0
woman	0	0	0	30

Table 2: Nearest-mean classification in the learned two-dimensional space.

The held-out confusion matrix is perfectly diagonal:

The important thing about this result is not that the classifier is complex. It is the opposite: classification is almost embarrassingly simple once the codec has organized the representations. The model activations land in a plane where the four concepts are cleanly separated, and the analogy is visible as geometry.

5.1 Loss Ablation

The full objective deliberately asks for both class separation and analogy geometry. To check how much of the result is imposed by the loss, we reran the codec training under five loss configurations while holding the data, seed, optimizer, learning rate, and number of steps fixed. Table 3 shows the result.

Training objective	Analogy cosine	Accuracy	Axis angle
Full loss	1.0000	100.0%	90.0°
Drop $\mathcal{L}_{analogy}$	0.9951	100.0%	90.0°
Drop $\mathcal{L}_{perp}$	0.9998	100.0%	98.5°
Drop $\mathcal{L}_{analogy}$ and $\mathcal{L}_{perp}$	0.9971	100.0%	96.4°
Classification only	0.9840	100.0%	105.8°

Table 3: Loss ablation with fixed seed, data, optimizer, learning rate, and 3000 training steps. The ablation uses the saved extracted activations and re-trains the small codec for each configuration.

The interesting result is the last row. Even with classification loss alone, the learned two-dimensional space still gives an analogy cosine of 0.9840. The full objective sharpens the geometry and makes the axes nearly orthogonal, but the analogy relation is not created entirely by the analogy term. Some of the structure is already available in the activations and can be recovered by a very small supervised interface.

5.2 Layer Sweep

The original run used one hidden layer, selected by an offset of 10 from the final layer. To check whether the analogy structure is already linearly recoverable before the explicit analogy loss, we repeated the extraction across all decoder-layer outputs and trained a classification-only codec at each layer. Figure 1 shows the analogy cosine and held-out classification accuracy by layer.

This version of the sweep is more informative than the full-loss version: classification accuracy is saturated everywhere, but analogy cosine varies substantially. Early layers, the original layer, and several late-middle layers retain strong analogy geometry under classification-only training. A few middle layers and the final layers are much weaker. This suggests that the analogy relation is not an artifact of one chosen layer, but it is also not uniformly legible throughout the model.

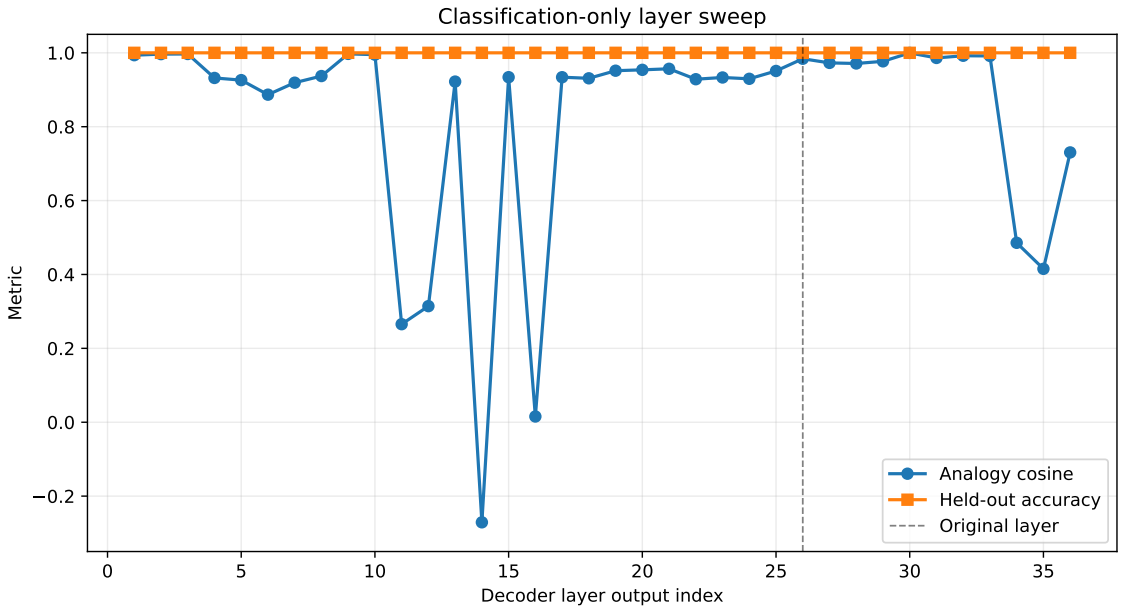

Figure 1: Layer sweep over decoder-layer outputs. The dashed line marks the original layer used in the main experiment. The sweep uses classification loss only, so analogy geometry is not directly optimized.

6 Held-Out Vocabulary

The four training words leave open an obvious concern: perhaps the codec has only memorized four labels. As a first check, we froze the trained codec and extracted new hidden states for eight held-out words using the same sentence template scheme:

$$\{\text{prince}, \text{princess}, \text{boy}, \text{girl}, \text{father}, \text{mother}, \text{duke}, \text{duchess}\}.$$

There was no retraining or refitting. Each held-out word was sampled in 50 contexts, projected through the existing encoder, and compared to the four trained class means.

Word	z_1	z_2	Quadrant	d_k	d_q	d_m	d_w	Nearest
prince	-5.09	-5.59	III	2.68	10.59	13.55	16.85	king
princess	-0.01	-5.45	III	6.73	5.97	14.37	13.94	queen
boy	-2.64	4.07	II	12.61	14.83	4.63	9.09	man
girl	3.46	4.59	I	15.90	13.29	9.25	3.56	woman
father	-1.49	-0.41	III	8.96	10.49	9.14	10.59	king
mother	4.69	-2.08	IV	12.43	6.54	14.07	9.34	queen
duke	-4.16	-4.65	III	3.95	10.03	12.66	15.53	king
duchess	3.03	-2.43	IV	10.81	6.50	13.23	10.02	queen

Table 4: Held-out vocabulary projected through the trained codec. Distances are to the trained class means k = king, q = queen, m = man, and w = woman.

The result is encouraging but not perfectly clean. The strongest pattern is in the royal words: *prince* and *duke* land closest to the king mean, while *princess* and *duchess* land closest to the queen mean. The basic gender words also behave as expected: *boy* is nearest to *man*, and *girl* is nearest to *woman*. The family terms are messier: *father* is pulled toward the king region and *mother* toward

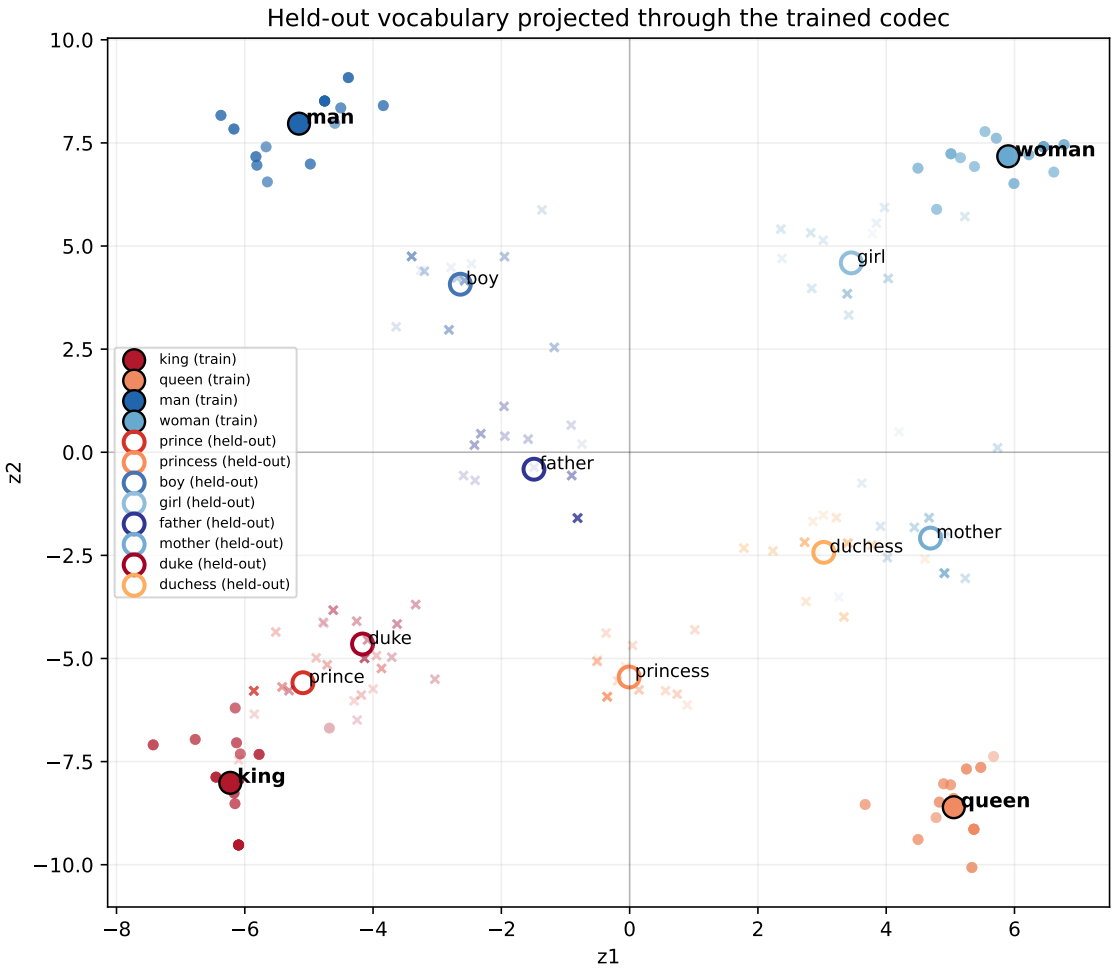

Figure 2: Trained classes are shown with filled markers; held-out classes are shown with open markers. The royal masculine held-out words *prince* and *duke* land nearest to *king*, while *princess* and *duchess* land nearest to *queen*.

the queen region, suggesting that the current plane is not a pure gender axis plus a pure royalty axis. Still, the held-out projection gives evidence that the codec is not merely memorizing four training labels; it transfers some structure to unseen but related vocabulary.

7 What This Shows

The experiment supports three modest claims.

First, the Qwen hidden states contain enough information to distinguish these four word classes across simple contexts. This is expected, but the clean two-dimensional separation makes it directly visible.

Second, the analogy structure can be made explicit with a small supervised projection. The relation does not have to be recovered by inspecting the full hidden state directly; it can be exposed through a tiny learned interface.

Third, the learned space is not just a classifier trick. The geometry has interpretable structure:

the analogy closes, and the two hand-named semantic directions are close to orthogonal.

This is why the experiment is interesting. It suggests that at least some semantic circuits inside a model can be treated as something like an API: a small codec translates opaque activation vectors into coordinates that can be tested, plotted, and reasoned about. This framing is compatible with the broader linear representation hypothesis: useful semantic variables may sometimes be represented in directions or low-dimensional subspaces that simple maps can recover [6].

8 Intervention

The previous sections treat the codec as a readout. A stronger question is whether it can also support a small write operation. We therefore ran a simple pseudoinverse intervention inspired by causal abstraction and model-editing work [8, 9, 10].

Let $g_{2d} = \mu_{\text{woman}} - \mu_{\text{man}}$ be the learned gender direction in the codec space. Starting at μ_{king}, we constructed targets

$$z_\alpha = \mu_{\text{king}} + \alpha g_{2d}, \quad \alpha \in \{0, 0.5, 1.0\}.$$

Using the Moore–Penrose pseudoinverse of the encoder matrix, we mapped the 2D shift back to a minimum-norm hidden-state perturbation:

$$\Delta x_\alpha = (\alpha g_{2d}) W_e^+.$$

We then ran Qwen on the prompt

> I saw the king today. The royal person was a

and added Δx_α to the hidden state of the *king* token at the codec layer before the later tokens attended to it. We measured continuation log-probabilities for six target words.

Continuation	Shift at $\alpha = 0.5$ (nats)	Shift at $\alpha = 1.0$ (nats)
king	-0.051	-0.105
queen	0.285	0.574
prince	-0.027	-0.066
princess	0.223	0.457
man	-0.012	-0.020
woman	0.285	0.559

Table 5: Continuation log-probability shifts relative to $\alpha = 0$. The intervention raises queen/woman/princess continuations and slightly lowers king/prince/man continuations.

This is not a dramatic steering result. The greedy completions were nearly unchanged: at $\alpha = 0$ and $\alpha = 0.5$, the model continued with "bit upset, but I tried to be," while at $\alpha = 1.0$ it continued with "bit upset, and I wanted to help." Still, the log-probability shifts are directional. The pseudoinverse intervention produces the minimum-norm perturbation matching the 2D target, which does not guarantee alignment with the model's natural use of those directions. The monotone downstream effect is therefore a modest piece of evidence rather than a clean steering result. Still, the same 2D shift that moves μ_{king} toward the queen side of the codec also increases the probabilities of queen-related and female-coded continuations in the downstream distribution. This is modest evidence that the codec is not only a readable interface, but a weak writeable one.

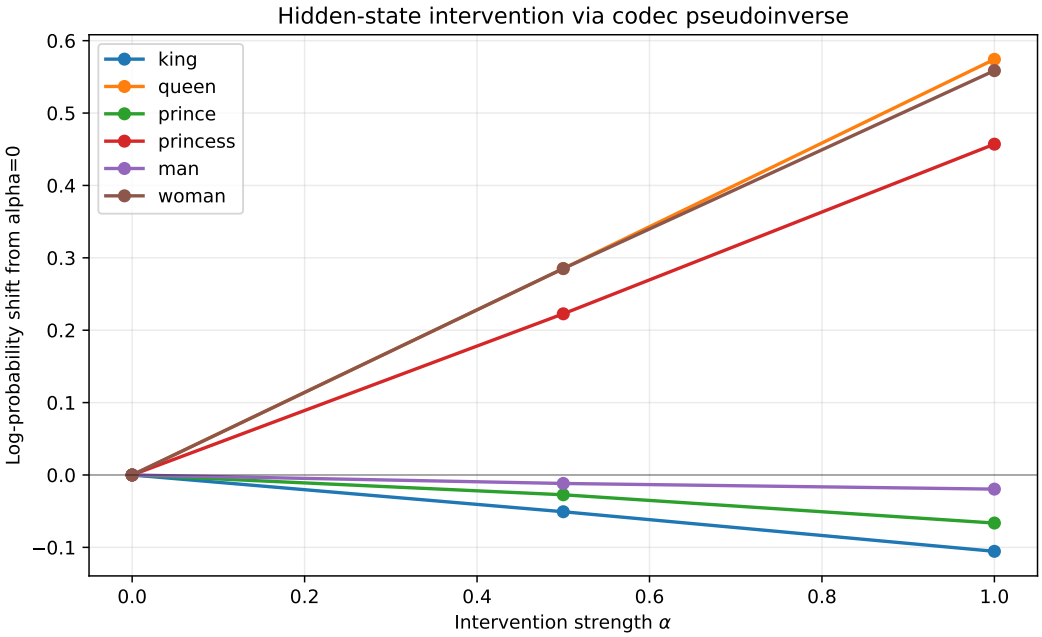

Figure 3: Log-probability shifts under the pseudoinverse hidden-state intervention. The effect is small, but it is monotone and directionally aligned with the intended 2D shift.

9 Limitations

This is a proof of concept, not a broad interpretability result.

The vocabulary is tiny. Four words are enough to test the idea, but not enough to claim that the same two axes generalize across language. A stronger version would include many royal terms, many gendered terms, neutral controls, and examples across languages and syntactic settings.

The classification gate is also not a strict generalization benchmark. The codec is trained using the full extracted dataset, and the held-out split is used afterward to check separability in the learned space. This is acceptable for a first demonstration, but future work should use a cleaner train, validation, and test protocol. More generally, probe results need controls: a sufficiently expressive probe can sometimes learn the task rather than reveal what was already explicit in the representation [4].

The sentence templates are simple. That is useful because it keeps the experiment controlled, but it also means the result may depend on a narrow distribution of contexts. More varied natural sentences would make the test more realistic.

Finally, the axis names should be read carefully. Calling one direction "gender" and another "royalty" is a useful interpretation of this particular four-word geometry. It is not evidence, by itself, that the model has a single global gender direction or royalty direction in the larger hidden space. Superposition also gives a reason to expect useful features to be entangled rather than cleanly isolated in general [7].

10 Next Steps

The most natural next step is to broaden the vocabulary in a more systematic way. The current held-out set is suggestive, but a stronger test would include neutral controls such as *person* and

human, unrelated nouns, and antonyms or role terms that should not be forced into the four training quadrants. Cross-lingual tests are also attractive: if the same codec geometry transfers to Japanese or other languages, the result would be harder to explain as a template-specific English artifact.

The intervention result should also be extended. The present experiment only tests one gender-direction shift at one token and one layer. A stronger followup would test royalty shifts, longer contexts, multiple prompts, and interventions at several layers. It would also compare the pseudoinverse perturbation against more natural perturbations estimated directly from hidden-state differences.

Another useful comparison point is unsupervised dictionary learning, which aims to recover sparse interpretable features from model activations at larger scale [11]. The present codec is much smaller and supervised, but it shares the same practical ambition: turning opaque activation spaces into interfaces with named, testable coordinates.

11 Conclusion

This experiment shows that a very small linear codec can extract a clean, interpretable analogy space from frozen Qwen hidden states. In that space, the relation king $-$ man $+$ woman $\approx$ queen holds almost exactly, the four classes are perfectly separable under a simple nearest-mean classifier, and the main semantic directions are nearly orthogonal.

The result is small but suggestive. It points toward a practical style of mechanistic work where model internals are not only probed, but wrapped in compact interfaces with explicit validation gates. The codec is not only readable: it generalizes partially to unseen royal and gendered vocabulary, and a pseudoinverse intervention produces small but coherent shifts in downstream token probabilities. Together these point toward interfaces that can be read, tested, and, at least weakly, written to.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[2] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751, 2013.

[3] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[4] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2733–2743, 2019.

[5] Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[6] K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[7] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[8] A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.

[9] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

[10] N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

[11] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jermyn, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Hume, J. Bills, B. Lee, N. Schiefer, J. Batson, M. Wattenberg, and C. Olah. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

[12] A. Yang, B. Yang, B. Hui, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.