Probing and Interpretability in Encoder-Decoder Models

HeiCAD Colloquium

Akhilesh Kakolu Ramarao

Department of English Language and Linguistics, Heinrich Heine University Düsseldorf

June 12, 2026

1.1. About me

Akhilesh Kakolu Ramarao

- I am a PhD student in the department of General Linguistics under the supervision of Prof. Dr. Kevin Tang, Prof. Dr. Wiebke Petersen, and Dr. Dinah Baer-Henney.
- I work on computational morphology – modeling morphological patterns using transformers
- Before academia, I worked in industry as a software engineer and a NLP researcher (for about 5 years).
- To know more about me: https://akkikek.xyz/about

1.2. Two parts

I	**≈ 15 min**	**Pair-work**
II	**≈ 30 min**	**Open discussion**
III	**≈ 15 min**	**Taxonomies to *think with***

≈ 15 minutes

3.1. Q1 | Can you explain a single output?

Could you explain why your model produced *that* output, and not another?

Some methods:

- **Input attribution**: which tokens drove it (saliency, LIME, SHAP) (Ribeiro 2016; Sundararajan 2017)
- **Contrastive**: why *this* token rather than that one (Yin & Neubig 2022)
- **Per-token attribution** for generation (Inseq: Sarti 2023)

3.2. Q2 | Did looking inside ever help?

When a model of yours misbehaved, did you look *inside*? Did it actually help?

Most debugging is still retrain, reweight, reprompt. But we can look inside using methods such as:

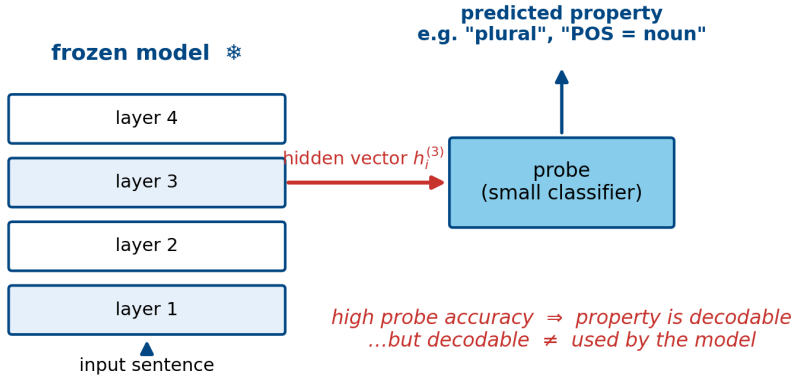
- **Context mixing**: which tokens really informed a representation (Mohebbi 2023) (raw attention will not tell you (Jain & Wallace 2019))
- **Activation patching / causal tracing**: localize where a behaviour lives (Vig 2020; Meng 2022)
- **Logit lens**: watch a prediction form across layers (nostalgebraist 2020; Tuned Lens: Belrose 2023)

3.3. Q3 | What counts as evidence?

"The model has learned / encodes / understands X." What evidence do you require?

Does the method **observe** or **intervene**?

- **Probing** shows a property is *decodable* (Tenney 2019) but decodable is not used
- **Amnesic probing**: remove it, see if behaviour breaks (Elazar 2021)
- Demand **faithful, complete, minimal** (Jacovi & Goldberg 2020) and **sanity checks** (Adebayo 2018)



probing reads a property out of frozen states

3.4. Q4 | Which tools do you trust?

Attention maps, saliency, probing accuracy, chain-of-thought, asking the model. Which would you put in a paper as evidence?

Each has a known failure mode:

- **Attention weights**: not explanation (Jain & Wallace 2019)
- **Saliency**: methods disagree (Atanasova 2020), some fail sanity checks (Adebayo 2018)
- **Chain-of-thought / self-explanations**: plausible, not faithful. The model's "I said that because.." is the same next-token machinery, optimized to convince (Turpin 2023)
- **Probing accuracy**: needs control tasks / MDL (Hewitt & Liang 2019; Voita & Titov 2020)

3.5. Q5 | Why access isn't understanding

With a model you have every weight, activation, and gradient. Why does that not settle it?

Access is necessary, not sufficient. Understanding means finding the right **abstraction**:

- **Sparse autoencoders** try to pull them apart into readable features (Cunningham 2024; Bricken 2023)
- **Circuits**: the unit that turns "all the numbers" into an algorithm (Olah 2020; Wang 2023)

3.6. Q6 | Does the "why" matter?

If your model tops the benchmark, does it matter *why* it works?

- **Robustness**: right for the wrong reasons collapses under distribution shift (Ribeiro 2016)
- **Trust & accountability**: high-stakes deployment, fairness
- **Science**: what did it actually learn about language?
- **Transfer**: does a finding on a toy model hold at scale? Circuits sometimes recur (Lieberum 2023), yet identical behaviour can hide *different* algorithms (Friedman 2024)

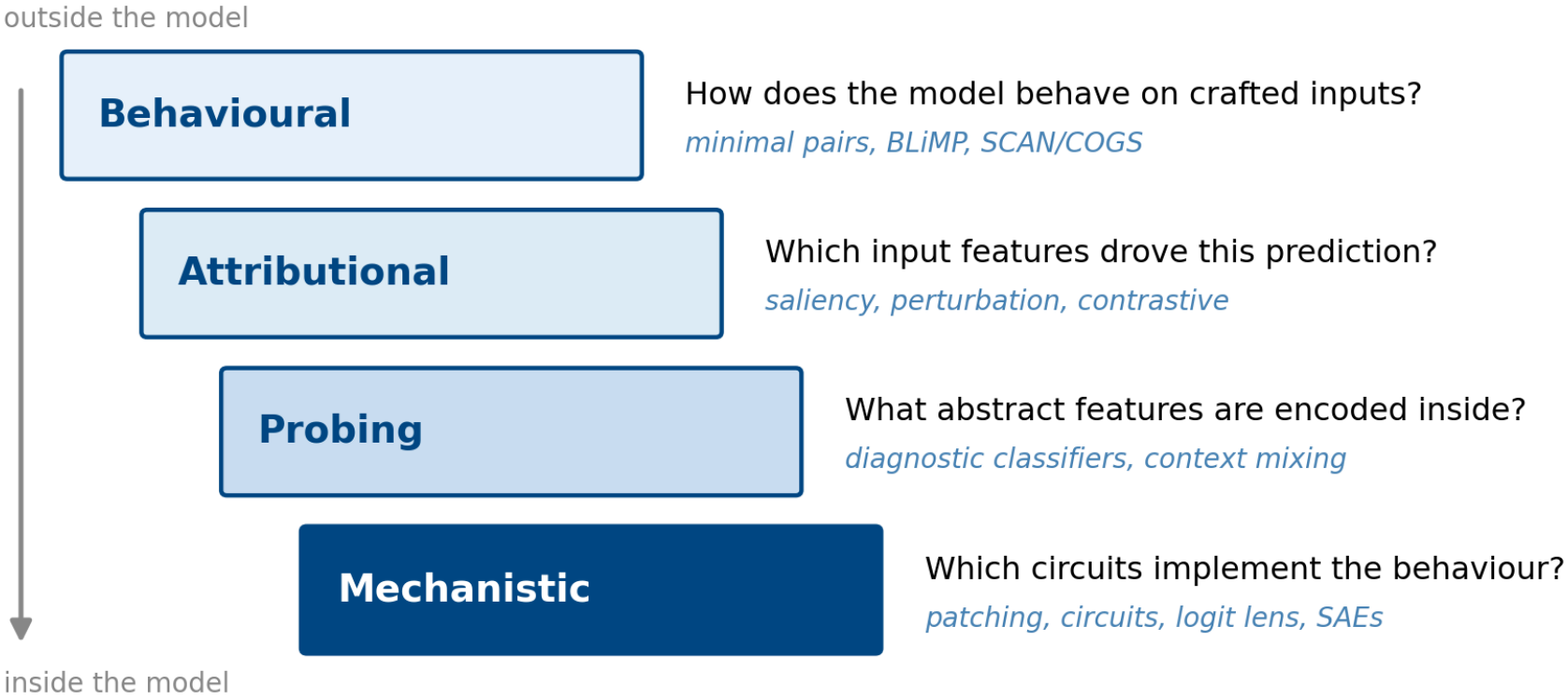
4.1. Marr's three levels

"How deep do you look?" goes back to David Marr: **any information-processing system** must be understood at **three loosely independent levels** (Marr 1982).

Computational	*what* problem is solved, and *why*? the goal and the logic behind it
Algorithmic	*how*: what representations and what procedure turn input into output?
Implementational	what *physical substrate* carries it out?

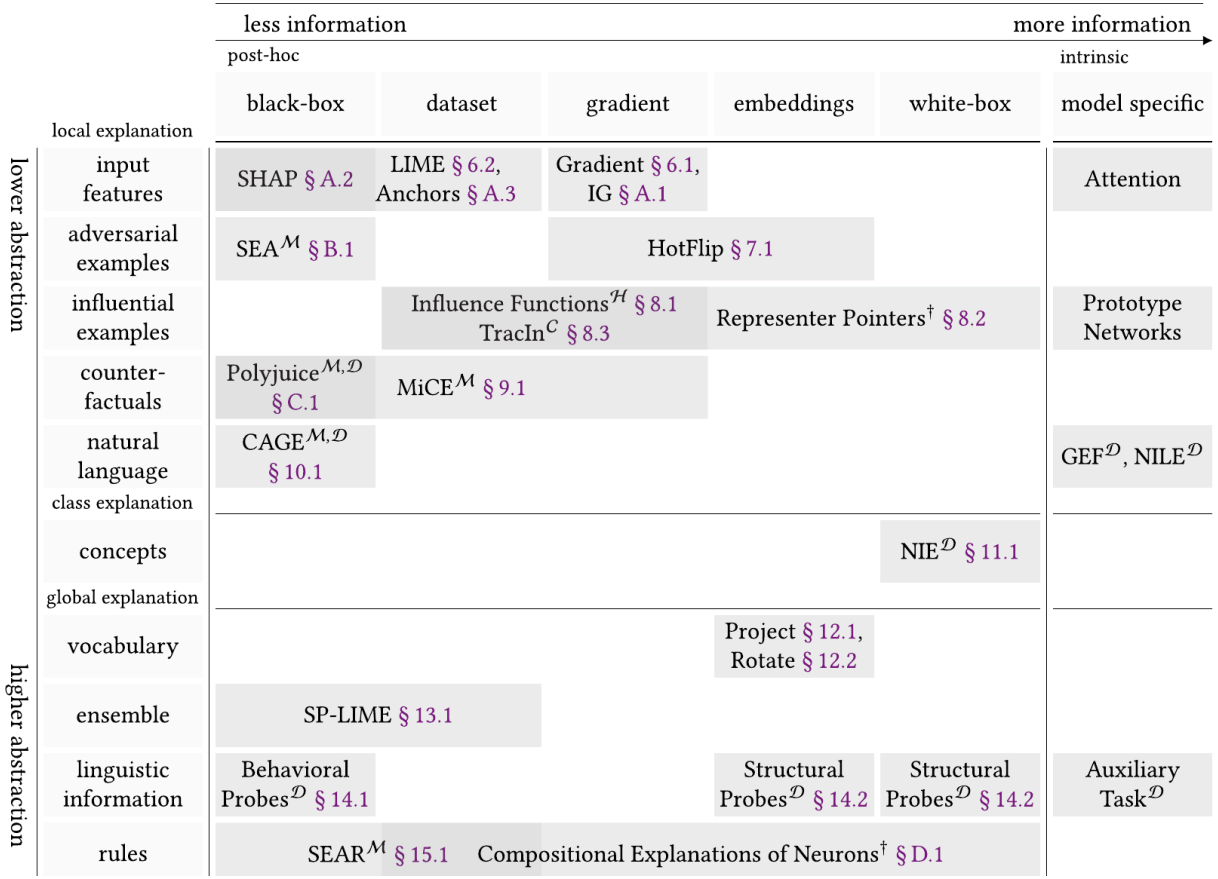
"Understanding perception by studying only neurons is like understanding bird flight by studying only feathers. You need aerodynamics first." (Marr 1982)

4.1. Marr's three levels



Sort by **depth**: how far inside you look, from behaviour to circuits.

4.1. Marr's three levels



Sort by **explanation** (Madsen 2022).

4.2. References

Motivation, faithfulness, taxonomy. Jacovi & Goldberg (2020). Towards Faithfully Interpretable NLP Systems. *ACL*. | Marr (1982). *Vision*. | Madsen, Reddy & Chandar (2022). Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*. | Ribeiro et al. (2016). LIME. *KDD*. | Turpin et al. (2023). Language Models Don't Always Say What They Think. *NeurIPS*.

Attribution. Sundararajan et al. (2017). Integrated Gradients. *ICML*. | Yin & Neubig (2022). Contrastive Explanations. *EMNLP*. | Atanasova et al. (2020). A Diagnostic Study of Explainability Techniques. *EMNLP*.

Probing. Tenney et al. (2019). BERT Rediscovers the Classical NLP Pipeline. *ACL*. | Hewitt & Liang (2019). Control Tasks. *EMNLP*. | Voita & Titov (2020). MDL Probing. *EMNLP*. | Elazar et al. (2021). Amnesic Probing. *TACL*.

Context mixing. Jain & Wallace (2019). Attention is not Explanation. *NAACL*. | Mohebbi et al. (2023). Value Zeroing. *EACL*; Homophone Disambiguation in Speech Transformers. *EMNLP*. | Sarti et al. (2023). Inseq. *ACL demo*.

Mechanistic. Olah et al. (2020). Zoom In: Circuits. *Distill*. | Vig et al. (2020). Causal Mediation Analysis. *NeurIPS*. | Meng et al. (2022). ROME. *NeurIPS*. | Wang et al. (2023). Interpretability in the Wild (IOI). *ICLR*. | nostalgebraist (2020). The Logit Lens. | Belrose et al. (2023). The Tuned Lens. *arXiv*. | Cunningham et al. (2024). Sparse Autoencoders. *ICLR*. | Bricken et al. (2023). Towards Monosemanticity. *Anthropic*.

Validity & transfer. Adebayo et al. (2018). Sanity Checks for Saliency Maps. *NeurIPS*. | Lieberum et al. (2023). Does Circuit Analysis Interpretability Scale? *arXiv*. | Friedman et al. (2024). Interpretability Illusions in the Generalization of Simplified Models. *ICML*.