

Construct Validity of Small-Sample Transformer Scoring Models: A Mechanistic Interpretability Approach

Michael P. Hemenway and Martha Bellows

Pearson

{michael.hemenway, martha.bellows}@pearson.com

Abstract

Transformer scoring models fine-tuned on small samples ($n=64$) can reach deployment-quality agreement with human raters, but agreement statistics cannot reveal whether models lean harder on construct-irrelevant surface features like response length. This paper explores Mechanistic Interpretability (MI) strategies as a framework for auditing scoring engines, evaluating six diagnostic techniques across 90 DeBERTa-v3-base models (6 ASAP2 prompts $\times$ 3 training sizes $\times$ 5 seeds). We distinguish between *correlational* techniques (saliency maps, probing, activation summaries) and *interventional* techniques (counterfactual edits, representational erasure). Correlational methods show suggestive trends—small-sample models exhibit higher length partial correlations ($d = 0.65$) and surface-dominated layer representations ($d = 0.60$)—but suffer from seed instability (80–96% seed noise) and lack power under multiplicity corrections ($q = .17$). Interventional methods provide definitive proof: surgically deleting length directions from a model’s 768-dimensional internal representation destroys 88% of agreement at $n=64$ versus 67% at $n=1024$ ($d = -0.75$, $p = .005$, $q = .049$), an effect independent of accuracy ($r = -0.018$). Counterfactual padding shows that duplicating sentences raises predicted scores for 86–92% of responses across all sizes. We synthesize these findings into an actionable MI audit protocol for measurement practitioners.

1 Introduction

Automated scoring engines are increasingly built by fine-tuning pre-trained transformer language models on small sets of human-scored responses. Recent work demonstrates that with systematic hyperparameter optimization, deployment-quality agreement (quadratic weighted kappa, $QWK \geq 0.70$) is attainable from as few as 64 training re-

sponses (Barber et al., 2025). The operational appeal is obvious: human scoring is the dominant cost in constructed-response assessment.

However, high agreement with human raters does not guarantee construct validity. Agreement is a purely behavioral criterion, and a model can reproduce human scores by exploiting surface heuristics that no scoring rubric endorses. Response length is the classic example: because length correlates strongly with human essay scores ($r \approx 0.70$ pooled), a length-only scorer can achieve non-trivial agreement (Perelman, 2014), and adversarial manipulations can game automated engines (Powers et al., 2002). When fine-tuned on small samples, models have fewer examples to discover complex construct signals and stronger incentives to rely on superficial shortcuts. Standard agreement metrics cannot detect when this occurs.

Introducing Mechanistic Interpretability to Measurement. In an attempt to move from “black box” to “glass box” models, the machine learning discipline has developed *Mechanistic Interpretability* (MI)—a subfield focused on reverse-engineering the internal representations, circuits, and decision logic of neural networks.

This paper serves a dual purpose for measurement scientists and educational ML practitioners. Substantively, we answer a pressing operational question: *Do small-sample scoring models rely more on construct-irrelevant surface features than models trained on larger samples?* Methodologically, we evaluate a battery of MI techniques, illustrating which strategies yield actionable validity evidence and which fail due to noise, unanchored activations, or uncalibrated baselines.

Scope and Validity Argument. We do not claim that MI replaces a comprehensive validity argument. Validity is a property of score interpretation and use, supported by evidence regarding

test content, response processes, internal structure, and consequences (Messick, 1989; Kane, 2013; American Educational Research Association et al., 2014). MI specifically informs two strands: *response processes* (how input features influence scores) and *internal structure* (how concepts are encoded in model representations).

We evaluate 90 models spanning six prompts, three training sizes ($n \in \{64, 512, 1024\}$), and five random seeds, analyzing them with six MI approaches categorized into *correlational* and *interventional* strategies.

2 Mechanistic Interpretability Strategies

We group MI techniques into three categories based on what they measure and why a measurement scientist might use them.

2.1 Data and Model Setup

We analyze the ASAP2 corpus of human-scored essays (scale 1–6) across six diverse prompts (Crossley et al., 2025). For each prompt, we fine-tune microsoft/deberta-v3-base (He et al., 2023) on training splits of $n \in \{64, 512, 1024\}$ across 5 independent random seeds ($6 \times 3 \times 5 = 90$ models). Validation ($n=400$) and test sets ($n=200$) are held fixed per prompt. Models are optimized using Optuna TPE search over 30 trials (Akiba et al., 2019) with frequency-balanced MSE loss, and mapped to score points via post-hoc affine calibration.

2.2 Category A: Feature Association and Probing

Partial Correlations (Behavioral Association).

Rationale: We want to know if model predictions align with surface features (word count, character count, punctuation) beyond what is justified by human scores. We compute the partial correlation $r(\hat{y}, x_j \mid y)$ between model predictions $\hat{y}$ and surface feature x_j , partialling out the human score y . Non-zero partial correlations indicate prediction variance tied to surface shortcuts that human raters do not endorse.

Probing Classifiers (Representational Content).

Rationale: Probing asks what information is stored inside intermediate neural layers (Belinkov, 2022). We extract mean-pooled hidden activations at layers $\{0, 3, 6, 9, 12\}$ and train logistic regression probes to predict quintile buckets of surface features versus construct targets (human score

quintiles). The *probing gap* (surface accuracy minus construct accuracy) tests whether internal representations carry more information about response appearance than essay quality.

2.3 Category B: Saliency and Activation Summaries

Token Attributions (Integrated Gradients & ADI).

Rationale: Attribution methods assign each token a score reflecting its contribution to the final prediction (Sundararajan et al., 2017), computed here using Captum (Kokhlikyan et al., 2020). Measurement researchers often view attribution maps as automated rubric checks. We compute the *Surface Feature Reliance Score* (SFRS) as the share of attribution mass landing on non-content tokens (function words and punctuation). To test consistency, the *Attribution Divergence Index* (ADI) measures token-attribution dissimilarity ($1 - \rho_{\text{Spearman}}$) between two models scoring the exact same response.

Attention Patterns and Logit Lens.

Rationale: Attention heads indicate where models focus, while the *logit lens* (Nostalgebraist, 2020) reads out intermediate predictions at each layer to see how early a score settles. We track attention entropy, content-token focus, and the layer at which predictions reach stability.

2.4 Category C: Interventional Causal Methods

Counterfactual Length Manipulation.

Rationale: Correlational methods show association, not causal use. To test if models *change scores* when length changes, we create sentence-aligned inputs: **Base** (original prefix), **Short** (randomly deletes 30% of sentences), and **Restored** (duplicates remaining sentences to match Base’s length without adding information). The *Length Compensation Effect* (LCE) is the score difference between Restored and Short: score points awarded for length alone.

Representational Erasure (Amnesic Probing).

Rationale: Before predicting a score, DeBERTa compresses a response into a 768-dimensional vector. Using iterative nullspace projection (Ravfogel et al., 2020; Elazar et al., 2021), we identify the directions in this 768-dimensional space encoding response length and surgically remove them before passing the edited vector to the scoring layer. Comparing agreement loss against a

	QWK _{raw}	QWK _{cal}	Exact	≥ 0.70
$n=64$	0.687	0.718 (0.056)	0.561	20/30
$n=512$	0.743	0.771 (0.036)	0.593	29/30
$n=1024$	0.759	0.787 (0.030)	0.602	30/30

Table 1: Test-set agreement across training sizes (mean over 30 models, SD in parentheses). Small-sample models frequently reach deployment thresholds ($\text{QWK} \geq 0.70$), setting up our interpretability comparison.

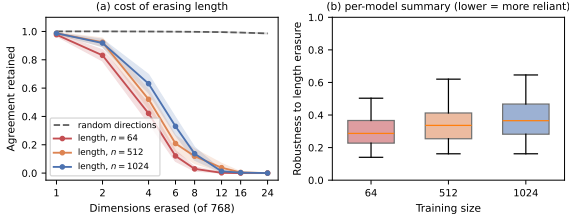


Figure 1: Representational erasure. (a) Agreement retained as length-encoding directions are removed from the 768-dimensional vector, compared to erasing random directions. (b) Per-model robustness ratio (retained agreement under length erasure vs. random control).

control that erases random directions tests whether length representation is causally load-bearing.

3 Substantive Findings

3.1 Behavioral Performance Baseline

Table 1 establishes the baseline. Mean calibrated QWK rises from 0.718 at $n=64$ to 0.787 at $n=1024$, with 20 of 30 small-sample models clearing the 0.70 deployment threshold. Because accuracy differs across sample sizes, we control for QWK throughout our analyses to ensure surface reliance is not merely a proxy for lower agreement.

3.2 Causal Evidence: Length is Load-Bearing

The most decisive finding comes from representational erasure (Figure 1 and Table 2). Erasing random directions from the 768-dimensional pooled vector is nearly harmless, retaining 98–99% of QWK at all sample sizes. In contrast, erasing length directions causes a catastrophic drop in agreement—and differentially so. Erasing just six length directions destroys 88% of human agreement for $n=64$ models (12% retained) versus 67% for $n=1024$ models (33% retained).

Across training sizes, the overall robustness

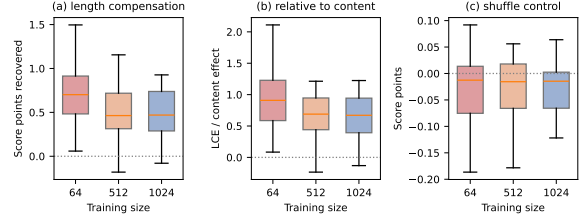


Figure 2: Counterfactual length manipulation. (a) Score points recovered by duplicating sentences. (b) Share of content-deletion penalty recovered. (c) Sentence-shuffle control.

score (retained agreement under length erasure divided by random control) rises from 0.299 at $n=64$ to 0.379 at $n=1024$ ($d = -0.75, p = .005, q = .049$). Crucially, this robustness score is uncorrelated with test QWK ($r = -0.018$). When restricting analysis to an accuracy-matched band ($0.697 \leq \text{QWK} \leq 0.804$), small-sample models remain significantly more length-dependent ($d = -0.81, p = .003$). This proves that small-sample models causally depend far more on internal length representations, independent of model accuracy.

3.3 Causal Evidence: Vulnerability to Surface Padding

The counterfactual edit provides a direct estimate of gaming vulnerability in score points (Figure 2 and Table 2). Restoring deleted length with duplicated sentences recovers +0.68 score points at $n=64$ compared to +0.48 points at $n=1024$ ($d = +0.54, p = .042$). Across all training sizes, 86–92% of individual essay predictions increase when padded with restatements, returning 89% of the content-deletion penalty at $n=64$ against 63% at $n=1024$.

Two controls confirm specificity: deleting content causes a uniform score drop (0.75–0.78 points across sizes), and shuffling sentence order has virtually no impact (-0.04 points), proving models respond to length and redundancy rather than general perturbation.

3.4 Correlational Evidence: Association and Probing

The correlational methods align with the causal findings but carry weaker statistical power.

Partial Correlations. Residual length correlation (Figure 3) drops from 0.555 at $n=64$ to 0.460 at $n=1024$ ($d = 0.65, p = .014$). Matching mod-

Measure	$n=64$	$n=1024$	d	90% CI	p_{diff}	q_{BH}	Verdict
Length compensation effect (score pts)	0.680	0.484	+0.54	[+0.11, +0.97]	.042	.087	suggestive
Length compensation per 100 tokens	0.545	0.393	+0.52	[+0.09, +0.95]	.048	.087	suggestive
Length reliance ratio	0.887	0.630	+0.52	[+0.09, +0.95]	.048	.087	suggestive
Content effect (score pts)	0.775	0.752	+0.18	[-0.26, +0.61]	.497	.527	inconclusive
Sentence-shuffle effect (score pts)	-0.031	-0.044	+0.16	[-0.27, +0.60]	.527	.527	inconclusive
Directions lost before agreement halves	4.933	6.133	-0.59	[-1.02, -0.16]	.025	.087	suggestive
Robustness to length erasure	0.299	0.379	-0.75	[-1.18, -0.31]	.005	.049	difference
Agreement lost after erasing 8 directions	0.704	0.677	+0.18	[-0.26, +0.61]	.497	.527	inconclusive
$r(\text{prediction, word count})$	0.777	0.753	+0.36	[-0.07, +0.79]	.170	.256	inconclusive

Table 2: Causal interventional experiments comparing $n=64$ and $n=1024$ models. Rows 1–5 report counterfactual length manipulations; rows 6–9 report amnesic representational erasure. Positive d indicates greater length reliance at $n=64$ (except rows 6–7 where lower values indicate greater reliance).

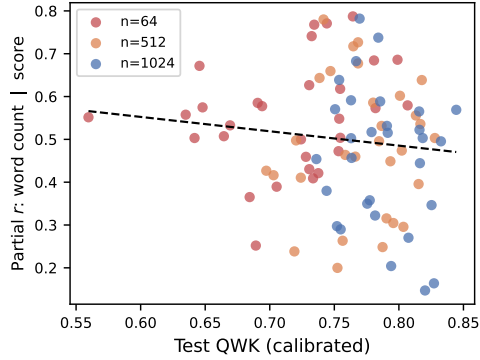


Figure 3: Residual length reliance (partial correlation between prediction and word count, controlling for human score) versus test QWK across all 90 models.

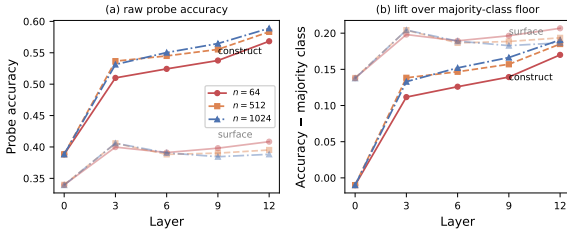


Figure 4: Probing classifier accuracy by layer. (a) Raw accuracy. (b) Lift over majority-class chance floor (construct floor = 0.398, surface floor = 0.202).

els on QWK preserves this difference (0.577 vs 0.481, $d = 0.64$, $p = .022$), confirming that two models with identical human agreement can differ in surface reliance.

Probing Gap. Scoring layer probes as lift over chance floors (Figure 4) shows that the probing gap shrinks from +0.148 at layer 0 (frozen embeddings) to +0.037 at layer 12 for $n=64$, +0.009 for $n=512$, and -0.004 for $n=1024$ ($d = 0.60$, $p = .023$). Fine-tuning converts a surface-dominated representation into a construct-dominated one, but only larger training sets complete this conversion

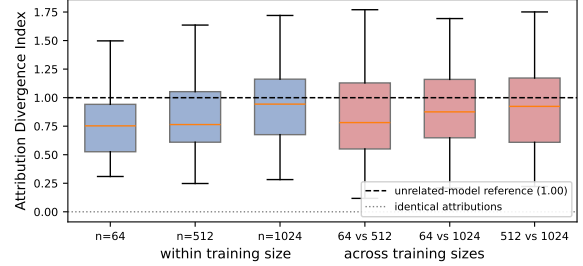


Figure 5: Attribution Divergence Index (ADI) between model seeds vs. between training conditions. Token attributions between seeds are as divergent (0.853) as attributions between sample sizes (0.881).

at the output layer.

4 Methodological Audit of MI Strategies

A key goal of this study is evaluating MI tools for educational assessment settings. Table 3 synthesizes our findings into a diagnostic guide for automated scoring contexts.

Lesson 1: Token Saliency Maps Suffer from Seed Instability.

Gradient-based token attributions (Integrated Gradients) are often promoted as automated rater checks. However, comparing attribution vectors across models reveals extreme instability (Figure 5). The divergence between two models trained on identical data with different random seeds (ADI = 0.853) is virtually indistinguishable from the divergence between small and large models (ADI = 0.881, $d = 0.08$, $p = .350$). Saliency maps reflect random initialization noise as much as learned representations (Kindermans et al., 2019; Adebayo et al., 2018). *Recommendation:* Token attribution maps must never be interpreted without reporting a same-condition seed baseline.

MI Strategy	Measurement Rationale	Empirical Result ($n=64$ vs 1024)	Diagnostic Utility & Recommendations
Partial Correlations (<i>Residual Alignment</i>)	Test if predictions align with surface features after controlling for human scores.	Monotonic decrease: $r = .555 \rightarrow .460$ ($d = 0.65, p = .014$).	Suggestive clue: Measures association, not causation; vulnerable to length-construct collinearity and battery multiplicity ($q = .17$).
Probing Classifiers (<i>Layer Representation</i>)	Audit whether internal layers encode surface features vs. construct targets.	Chance-corrected gap: $+.078 \rightarrow +.054$ ($d = 0.60, p = .023$).	Requires chance floors: Raw accuracy is misleading due to score distribution skew. Score probes as lift over majority-class floor.
Token Attributions (<i>Saliency / ADI / SFRS</i>)	Inspect token-level attributions for function/punctuation vs content tokens.	Seed vs seed ADI = 0.853; condition ADI = 0.881 ($d = 0.08$).	High seed instability: Saliency maps vary wildly across random seeds (80–96% seed noise). Uninformative without a seed baseline.
Attention & Logit Lens (<i>Internal Flow</i>)	Track head entropy, content focus, and intermediate prediction settlement.	Content mass $d = -0.08$; stabilization layer $d = +0.09$.	Unanchored activations: Activation summaries lack task controls; power was inadequate for unconstrained summaries ($d_{\min} = 0.74$).
Counterfactual Edits (<i>Length Padding</i>)	Test if adding length without information alters predicted scores.	Score recovery: $+0.68$ vs $+0.48$ pts; 86–92% gain score from padding.	Actionable in score points: Directly measures model gaming vulnerability and susceptibility to restatements.
Representational Erasure (<i>Amnesic Probing</i>)	Surgically delete length directions from 768-d pooled representation.	Agreement retained: 12% vs 33% ($d = -0.75, p = .005, q = .049$).	Causally diagnostic: Proves length is load-bearing; independent of QWK ($r = -0.018$) and holds under QWK-matching ($d = -0.81$).

Table 3: Diagnostic evaluation of Mechanistic Interpretability (MI) techniques for measurement applications. Interventional strategies provide actionable validity evidence, whereas unanchored correlational summaries suffer from seed noise and multiplicity.

Lesson 2: Probing Classifiers Require Chance Floors. Raw probing accuracy suggested models encode construct better than surface features. However, construct targets (score quintiles collapsed to 4 categories) carried an artificial chance floor of 0.398, whereas surface targets (balanced quintiles) carried a floor of 0.202. Raw probing accuracy was completely uninterpretable. Subtracting chance floors reversed the conclusion, showing surface dominance in early and middle layers (Figure 4b), aligning with control-task principles for probing (Hewitt and Liang, 2019; Voita and Titov, 2020). *Recommendation:* Probing accuracies must always be reported as lift over majority-class chance baselines.

Lesson 3: Unconstrained Batteries Suffer Multiplicity. Running a broad battery of 15 correlational metrics across activations without a pre-specified primary endpoint leads to severe multiplicity penalties under procedures such as Benjamini–Hochberg (Benjamini and Hochberg, 1995). While individual partial correlations and probing gaps showed nominal significance ($p \leq .023$), none survived false-discovery-rate correction ($q = .17$). In contrast, causal interventions

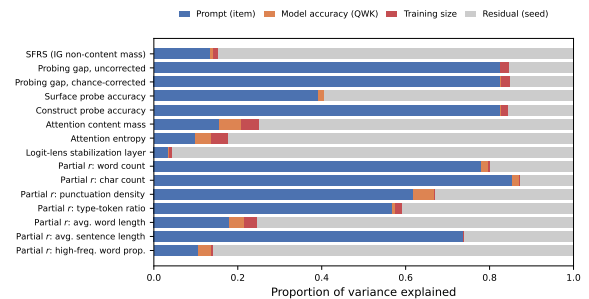


Figure 6: Variance decomposition across prompt identity, model accuracy (test QWK), training size, and seed noise.

yield strong, unconfounded signals that survive correction.

Lesson 4: Item Identity Dominates Variance. Partitioning variance across all metrics (Figure 6) reveals that prompt/item identity accounts for 50–83% of total variance, whereas training size accounts for only 0.1–5%. *Recommendation:* Audit protocols must evaluate models across diverse items rather than pooling across prompts.

5 Practical Guidance for Practitioners

Actionable Audit Protocol. For measurement programs evaluating automated scoring engines, we recommend a 4-step validity audit:

1. **Measure Seed Noise:** Establish same-condition seed variance before drawing conclusions about model differences.
2. **Apply Counterfactual Padding Tests:** Test model predictions against restatements and duplicated text to quantify gaming vulnerability in score points.
3. **Perform Representational Erasure:** Use nullspace projection on pooled embeddings to verify whether surface shortcuts are causally load-bearing, independent of QWK.
4. **Evaluate Item-Level Variation:** Organize validity checks by item rather than relying on aggregate summary metrics.

Limitations. ASAP2 carries a single human score per response, preventing human–human baseline comparisons. While ASAP2 includes student demographic metadata (e.g., race/ethnicity, ELL status, disability, and economic status), subgroup fairness analysis was beyond the scope of this mechanistic length-reliance audit; future work will conduct intersectional subgroup analyses. Erasure removes linear directions, serving as a lower bound on feature dependence.

6 Conclusion

Mechanistic interpretability provides powerful tools for validating automated scoring engines, but implementation must be carefully designed and can benefit from collaboration between AI Scientists and Psychometricians. Small-sample transformer scoring models ($n=64$) reach high human agreement but causally depend more on construct-irrelevant surface length representations than models trained on larger samples. While passive correlational summaries suffer from seed noise and multiplicity, interventional techniques—specifically representational erasure and counterfactual edits—offer robust, actionable validity evidence for assessment.

References

- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. [Sanity checks for saliency maps](#). In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 9505–9515, Montréal, Canada.
- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. [Op-tuna: A next-generation hyperparameter optimization framework](#). In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Justin O. Barber, Michael P. Hemenway, and Edward Wolfe. 2025. [When does active learning actually help? Empirical insights with transformer-based automated scoring](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 1–8, Pittsburgh, Pennsylvania. National Council on Measurement in Education.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. [A large-scale corpus for assessing source-based writing quality: ASAP 2.0. Assessing Writing](#), 65:100954.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

- Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.
- Pieter-Jan Kindermans, Sara Hooker, Julius Adedbayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. [The \(un\)reliability of saliency methods](#). In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer, Cham.
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. [Captum: A unified and generic model interpretability library for PyTorch](#). arXiv preprint arXiv:2009.07896.
- Samuel Messick. 1989. Validity. In Robert L. Linn, editor, *Educational Measurement*, 3rd edition, pages 13–103. American Council on Education / Macmillan, New York, NY.
- Nostalgebraist. 2020. [Interpreting GPT: The logit lens](#). LessWrong.
- Les Perelman. 2014. [When “the state of the art” is counting words](#). *Assessing Writing*, 21:104–111.
- Donald E. Powers, Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. 2002. [Stumping e-rater:challenging the validity of automated essay scoring](#). *Computers in Human Behavior*, 18(2):103–134.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML’17*, pages 3319–3328. JMLR.org.
- Elena Voita and Ivan Titov. 2020. [Information-theoretic probing with minimum description length](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.