AIMLAI Tutorial

Part 1: Interpretable Machine Learning for Sequential Data

Jérôme Fink

UNIVERSITÉ DE NAMUR

Outlines

- Explainability taxonomy and its ambiguities
- What are sequential data?
- Explainability methods for sequential data

Feel free to interrupt me!

Explainability Taxonomy

Field of explainability is broad and straddles several fields

- **Machine Learning / AI**: How to provide models that the user would trust
- **Human Computer Interfaces**: What is the best form of explainability for the end user? How to display it?
- **Law**: How to develop explainability methods that are compliant with regulations?

The field is also evolving. Models start to explain themselves (or do they? Stay tuned!)

Explainability Taxonomy

The vocabulary used in each field is not always the same.

The focus and concern of each domain are not the same neither.

Therefore it is hard to construct a unique taxonomy for the field of interpretable ML.

Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

3 approaches were identified:

- Functioning-based approach
- Result-based approach
- Conceptual approach

Functioning Based Taxonomy

Classifies the explainability methods by looking at how they work.

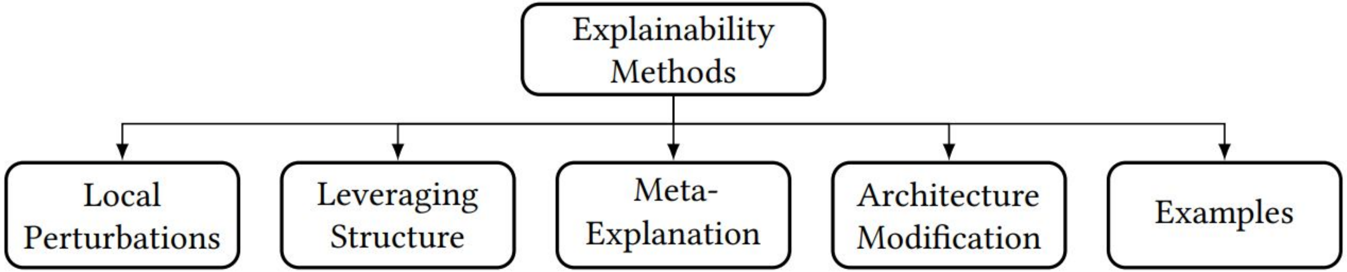


Figure reproduced from Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

Result Based Taxonomy

Sort the explainability methods by their results

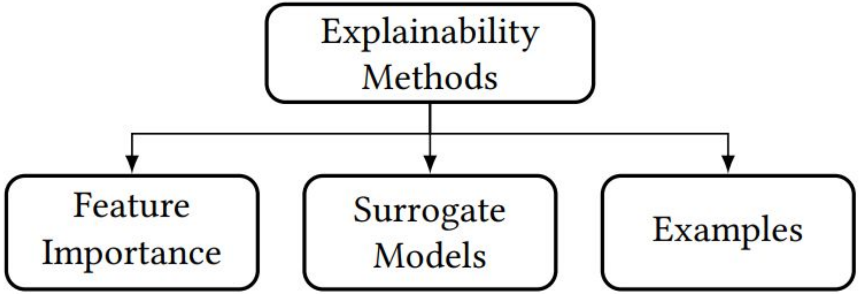


Figure reproduced from Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

Conceptual Taxonomy

More complex taxonomy mapping the explainability methods to several concepts.

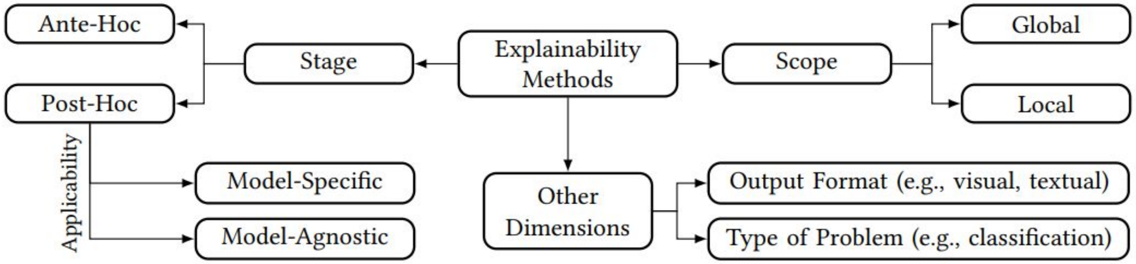


Figure reproduced from Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

Proposed Unified Taxonomy

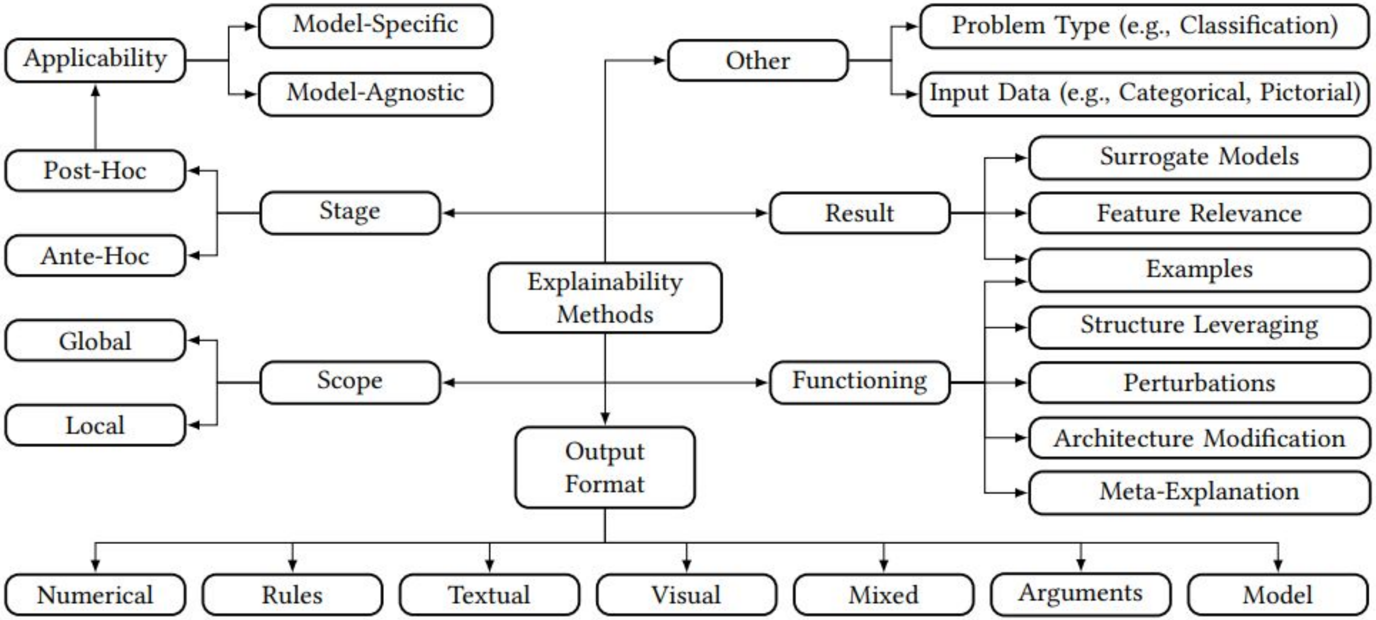


Figure reproduced from Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

My Taxonomy

I figured out that I naturally classify the methods using a subset of the conceptual approach.

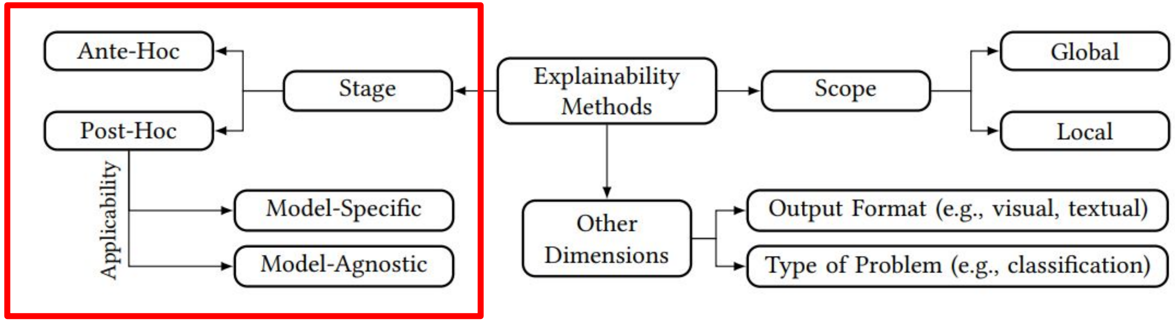


Figure reproduced from Speith, T. (2022). **A review of taxonomies of explainable artificial intelligence (XAI) methods.** In *Proceedings of the ACM conference on fairness, accountability, and transparency* (pp. 2239-2250).

My Taxonomy

Ante-hoc: Methods implemented before training the model to improve its explainability

Post-hoc: Trying to make sense of an already trained model

- **model-specific**: methods only applicable to a specific family of models
- **model-agnostic**: methods that does not rely on a specific family of models

I am biased toward gradient-based approaches!

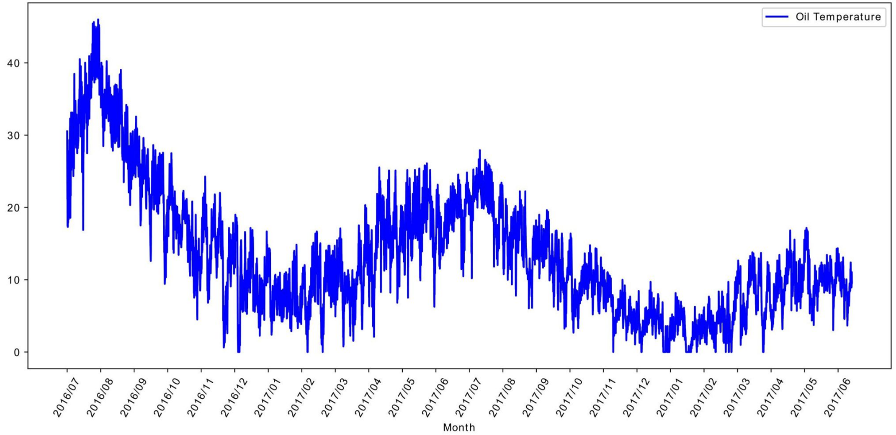
Model Taxonomy

Do you have any questions or comments?

What are Sequential Data?

Sequential Data are data arranged in a sequence where order matters.

Typically time ordered data (a.k.a time series)



Sample of the ETTh1 dataset

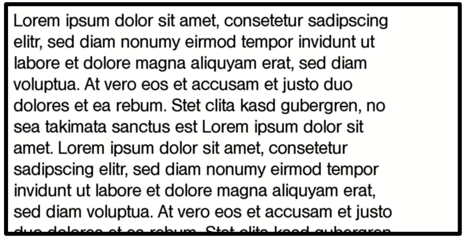
What are Sequential Data?

Sequential Data are data arranged in a sequence where order matters.

But not only…

Textual data:

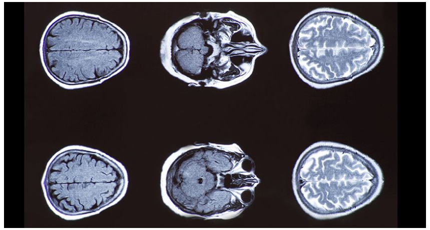
“the cat eats the fish” ≠ “the fish eats the cat”


Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo

Biological:

DNA Sequence: The sequence determine the protein

CT-Scan: Scan layer by layer

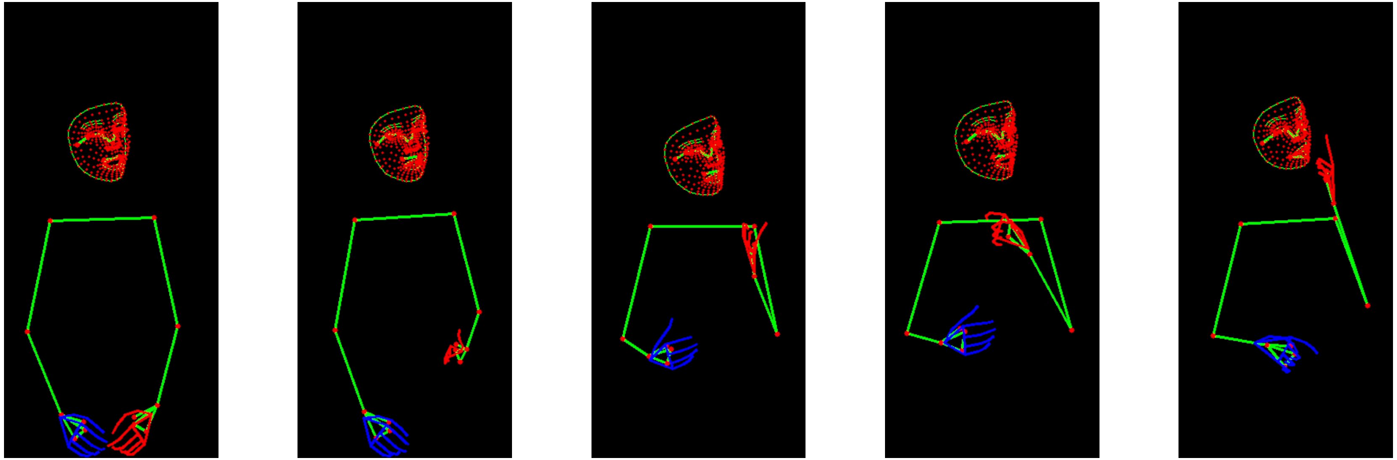


What are Sequential Data?

Sequential Data are data arranged in a sequence where order matters.

But not only…

Video data:



What are Sequential Data?

Sequential Data are data arranged in a sequence where order matters.

But not only…

Audio data:

Audio is a special case as a lot of methods process the spectrogram as an image!

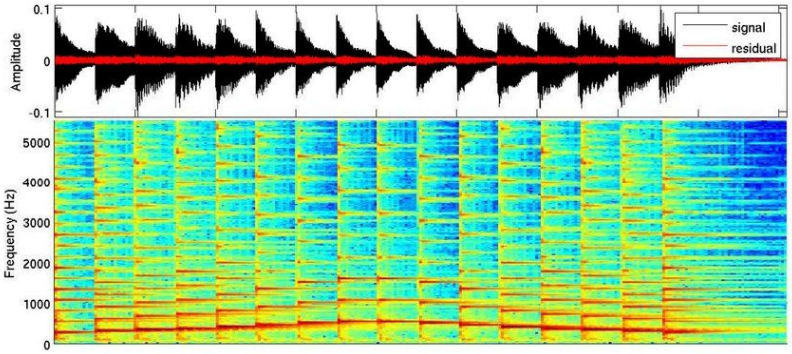


Figure reproduced from FOURER, D. (2009). **Amélioration et évaluation de systèmes de transcription polyphonique.**

What are Sequential Data?

Sequential Data are data arranged in a sequence where order matters.

But not only…

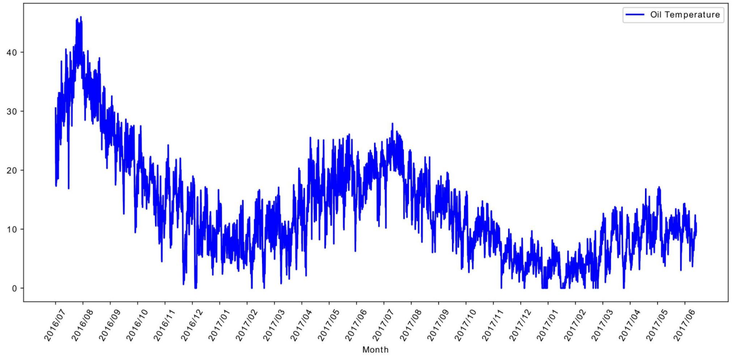
Image data:

Images can be seen as a sequence of pixels and, therefore, can be considered as a sequence.

But in which order should the pixels be read?

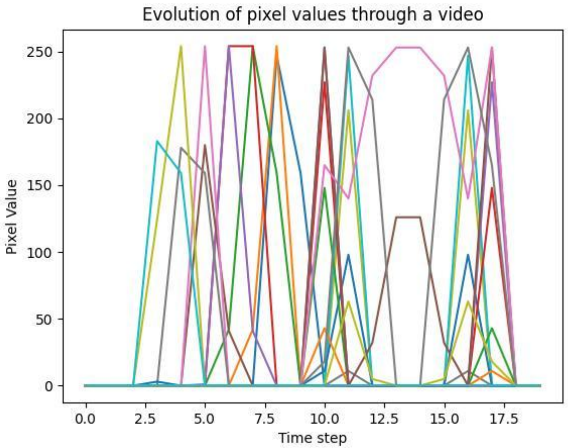
What are Sequential Data?

Univariate



Sample of the ETTh1 dataset

Multivariate



Moving MNIST dataset

What are Sequential Data?

Features of sequential data are strongly correlated.

It is harder to perform data augmentation on sequential data.

You must be careful when perturbing the input.

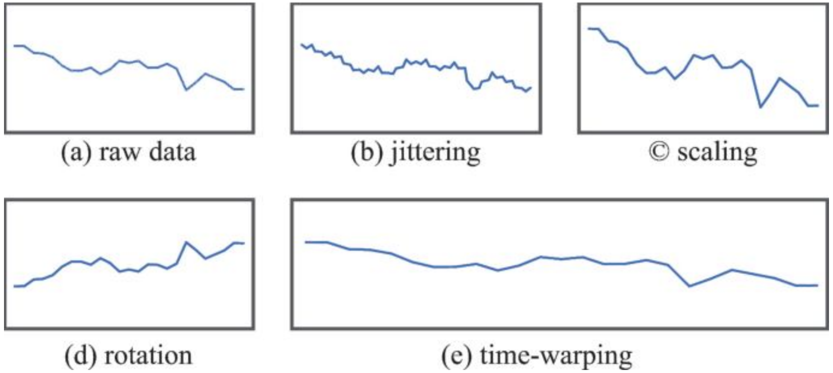


Figure reproduced from Flores, A., Tito-Chura, H., & Apaza-Alanoca, H. (2021). **Data augmentation for short-term time series prediction with deep learning**. In *Arai, K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, 284. Springer, Cham.*

Autoregressive Property

$$X_1 \Rightarrow X_2 \Rightarrow X_3 \Rightarrow \dots \Rightarrow X_t$$

An sequence is *autoregressive* when the point X_n could be predicted using the point X_{n-1}

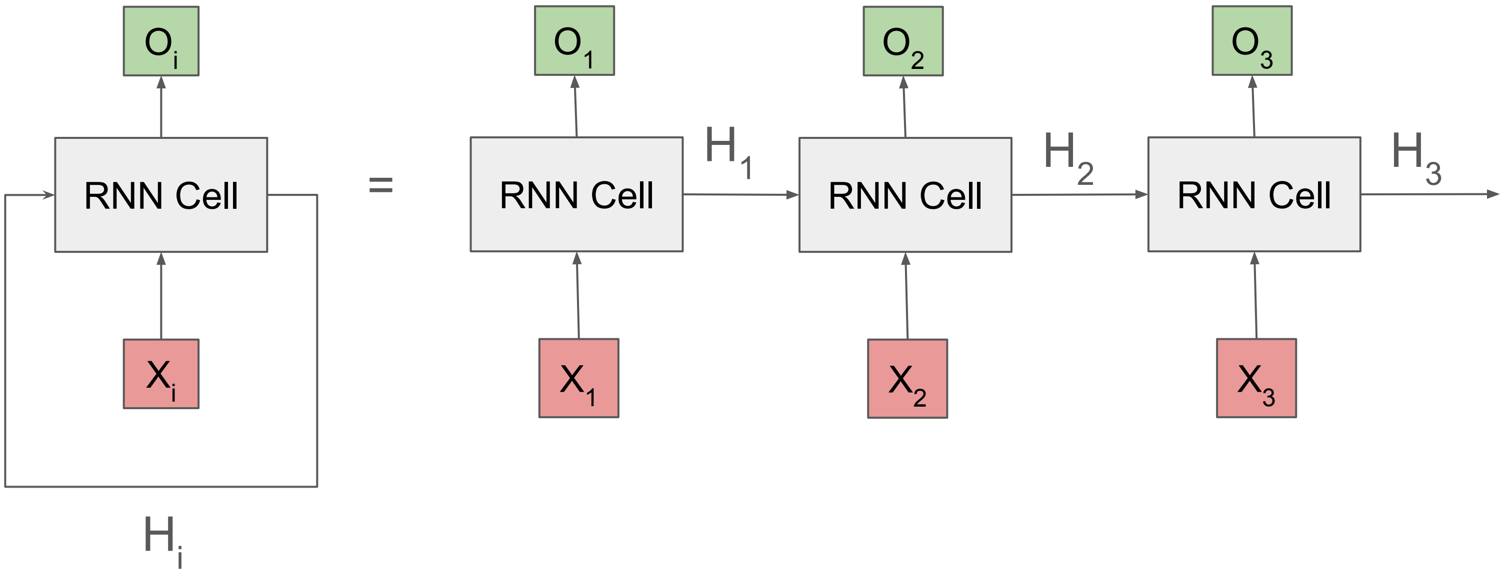
This property is handy for **online inference**

It is an inherent property of several architecture designed for sequence.

Property exploited by PixelRNN a first noticeable approach to generate images.

Sequential Architecture

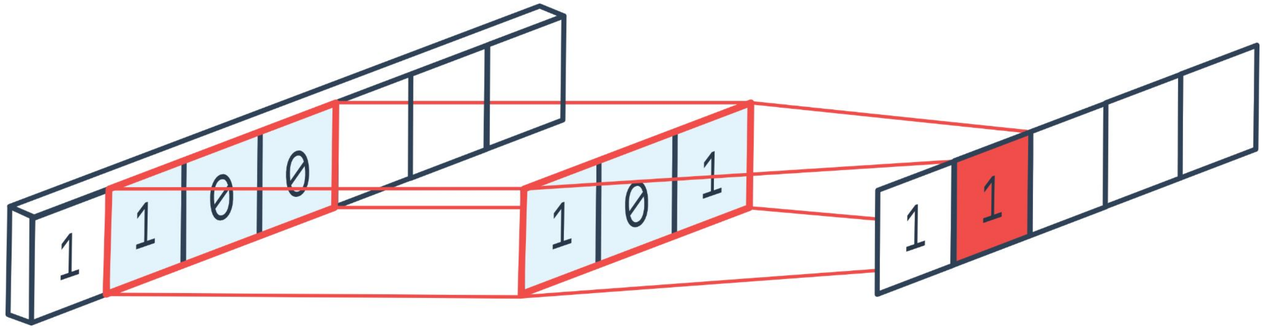
Recurrent Neural Network and co (LSTM): autoregressive



Sequential Architecture

Transformer architecture (attention mechanism): Not autoregressive by default

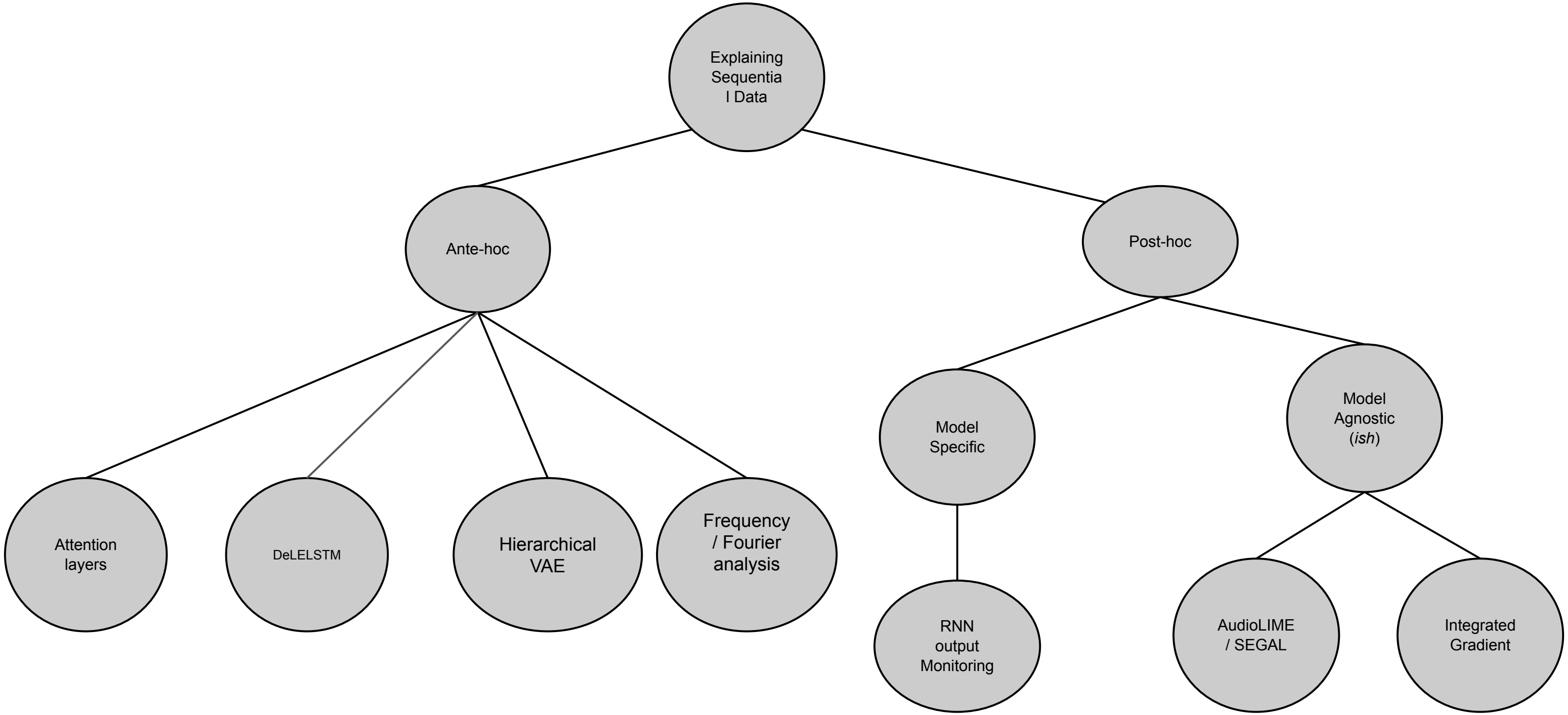
Convolutional models: Not autoregressive by default



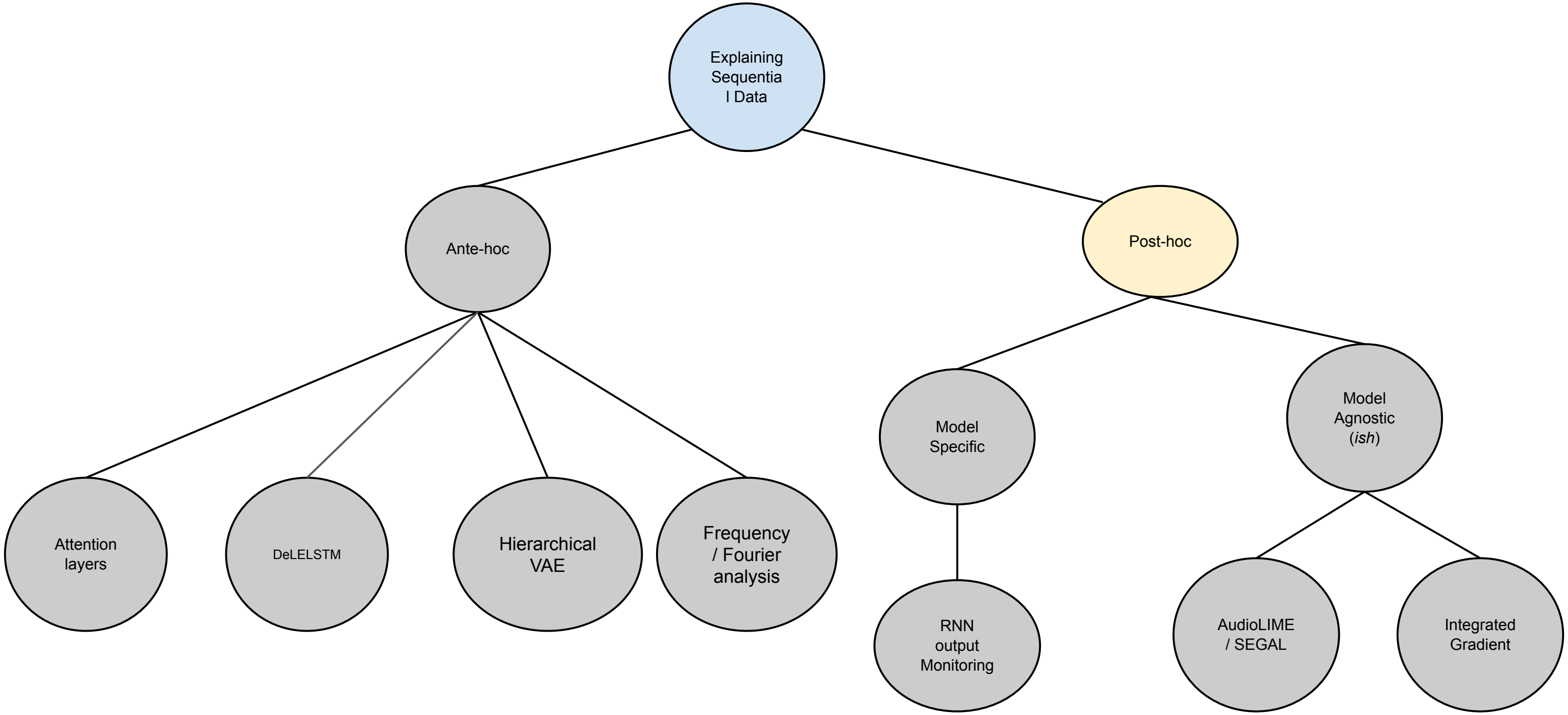
Sequential Data & Architecture

Do you have any questions or comments?

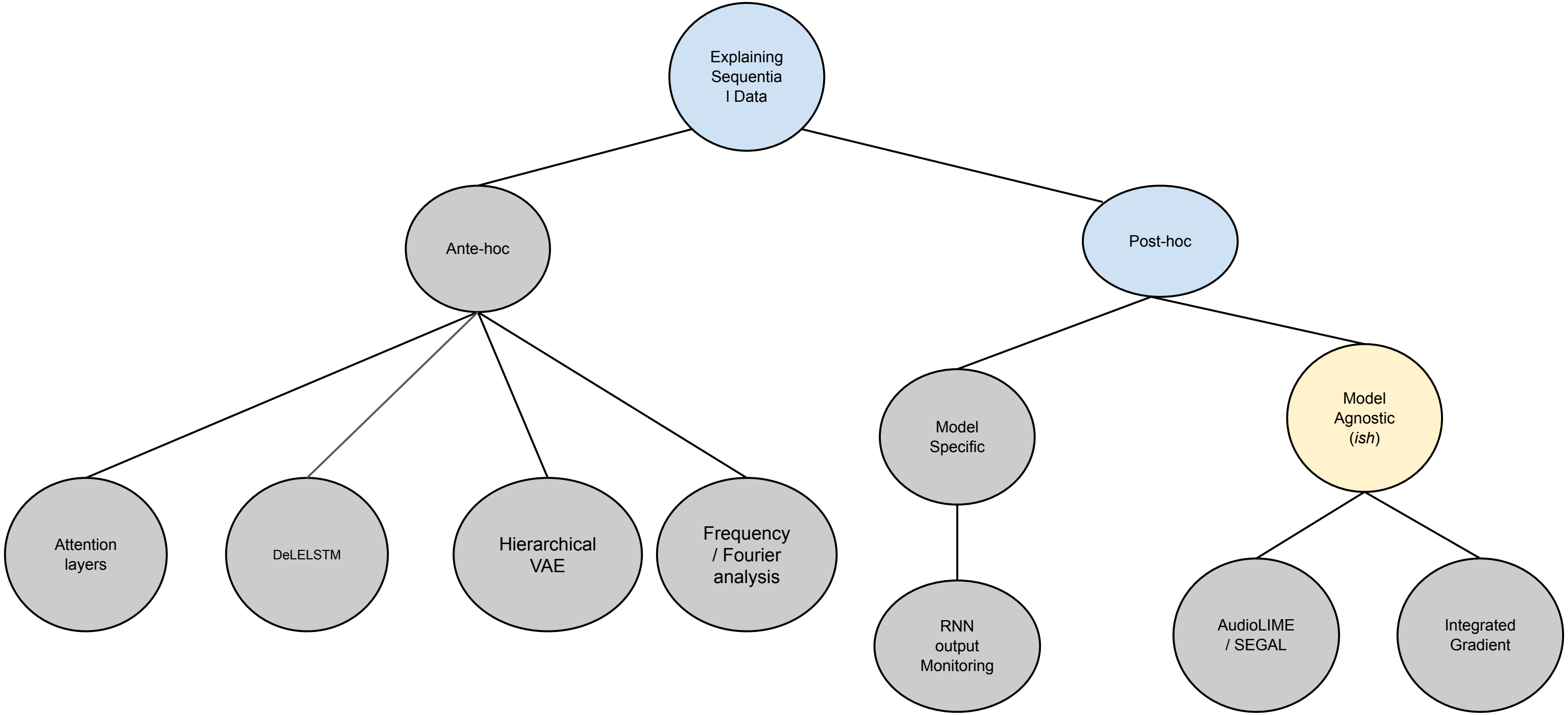
Interpretability of Sequential Data



Interpretability of Sequential Data



Interpretability of Sequential Data



Integrated Gradient & Shapley

- Local explanation
- Compute a saliency for the input data
- Can be leveraged for any model or task

Those methods proved to be particularly efficient to highlight relevant features used by LSTMs. Less conclusive on transformer models.

Turbé, H., Bjelogrlic, M., Lovis, C., & Mengaldo, G. (2023). **Evaluation of post-hoc interpretability methods in time-series classification**. *Nature Machine Intelligence*, *5*(3), 250-260.

Integrated Gradient

Approach leveraging signal from the back-propagation to compute an attribution of a feature to the output. Relies on two axioms :

- **Sensitivity** : *"An attribution method satisfies Sensitivity if, for every input and baseline that differ in ont feature but have different prediction, the differing feature should be given an non-zero attribution."*
- **Implementation Invariance** : "*Two networks are functionally equivalent if their outputs are equal for all inputs, despite having different implementation. Attribution method should satisfy implementation invariance, i.e., the attribution are always identical for two functionally equivalent networks."*

Sundararajan, M., Taly, A., & Yan, Q. (2017). **Axiomatic attribution for deep networks**. In *International conference on machine learning* (pp. 3319-3328).

Integrated Gradient

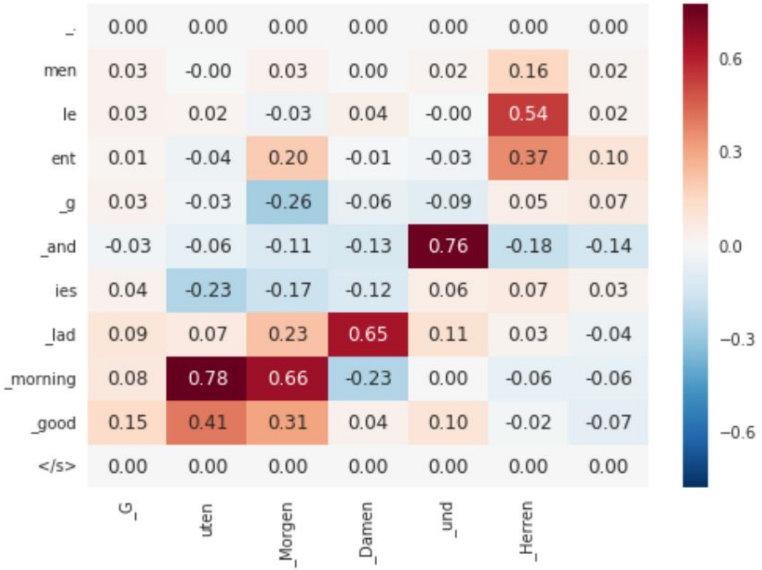


Figure reproduced from Sundararajan, M., Taly, A., & Yan, Q. (2017). **Axiomatic attribution for deep networks**. In *International conference on machine learning* (pp. 3319-3328).

Shapley

Also compute an attribution value for each feature given an output. Not restricted to gradient-based method.

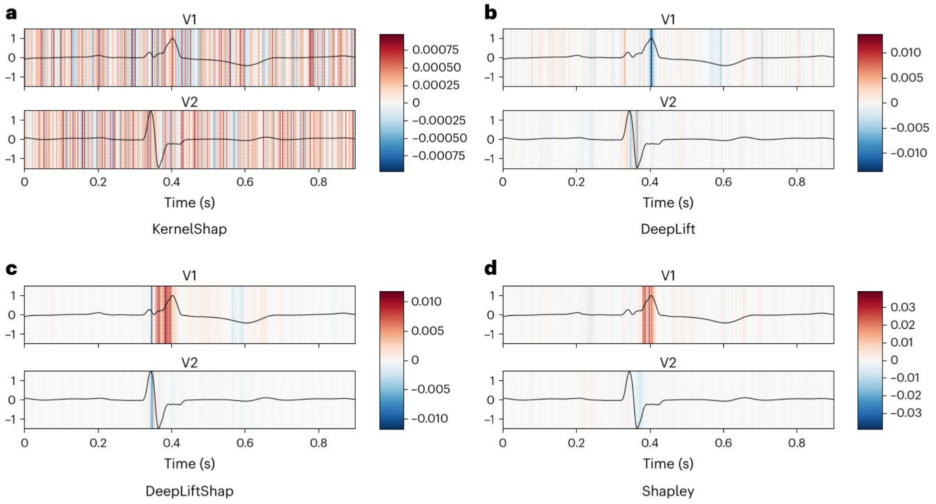


Figure reproduced from Turbé, H., Bjelogrlic, M., Lovis, C., & Mengaldo, G. (2023). **Evaluation of post-hoc interpretability methods in time-series classification**. *Nature Machine Intelligence*, *5*(3), 250-260.

Audio Lime / SEGAL

- Two specialized version of Lime for different kinds of sequential data
- Local explanation
- Highlight the interest of adapting existing methods to the kind of data you manipulate

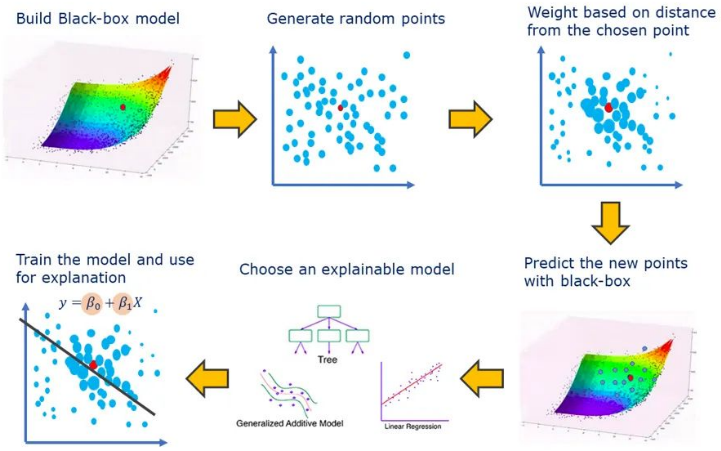
Haunschmid, V., Manilow, E., & Widmer, G. (2020). **audioLIME: Listenable explanations using source separation**. *arXiv:2008.00582*.

Meng, H., Wagner, C., & Triguero, I. (2024). **SEGAL time series classification – Stable explanations using a generative model and an adaptive weighting method for LIME**. *Neural Networks*, *176*, 106345.

LIME

Local Interpretable Model-agnostic Explanation

Provide local explanation for a given point in the dataset by constructing an interpretable surrogate model



Picture by Giorgio Visani

Audio Lime

The audio signal is divided into several channels. Those are used as feature by the LIME algorithm. The end user can listen to the most relevant feature.

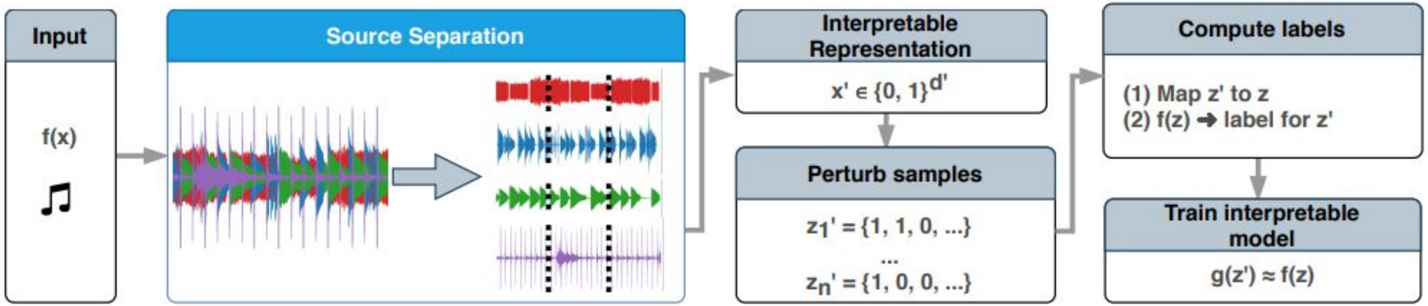


Figure reproduced from Haunschmid, V., Manilow, E., & Widmer, G. (2020). **audioLIME: Listenable explanations using source separation**. *arXiv:2008.00582*.

SEGAL

LIME create samples without considering the data distribution.

Stable Explanation using Generative model and an Adaptive weighting method for Lime (SEGAL) uses generative model to create altered version of sequences (multivariate time series).

SEGAL

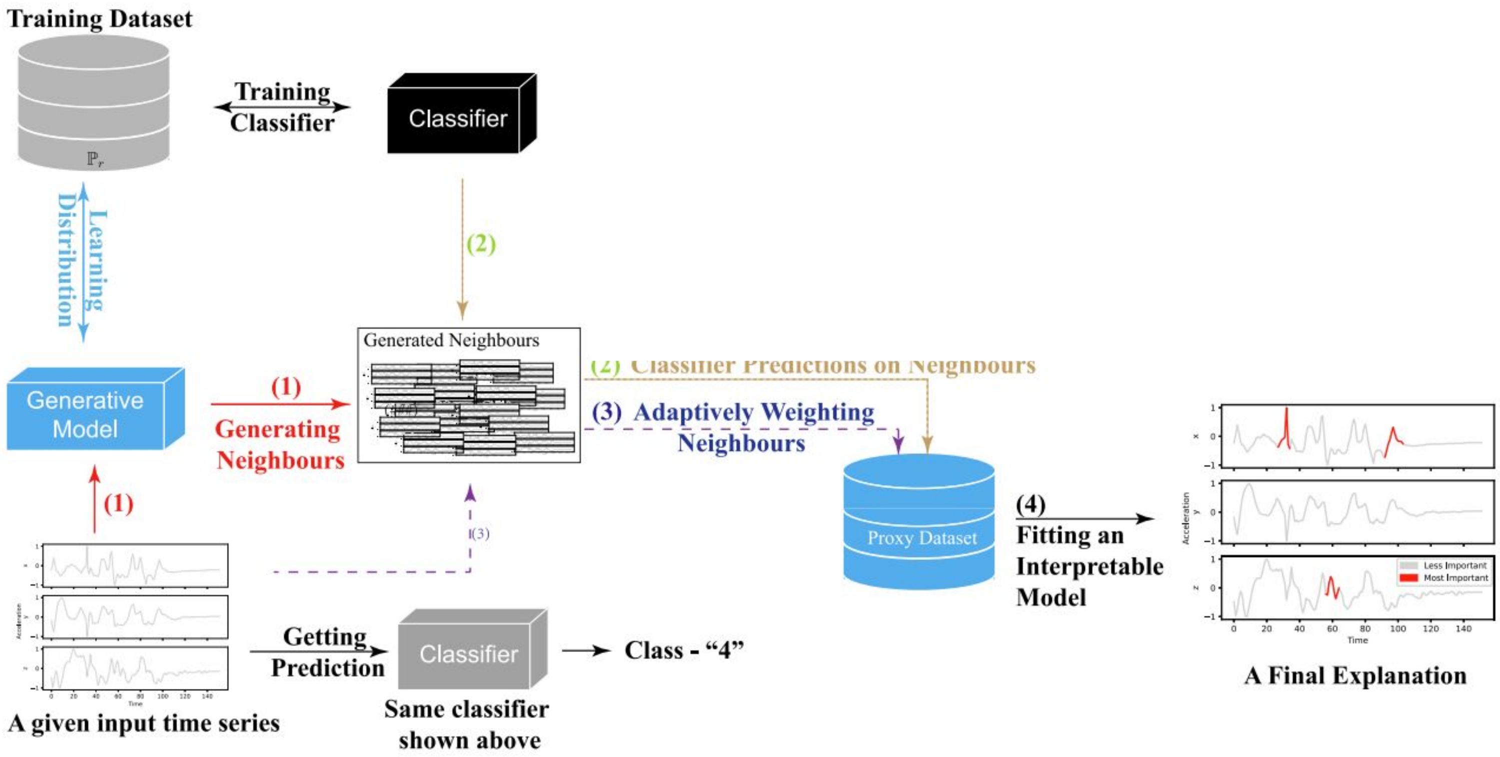
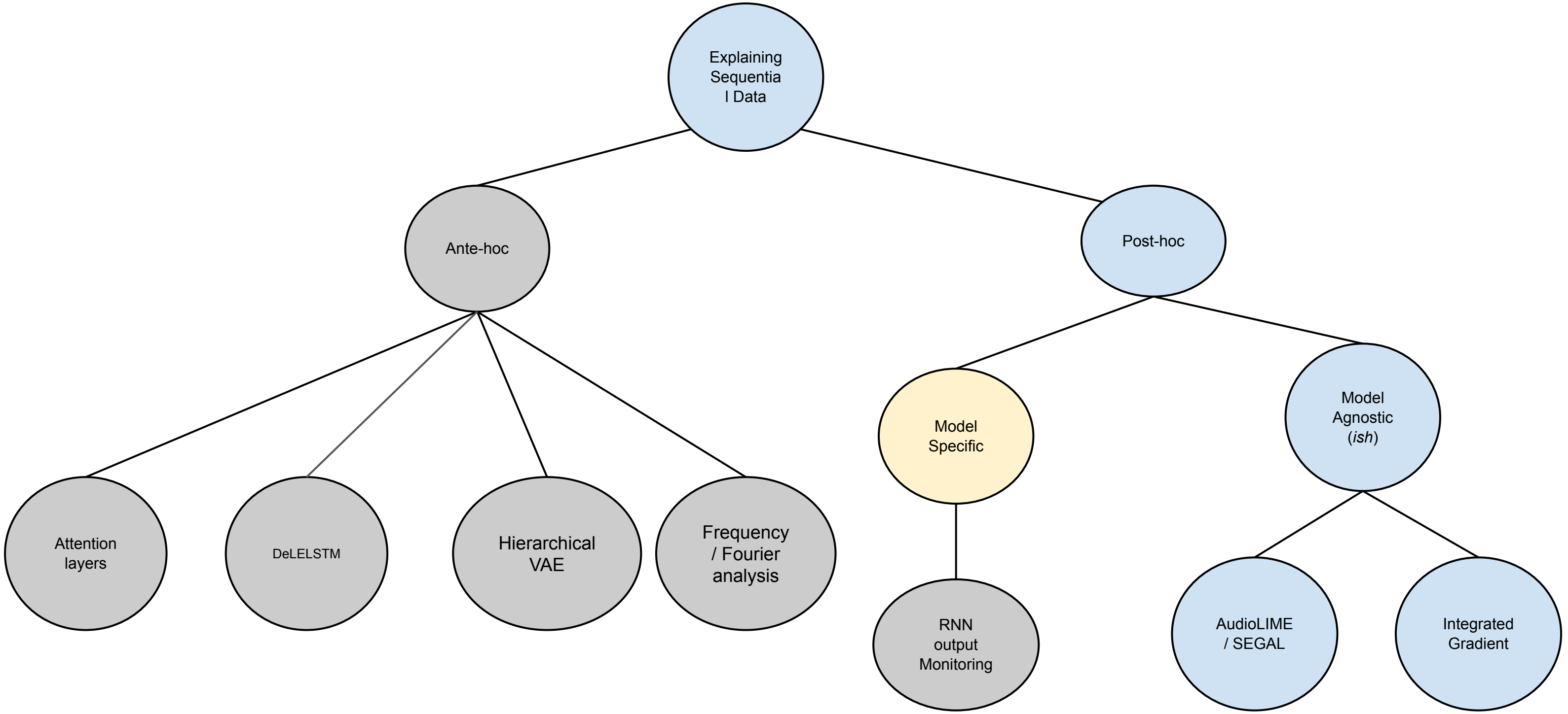


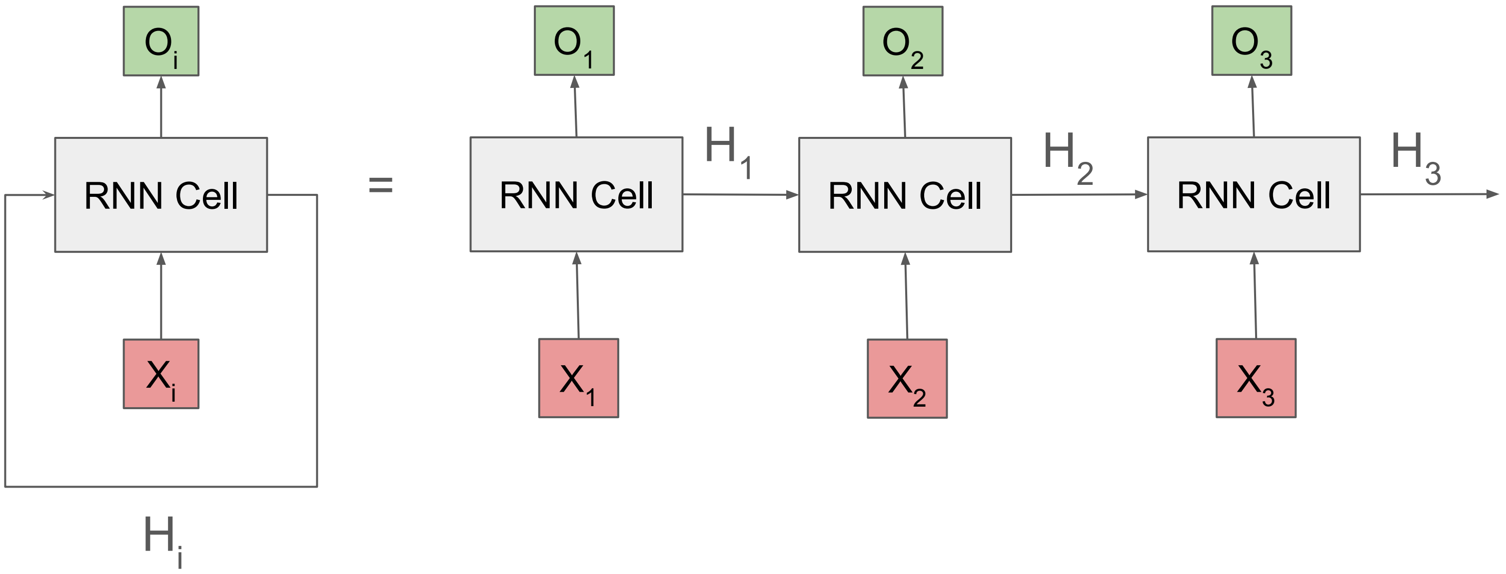
Figure reproduced from Meng, H., Wagner, C., & Triguero, I. (2024). **SEGAL time series classification—Stable explanations using a generative model and an adaptive weighting method for LIME**. *Neural Networks*, *176*, 106345.

Interpretability of Sequential Data



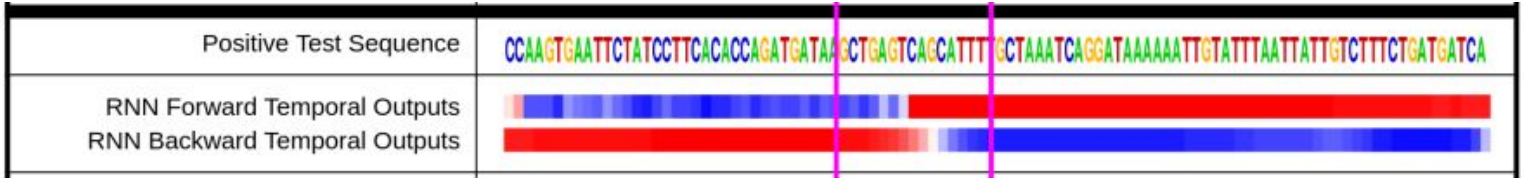
Output Monitoring

RNN outputs a lot of information while processing a sequence



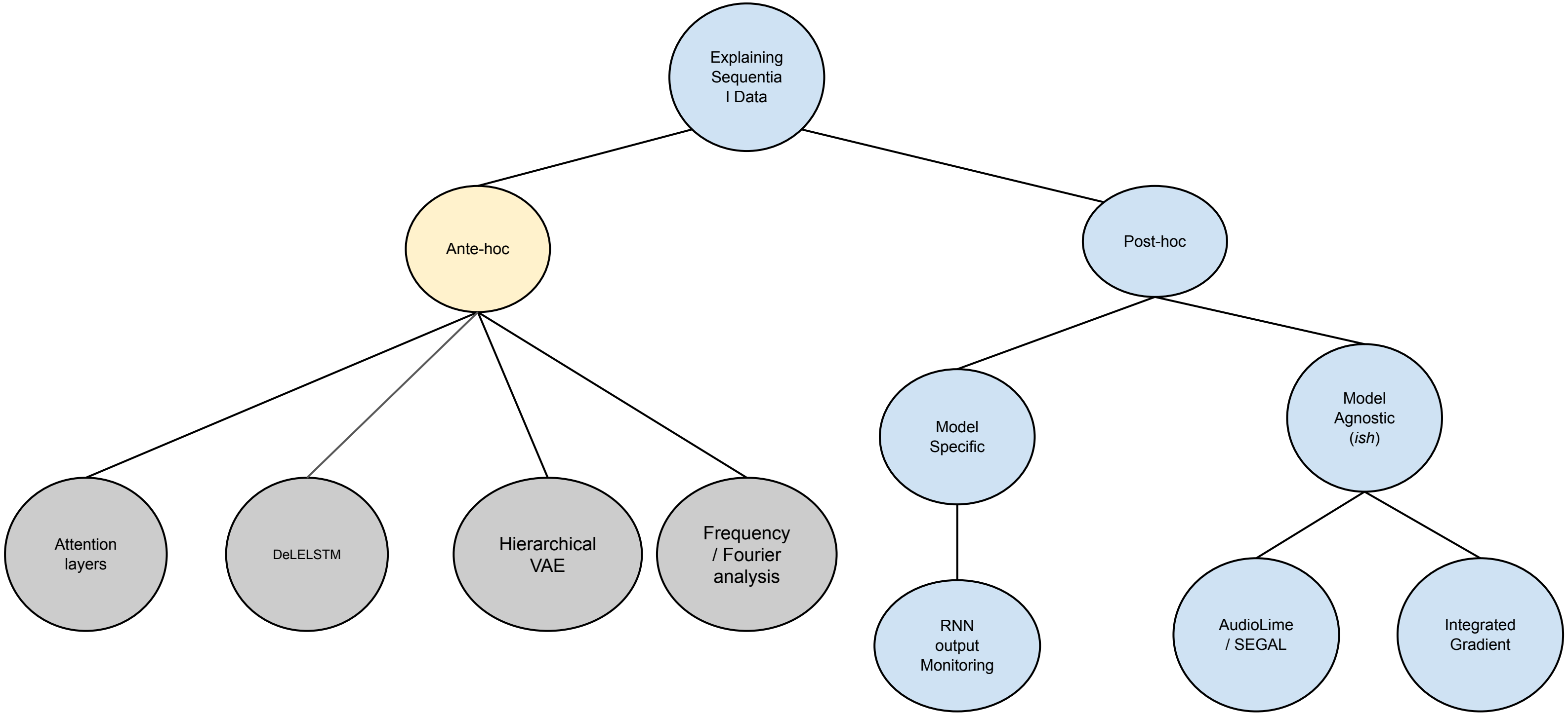
Output Monitoring

Monitoring fluctuation in the output allows to identify interesting part of the sequence.



Example reproduced from Lanchantin, J., Singh, R., Wang, B., & Qi, Y. (2017). **Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks.** In *Pacific symposium on biocomputing 2017* (pp. 254-265).

Interpretability of Sequential Data



Ante-hoc methods

Instead of trying to interpret a black box, we could create more transparent boxes.

Rudin, C. (2019). **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**. *Nature machine intelligence*, 1(5), 206-215

"[...] Trying to *explain* a black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society"

Cynthia Rudin

Black-box Interpretation Issues

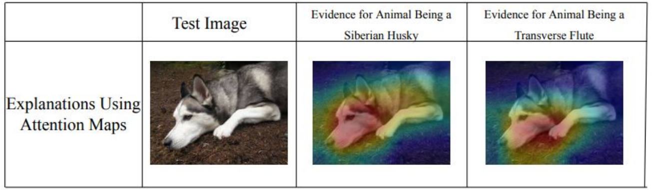


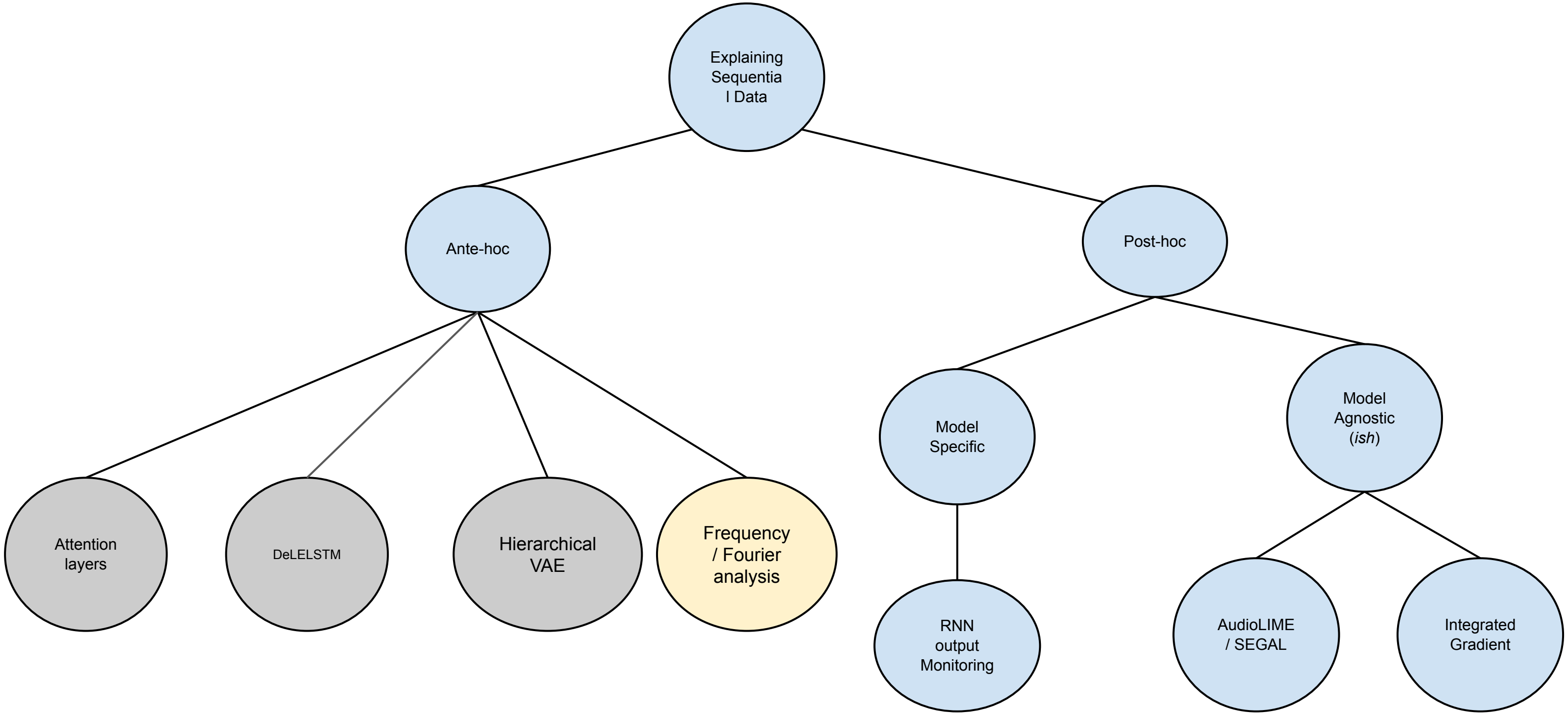
Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

Example reproduced from Rudin, C. (2019). **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**. *Nature machine intelligence*, 1(5), 206-215

(a) Husky classified as wolf (b) Explanation

Example reproduced from Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). **"Why should I trust you?" Explaining the predictions of any classifier**. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Interpretability of Sequential Data



Frequency Analysis

Over a certain length, sequences tend to exhibit a **trend** and a **seasonality**.

Trend: Represent a persistent, long-term change in the mean of the series.

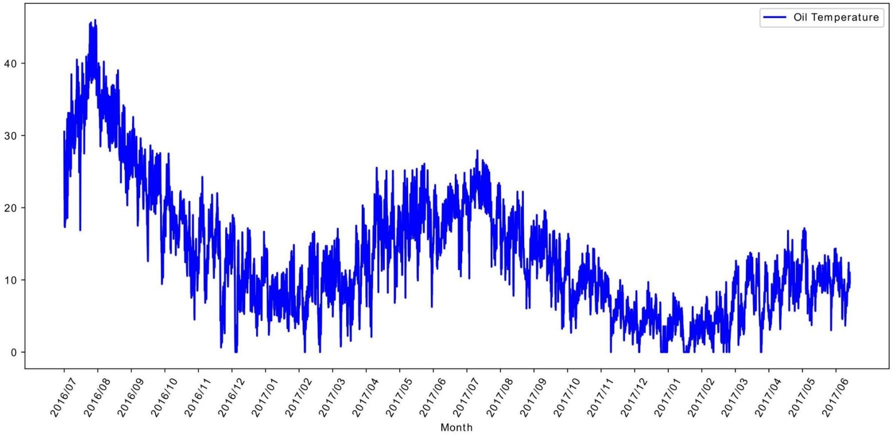
Seasonality: Represent the presence of a regular, periodic change in the mean of the serie.

Frequency Analysis

Over a certain length, sequences tend to exhibit a **trend** and a **seasonality**.

Trend: Represent a persistent, long-term change in the mean of the series.

Seasonality: Represent the presence of a regular, periodic change in the mean of the serie.

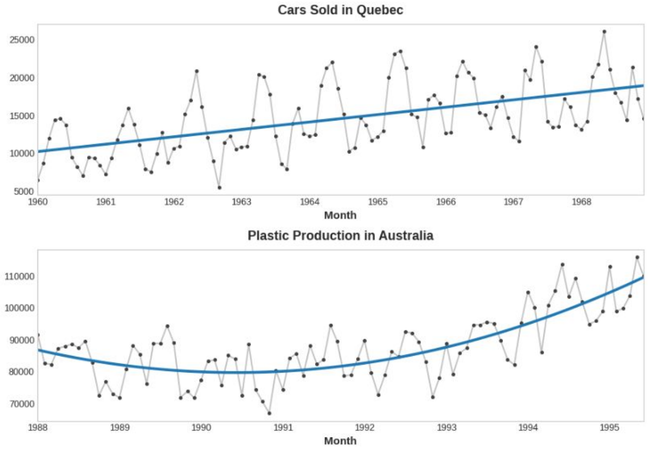


Sample of the ETTh1 dataset

Frequency Analysis

If the sequence is continuous.

- The **trends** of a sequence can be characterized by a linear or a polynomial regression.

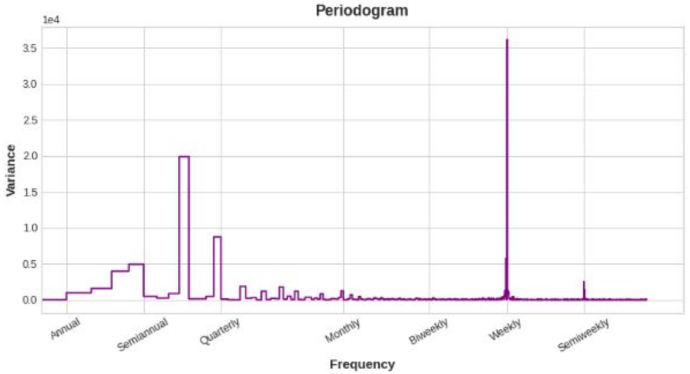


Example reproduced from https://www.kaggle.com/code/ryanholbrook/trend

Frequency Analysis

If the sequence is continuous.

- The **trends** of a sequence can be characterized by a linear or a polynomial regression.
- **Seasons** could be decomposed with a fourier transforms.



Example reproduced from https://www.kaggle.com/code/ryanholbrook/seasonality

Frequency Analysis

If the sequence is continuous.

- The **trends** of a sequence can be characterized by a linear or a polynomial regression.
- **Seasons** could be decomposed with a fourier transforms.

For **sequence prediction**, those parameters can be leveraged to predict the next points in the sequence

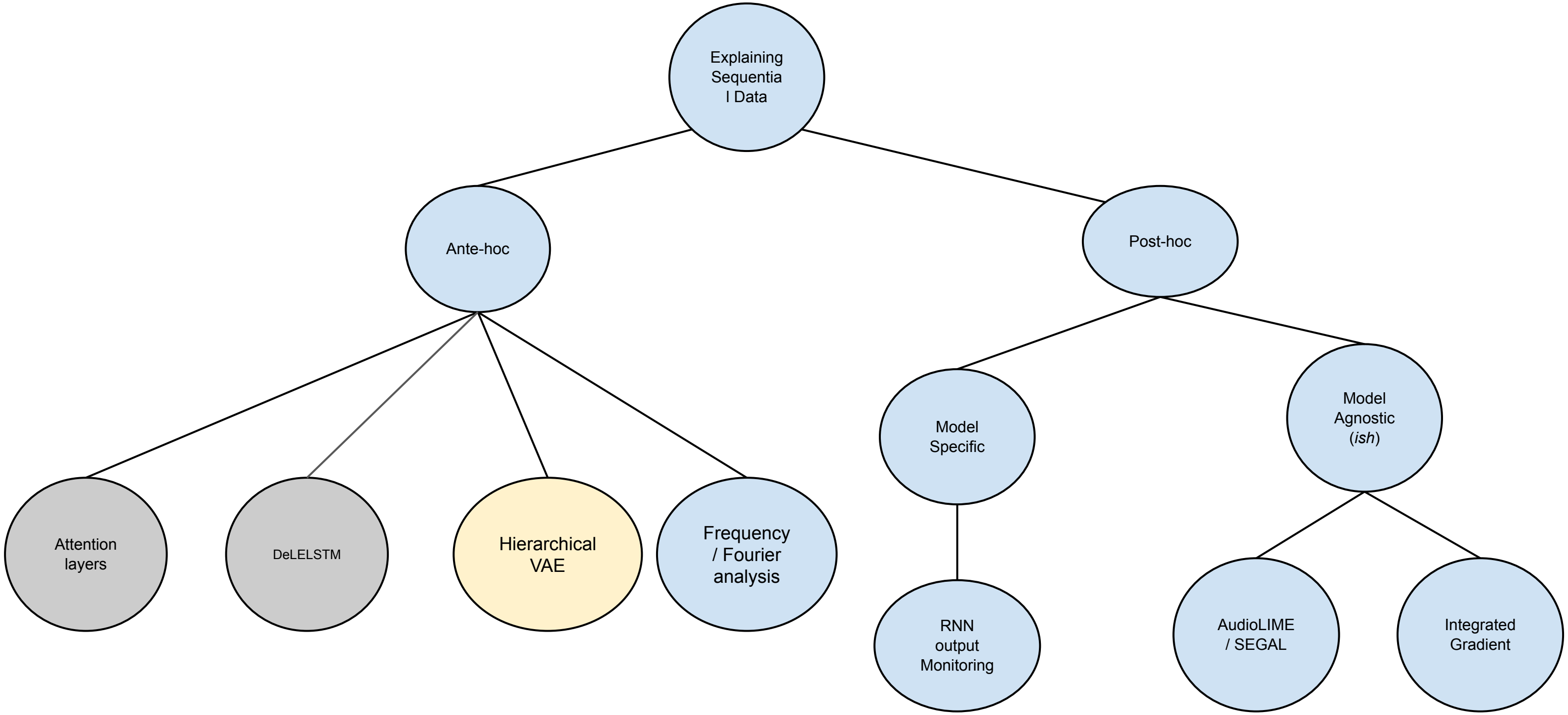
For **classification** or **anomaly detection** this allows the detection of fluctuations in the sequence.

Frequency Analysis

- Trends and Seasons are easy to interpret on dataset with a reasonable amount of features.
- It provides a global explanation of the dynamic of the process generating the sequence

- Assumes that there is a long lasting trend in the data
- Requires knowledge to select the right seasonal features

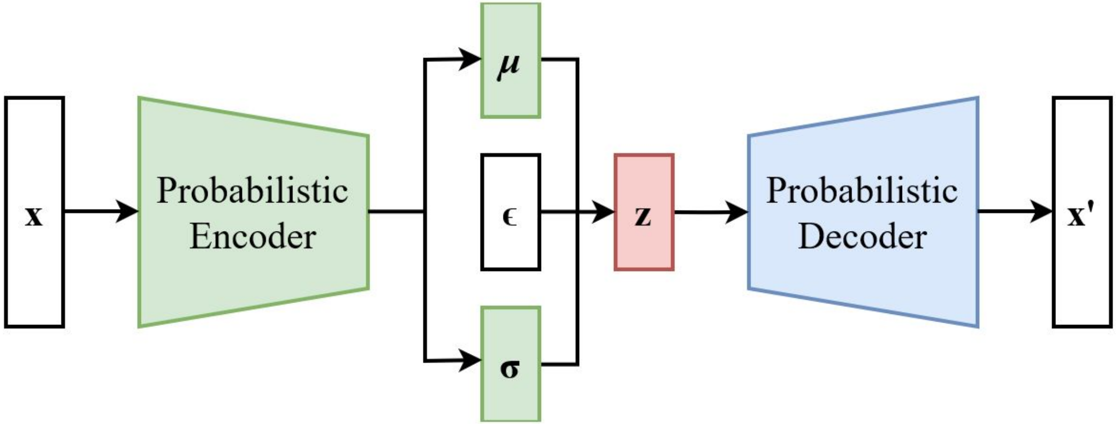
Interpretability of Sequential Data



Hierarchical VAE

Variational Auto Encoder (VAE) is an encoder/decoder architecture generating, from an input, a latent variable *z* following a given distribution.

This *forces* the network to only encode the most relevant information for reconstruction the data.



Hierarchical VAE

In sequence database, there are properties shared by all the sequence of the dataset and properties specifics for each instances.

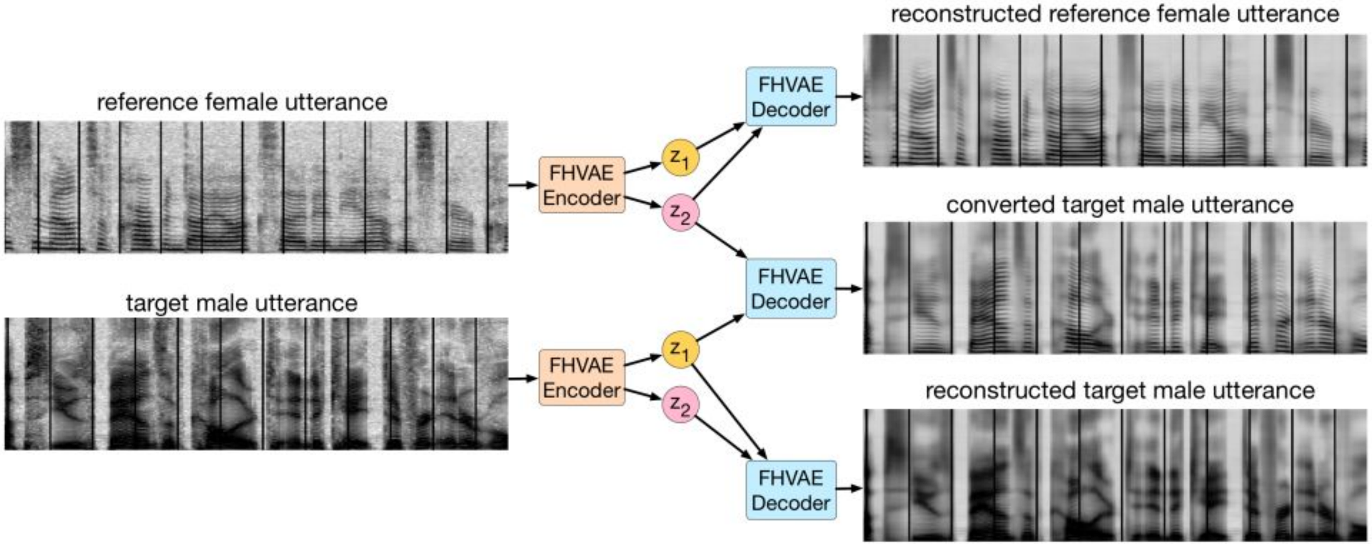
Example: english voice dataset

- The phonetic content of speech is global to all the database.
- Pitch and volume is specific to a subset of frequencies.

Hierarchical VAE allows to compute a separate latent vector for those two aspects. Hsu et al. (2017) developed a training method to make the latent space interpretable.

Hsu, W. N., Zhang, Y., & Glass, J. (2017). **Unsupervised learning of disentangled and interpretable representations from sequential data**. *Advances in neural information processing systems*, 30, *(pp. 1876-1887)*.

Hierarchical VAE



Hsu, W. N., Zhang, Y., & Glass, J. (2017). **Unsupervised learning of disentangled and interpretable representations from sequential data**. *Advances in neural information processing systems*, *30, (pp. 1876-1887)*.

Hierarchical VAE

How to make the latent space interpretable?

- Force the latent vector to follow a prior distribution. This allows to sample in that distribution.
- Add a discriminative objective to the loss function encouraging the model to clearly separate the sequence level attributes and the segment level attributes.

Hierarchical VAE

How to make the latent space interpretable?

- Force the latent vector to follow a prior distribution. This allows to sample in that distribution.
- Add a discriminative objective to the loss function encouraging the model to clearly separate the sequence level attributes and the segment level attributes.

Apart from these two mechanisms, nothing constraint the latent space.
Thus, there is no real guarantee of interpretability.

Empirically, latent variables seems to encode useful representation

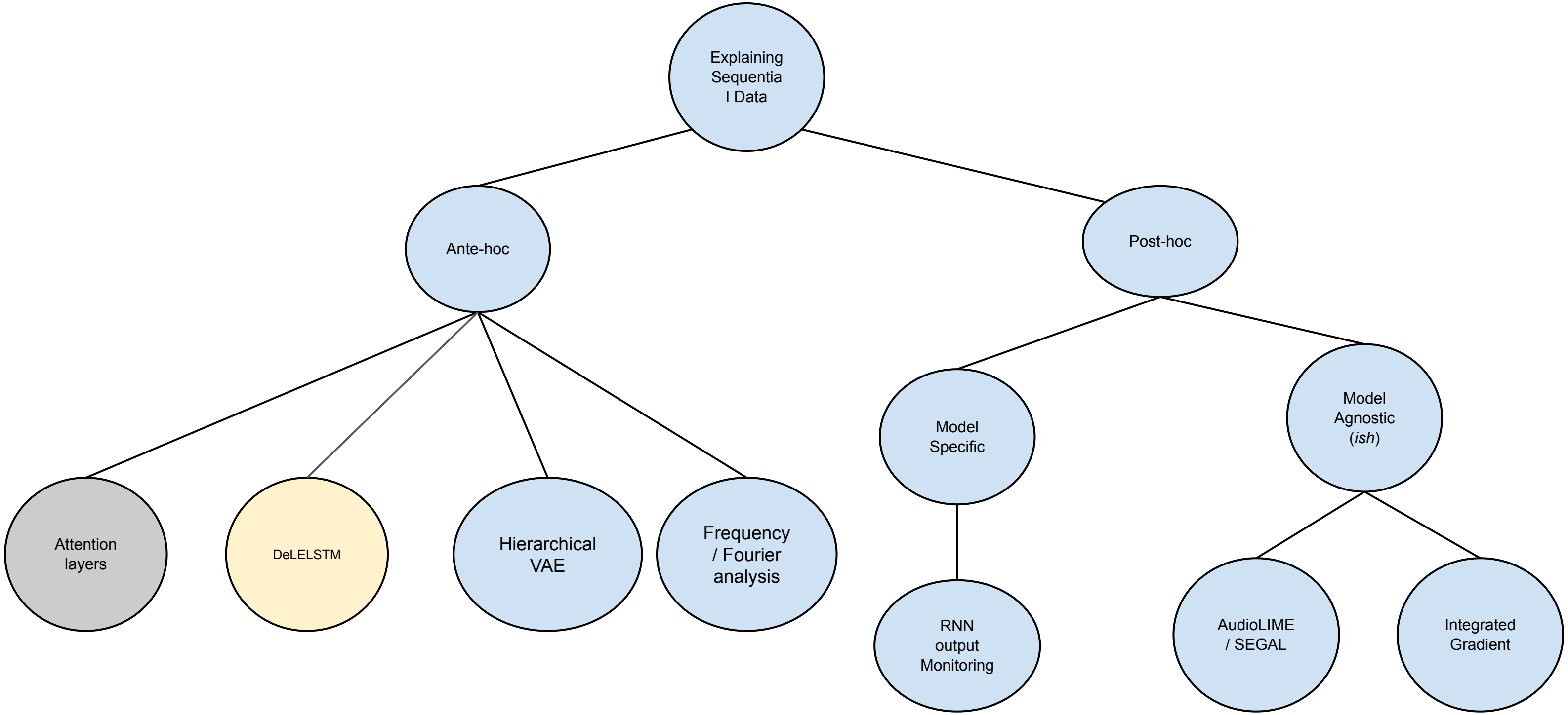
Hierarchical VAE

To evaluate the quality of the learnt latent space, they train a speaker verification model from the Z_1 and Z_2 latent space.

Better performances are reach using the Z_2 vector as it was train to encode speaker specific features.

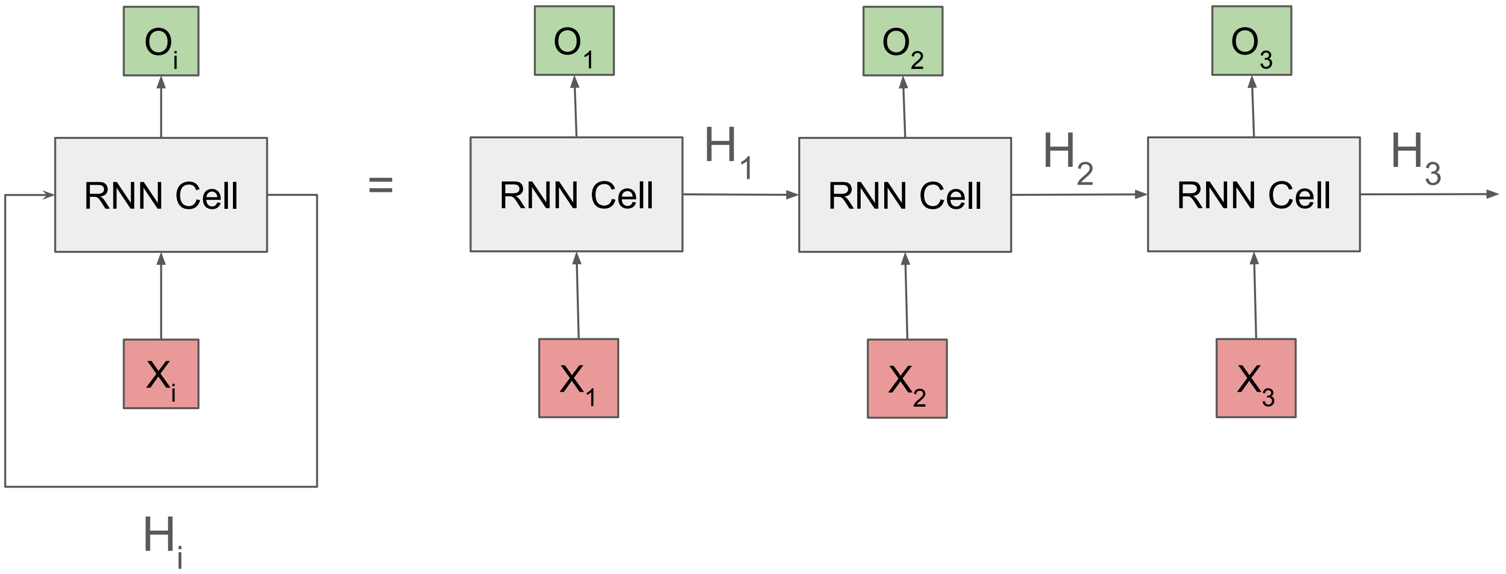
Does it make it interpretable? How to convey that information to non expert users?

Interpretability of Sequential Data



DeLELSTM

Hidden State: encode long term contextual information about the sequence



DeLELSTM Model

What if it was possible to train more interpretable Hidden State?

Decomposition-base Linear Explainable LSTM is an architecture that forces the hidden state to contain a linear combination of the past information.

Wang, C., Li, Y., Sun, X., Wu, Q., Wang, D., & Huang, Z. (2023). **DeLELSTM: decomposition-based linear explainable LSTM to Capture Instantaneous and Long-term Effects in Time Series**. *arXiv:2308.13797*.

DeLELSTM

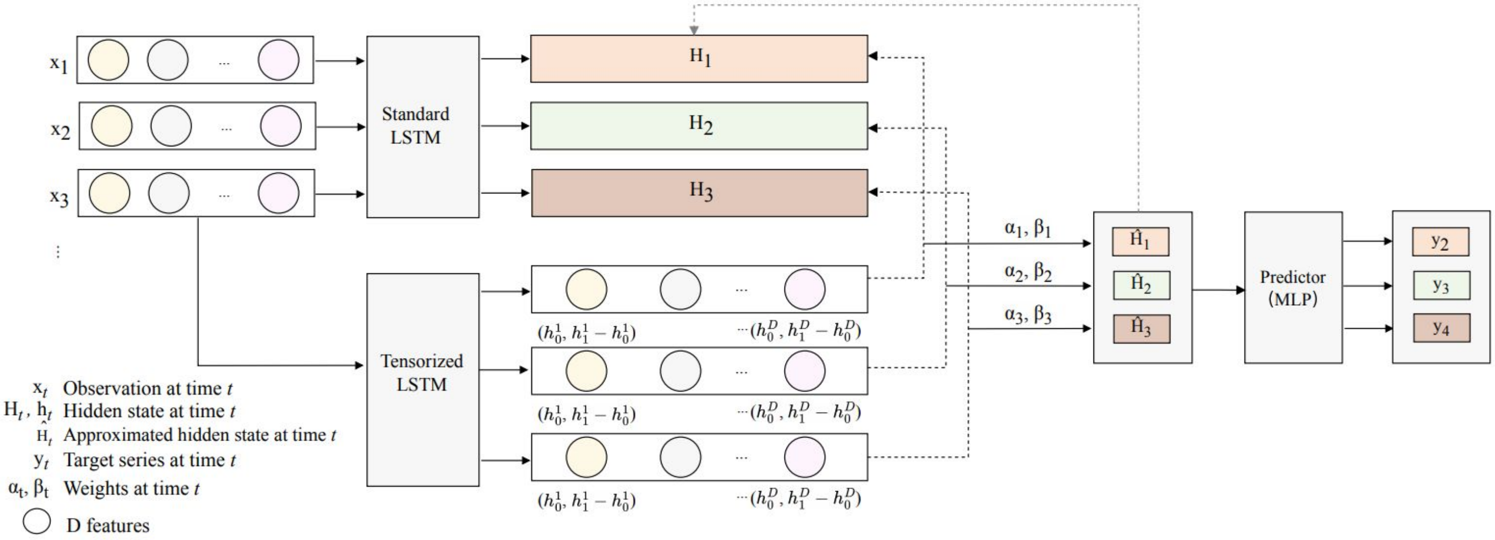


Figure reproduced from Wang, C., Li, Y., Sun, X., Wu, Q., Wang, D., & Huang, Z. (2023). **DeLELSTM: decomposition-based linear explainable LSTM to Capture Instantaneous and Long-term Effects in Time Series**. *arXiv:2308.13797*.

DeLELSTM

The hidden state computed by the Tensorized LSTM allows to build this kind of visualisation. This example investigate a prediction task on electricity consumption

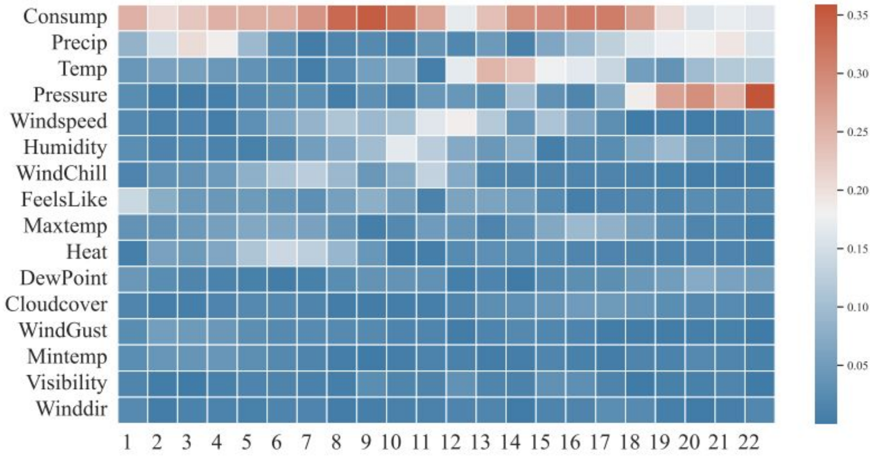


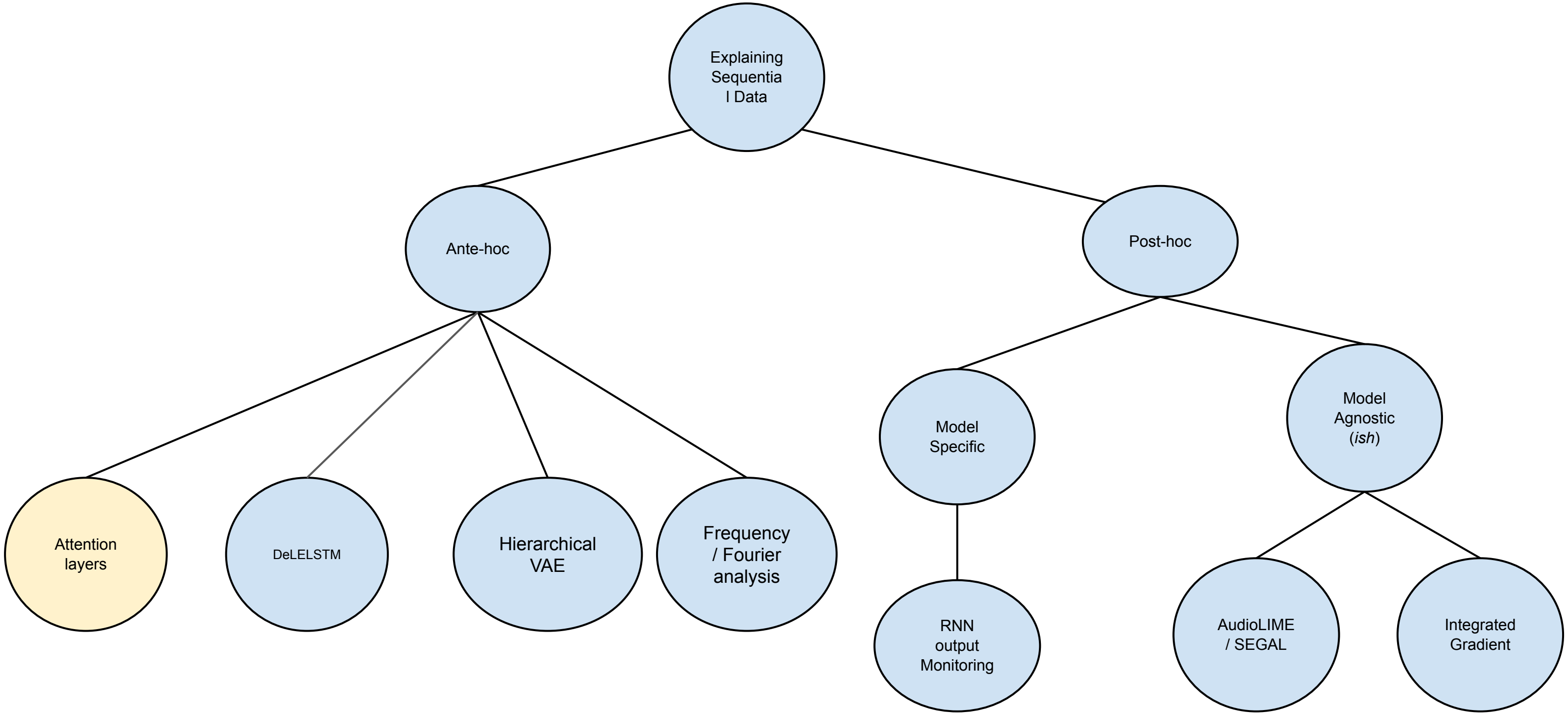
Figure reproduced from Wang, C., Li, Y., Sun, X., Wu, Q., Wang, D., & Huang, Z. (2023). **DeLELSTM: decomposition-based linear explainable LSTM to Capture Instantaneous and Long-term Effects in Time Series**. *arXiv:2308.13797*.

DeLELSTM

It is easy to compute the importance of each feature for a particular time step.

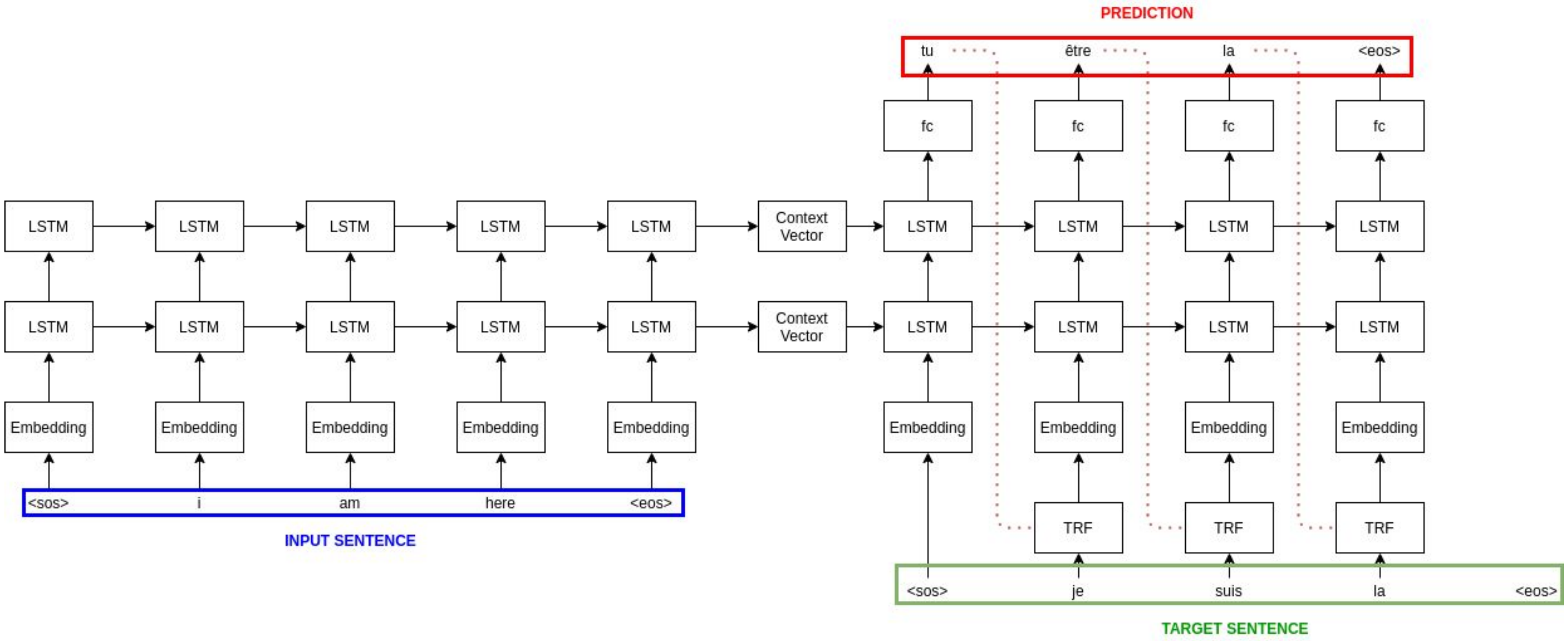
However the hidden state is bigger and harder to compute.

Interpretability of Sequential Data



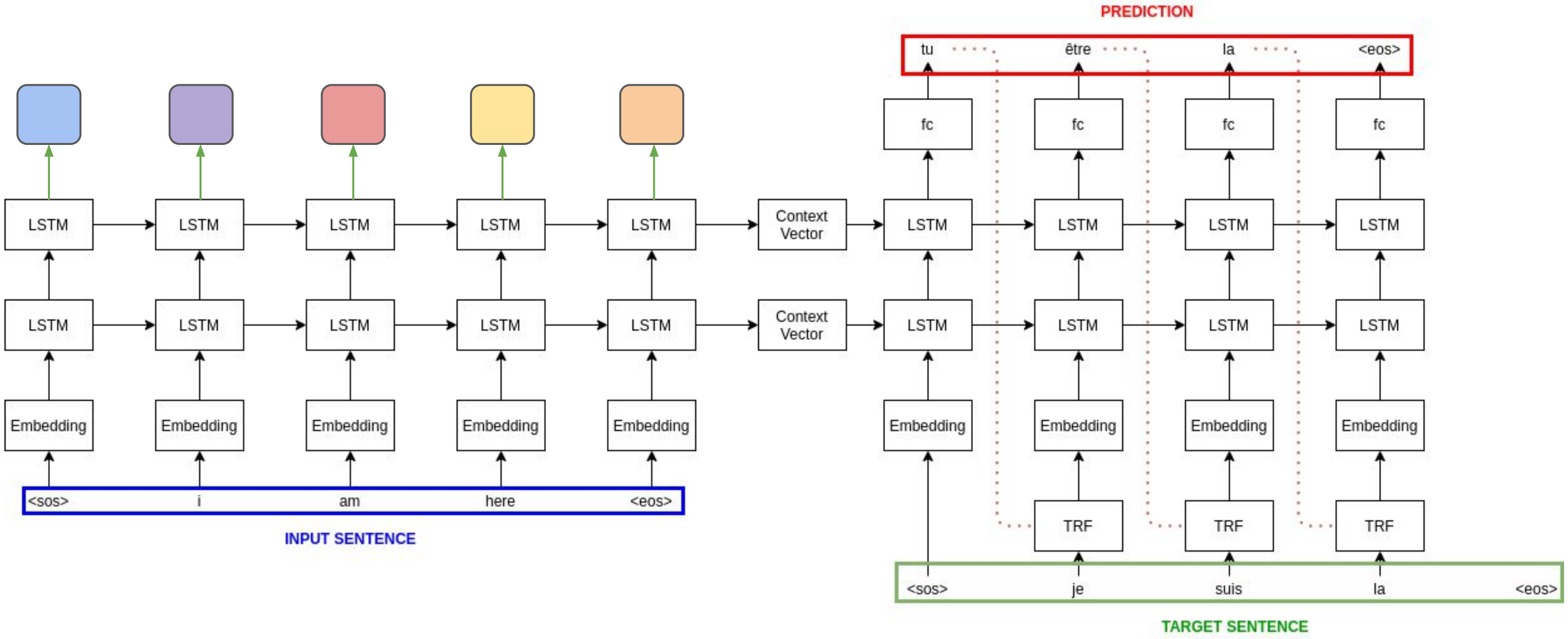
Attention Layers

Add an learnable *output monitoring* module in the RNN architecture. Primarily developed for seq2seq models.



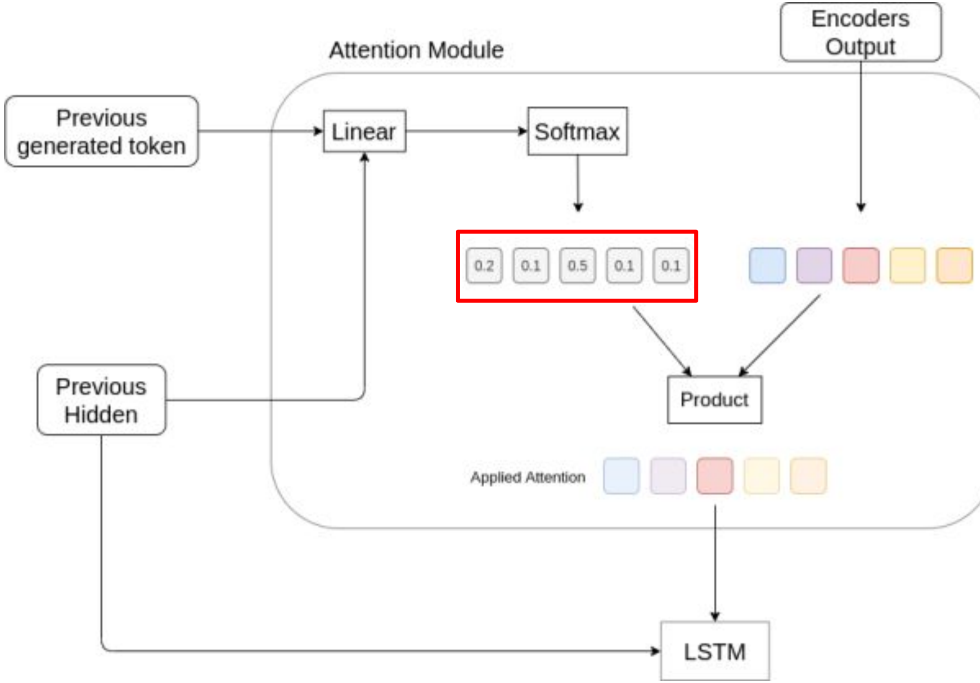
Attention Layers

Add an learnable *output monitoring* module in the RNN architecture. Primarily developed for seq2seq models.



Attention Layers

Add an learnable *output monitoring* module in the RNN architecture. Primarily developed for seq2seq models.



Attention Layers

Experiments showed that the attention mechanism is enough to create an efficient seq2seq model. Those findings started the transformers trends.

Are attention based models interpretable?

Serrano, S., & Smith, N. A. (2019). **Is Attention Interpretable?** In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 2931-2951).

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). **Is attention explanation? An introduction to the debate**. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 3889-3900).

Attention Layers

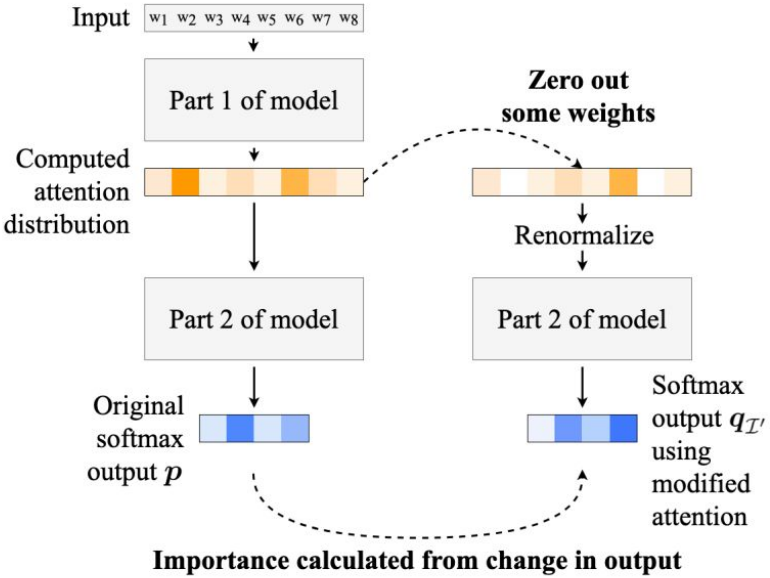
Are attention based models interpretable?

Most effective way to flip a model decision

1) Set to zero the attention weight receiving the larger gradient
2) Set to zero the largest attention weight
3) Set to zero random attention weight

Attention Layers

Are attention based models interpretable?



Serrano, S., & Smith, N. A. (2019). **Is Attention Interpretable?** In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 2931-2951).

Attention Layers

Are attention based models interpretable?

Most effective way to flip a model decision

1) Set to zero the attention weight receiving the larger gradient
2) Set to zero the largest attention weight
3) Set to zero random attention weight

"Attention weights are only a noisy predictors of component's importance"

Take home messages

- Before starting your research, you need to clarify which aspect of the field is valuable for you. HCI? Compliance? Create new methods / architecture?
- Generic methods (e.g., LIME) work well. But adapting them to better handle the type of data you manipulate could be valuable for end users.
- Explaining black boxes will always hide part of the story.

We just scratch the surface of the literature on the subject

Do you have any questions or comments?

AIMLAI Tutorial

Part 2: Explainability of LLMs

Adrien Bibal

InferLink

What this part of the tutorial will **not** cover

What this part of the tutorial will **not** cover

- Solutions only proposed for smaller language models (e.g., BERT)
  - Examples:
    - Analyzing attention heads in small(er) models
    - How to map concepts to neurons in small(er) models
    - Etc.

What this part of the tutorial will **not** cover

- Solutions only proposed for smaller language models (e.g., BERT)
  - Examples:
    - Analyzing attention heads in small(er) models
    - How to map concepts to neurons in small(er) models
    - Etc.
  - Some exceptions, like with GPT-2

What this part of the tutorial will **not** cover

- Solutions only proposed for smaller language models (e.g., BERT)
  - Examples:
    - Analyzing attention heads in small(er) models
    - How to map concepts to neurons in small(er) models
    - Etc.
  - Some exceptions, like with GPT-2
- Solutions where LLMs are used as an explainer, but not to explain LLMs

What this part of the tutorial will **not** cover

- Solutions only proposed for smaller language models (e.g., BERT)
    - Examples:
        - Analyzing attention heads in small(er) models
        - How to map concepts to neurons in small(er) models
        - Etc.
    - Some exceptions, like with GPT-2
- Solutions where LLMs are used as an explainer, but not to explain LLMs
    - Tutorial focused on how to explain LLMs
    - Not how to use LLMs to explain
    - Except when LLMs are used to explain LLMs

Additional introductory notes

- Speaking of LLMs used to get explanations

Additional introductory notes

- Speaking of LLMs used to get explanations

  LLMs can provide verbal explanations (also called rationalizations), but not always trustworthy because of hallucinations

  Turpin, M., Michael, J., Perez, E., & Bowman, S. (2024). **Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.** *In Proceedings of NeurIPS*, *36*, 74952-74965.

Additional introductory notes

- Speaking of LLMs used to get explanations

  LLMs can provide verbal explanations (also called rationalizations), but not always trustworthy because of hallucinations

  Turpin, M., Michael, J., Perez, E., & Bowman, S. (2024). **Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.** *In Proceedings of NeurIPS*, *36*, 74952-74965.

- Previously, (local) explainability was mostly about "why output given input"?

Additional introductory notes

- Speaking of LLMs used to get explanations

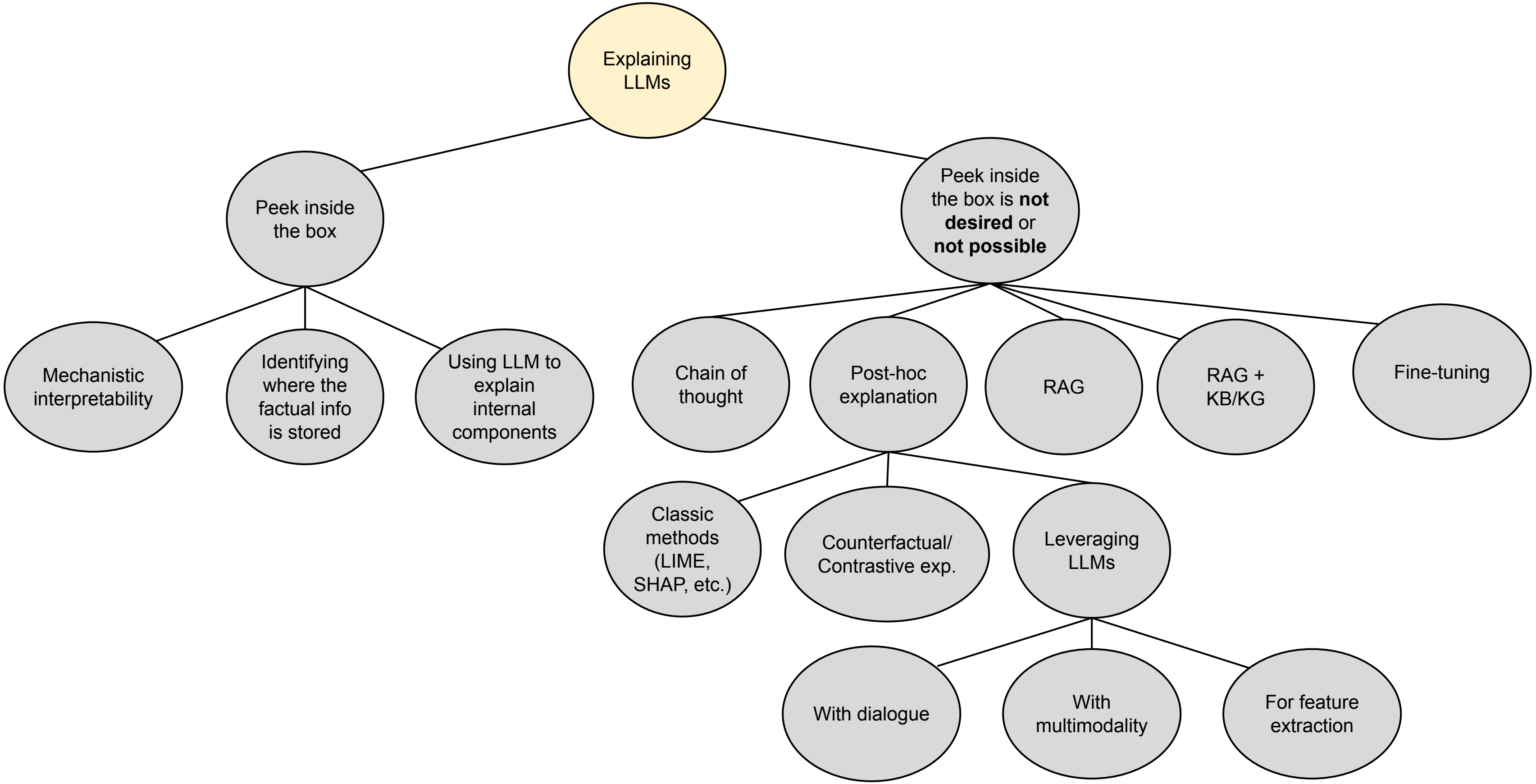
  LLMs can provide verbal explanations (also called rationalizations), but not always trustworthy because of hallucinations

  Turpin, M., Michael, J., Perez, E., & Bowman, S. (2024). **Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.** *In Proceedings of NeurIPS*, *36*, 74952-74965.

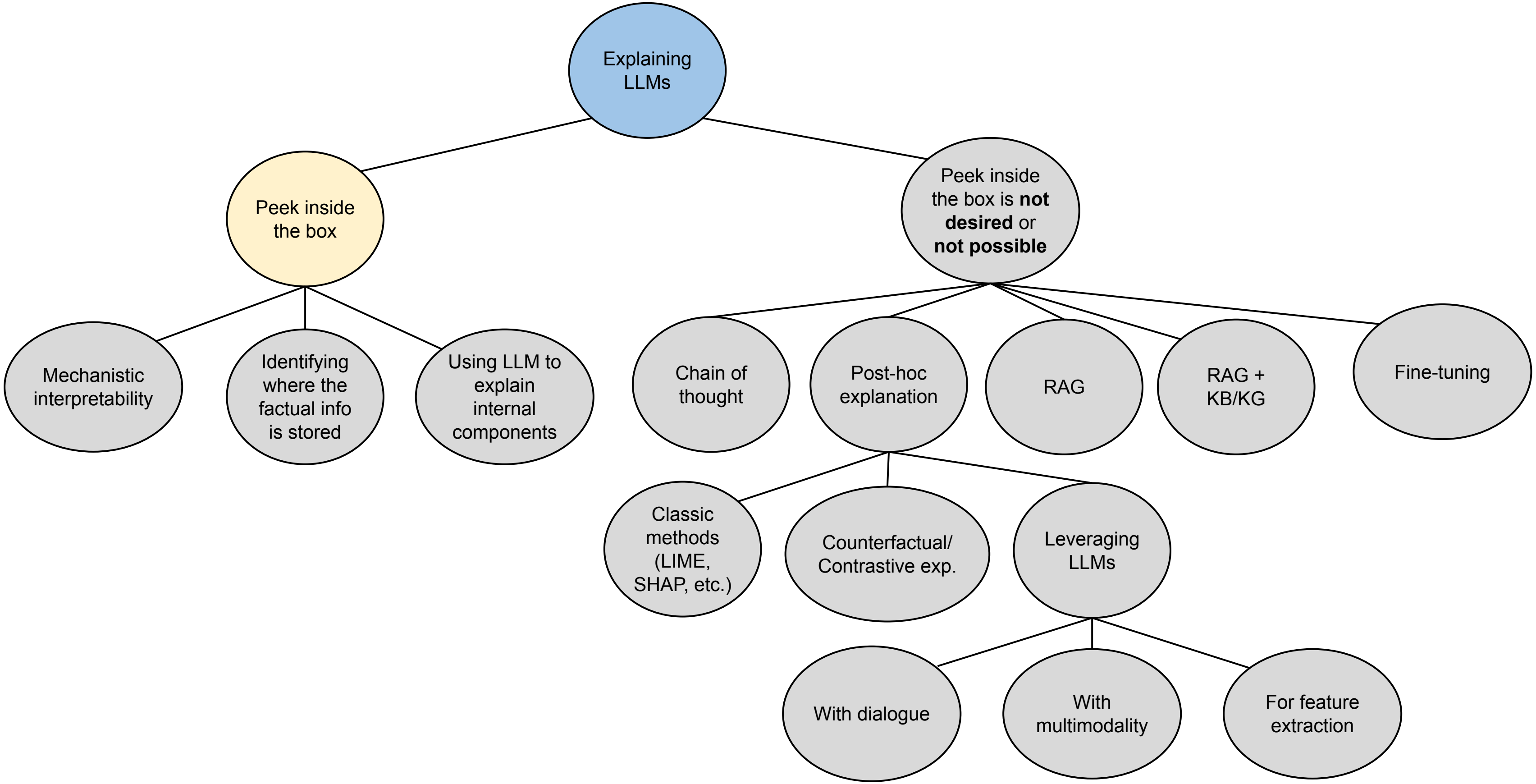
- Previously, (local) explainability was mostly about "why output given input"?

  But in the LLM era, the training data is as important as the input

End-goal of this part of the tutorial



Let's start with situations where peek inside the box is desirable



What if the LLM black box can be opened?

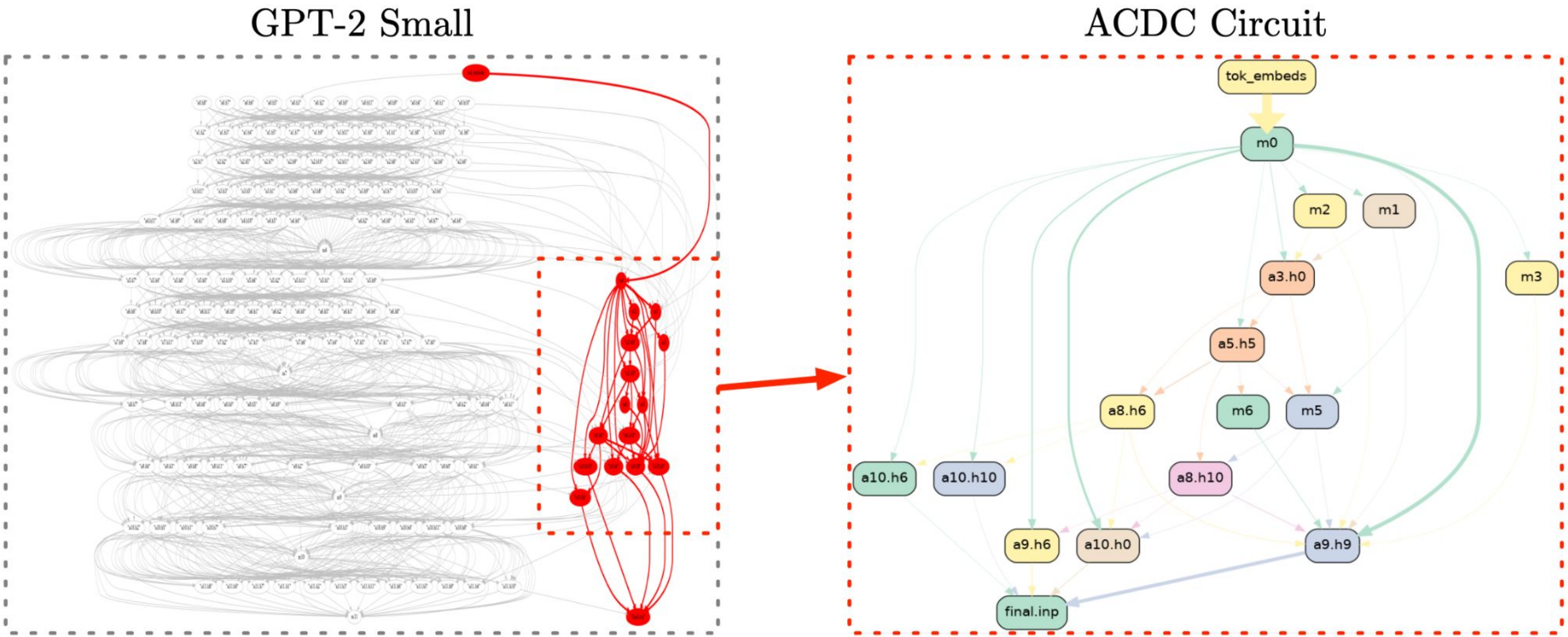
3 main options:

- Mechanistic interpretability
- Studying stored information in the LLM's neurons
- Using an LLM to explain the internal components of another LLM

White box – Mechanistic interpretability

- Global explanations
- Focus on understanding the internal mechanisms used by LLMs
- Generally studied for smaller models (e.g., GPT-2),
  with the hope that the conclusions also hold for larger models
- Examples of that:
  - Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... & Olah, C. (2022). **In-context learning and induction heads**. *arXiv:2209.11895*.
  - Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). **Towards automated circuit discovery for mechanistic interpretability**. In *Proceedings of NeurIPS*, *36*, 16318-16352.
  - Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2023). **Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small.** In *Proceedings of ICLR*.
  - Wu, Z., Geiger, A., Icard, T., Potts, C., & Goodman, N. (2024). **Interpretability at scale: Identifying causal mechanisms in alpaca**. In *Proceedings of NeurIPS*, *36, 78205-78226*.

White box – Mechanistic interpretability – Example



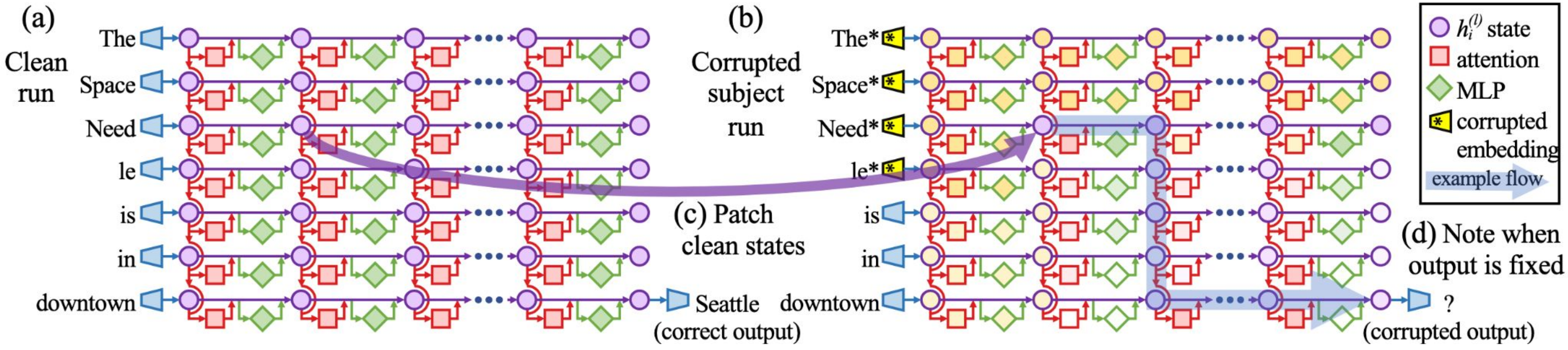
Using activation patching to discover circuits for specific behaviors

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *In Proceedings of NeurIPS*, *36*, 16318-16352.

White box – Studying stored information

- Global explanations
- Focus on *what* is stored, *how* it is stored, and *how* it is retrieved
- Again, smaller models (e.g., GPT-2) are generally studied, with the hope that the conclusions also hold for larger models
- Examples:
    - Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). **Locating and editing factual associations in GPT.** In *Proceedings of NeurIPS*, *35*, 17359-17372.
    - Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). **Dissecting recall of factual associations in auto-regressive language models.** In *Proceedings of EMNLP* (pp. 12216-12235).
    - Katz, S., & Belinkov, Y. (2023, December). **VISIT: Visualizing and interpreting the semantic information flow of transformers.** In *Findings of EMNLP* (pp. 14094-14113).

White box – Studying stored information – Example



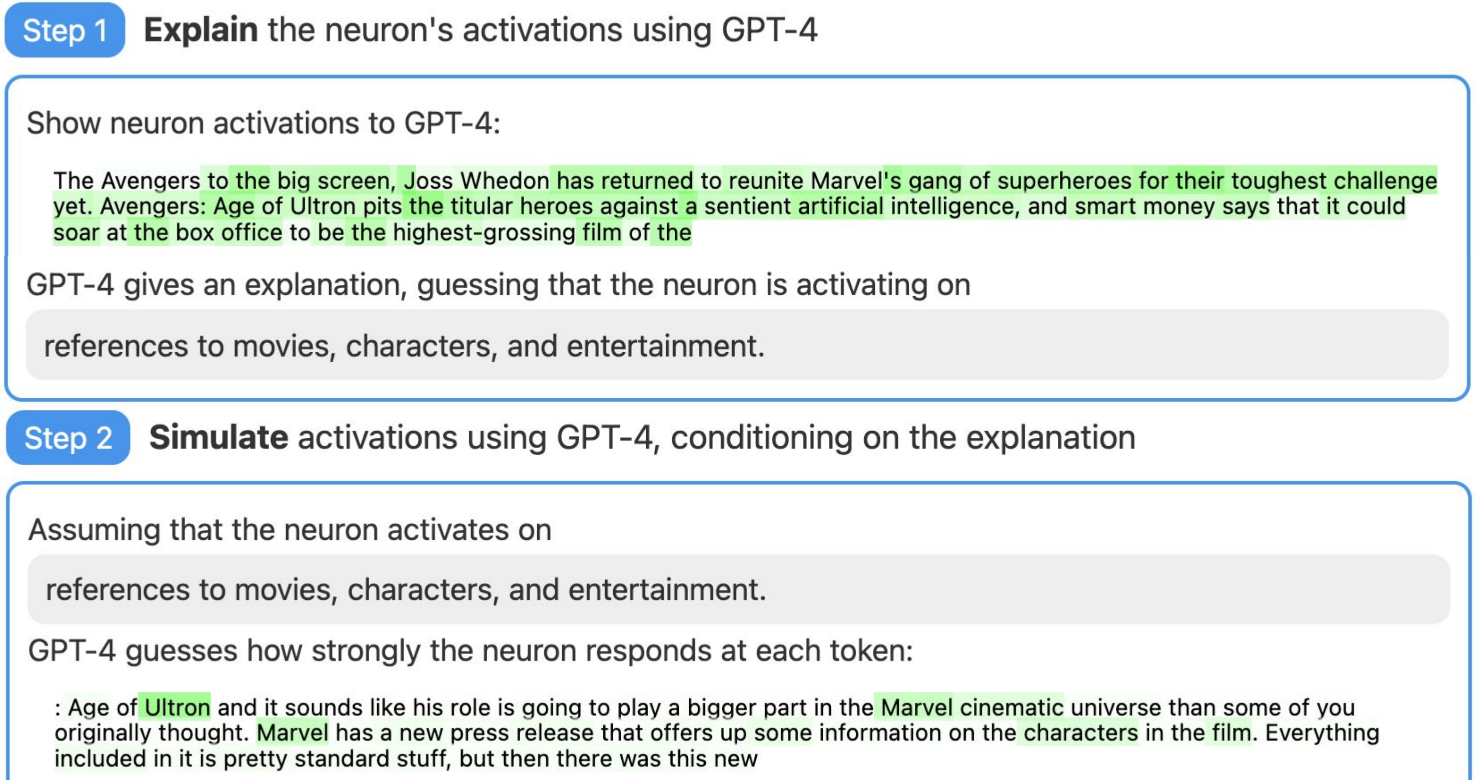
Corrupt input embeddings and and restore some states to study stored information

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*, 35, 17359-17372.

White box – Using an LLM to explain LLM's neurons

- Global explanations
- Use powerful LLMs to understand neurons in a given LLM
- Again and again, generally performed on smaller models (e.g., GPT-2), with the hope that the conclusions also hold for larger models
- Examples:
    - Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., ... & Saunders, W. (2023). **Language models can explain neurons in language models.** *https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html*
    - Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., & Geva, M. (2024). **Patchscopes: A unifying framework for inspecting hidden representations of language models.** In *Proceedings of ICML*.

White box – Using an LLM to explain an LLM's neurons – Example



Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., ... & Saunders, W. (2023). Language models can explain neurons in language models. *https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html*

White box – Using an LLM to explain an LLM's neurons – Example

Step 3 **Score** the explanation by comparing the simulated and real activations

Real activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

Simulated activations:

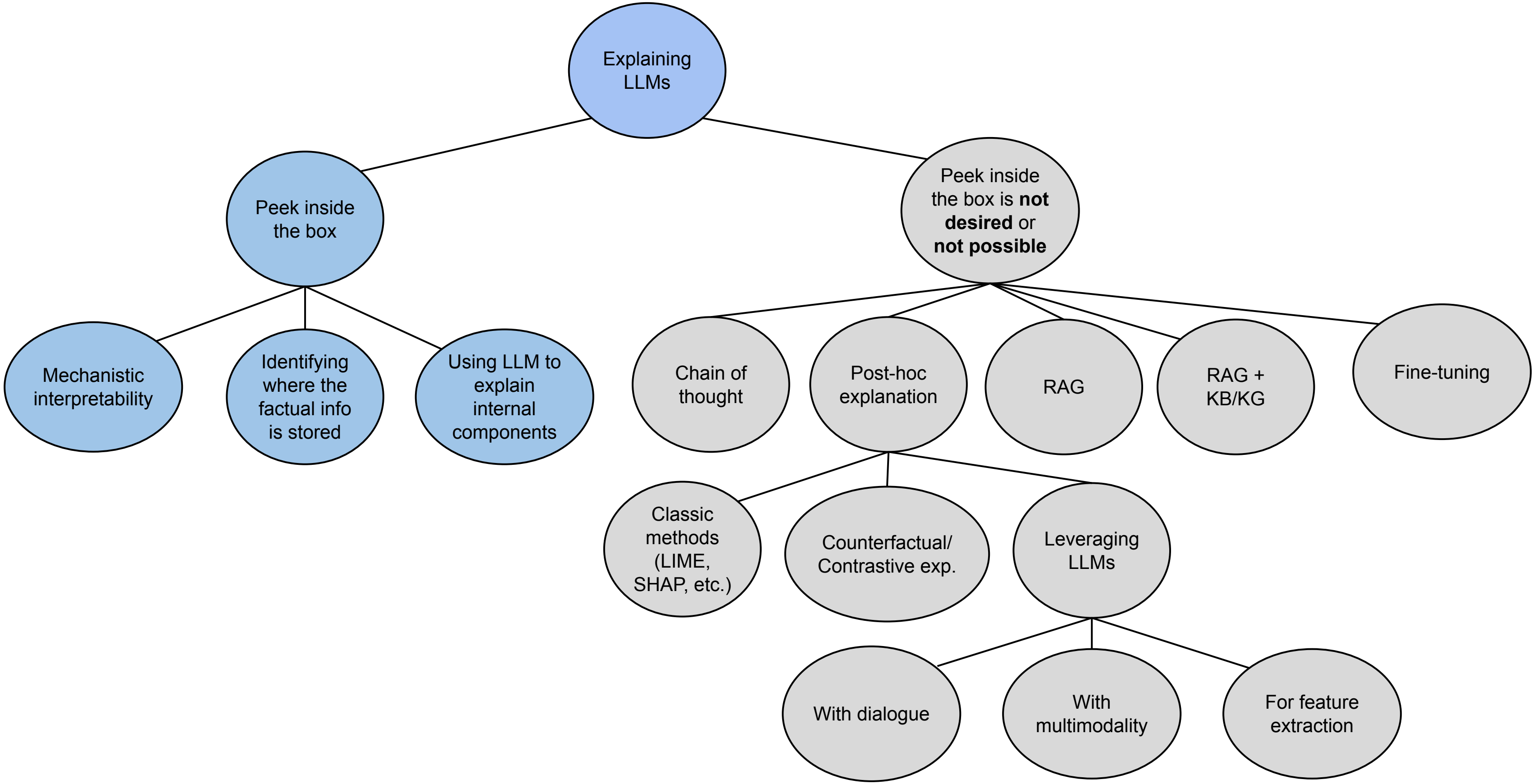
: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

Comparing the simulated and real activations to see how closely they match, we derive a score:

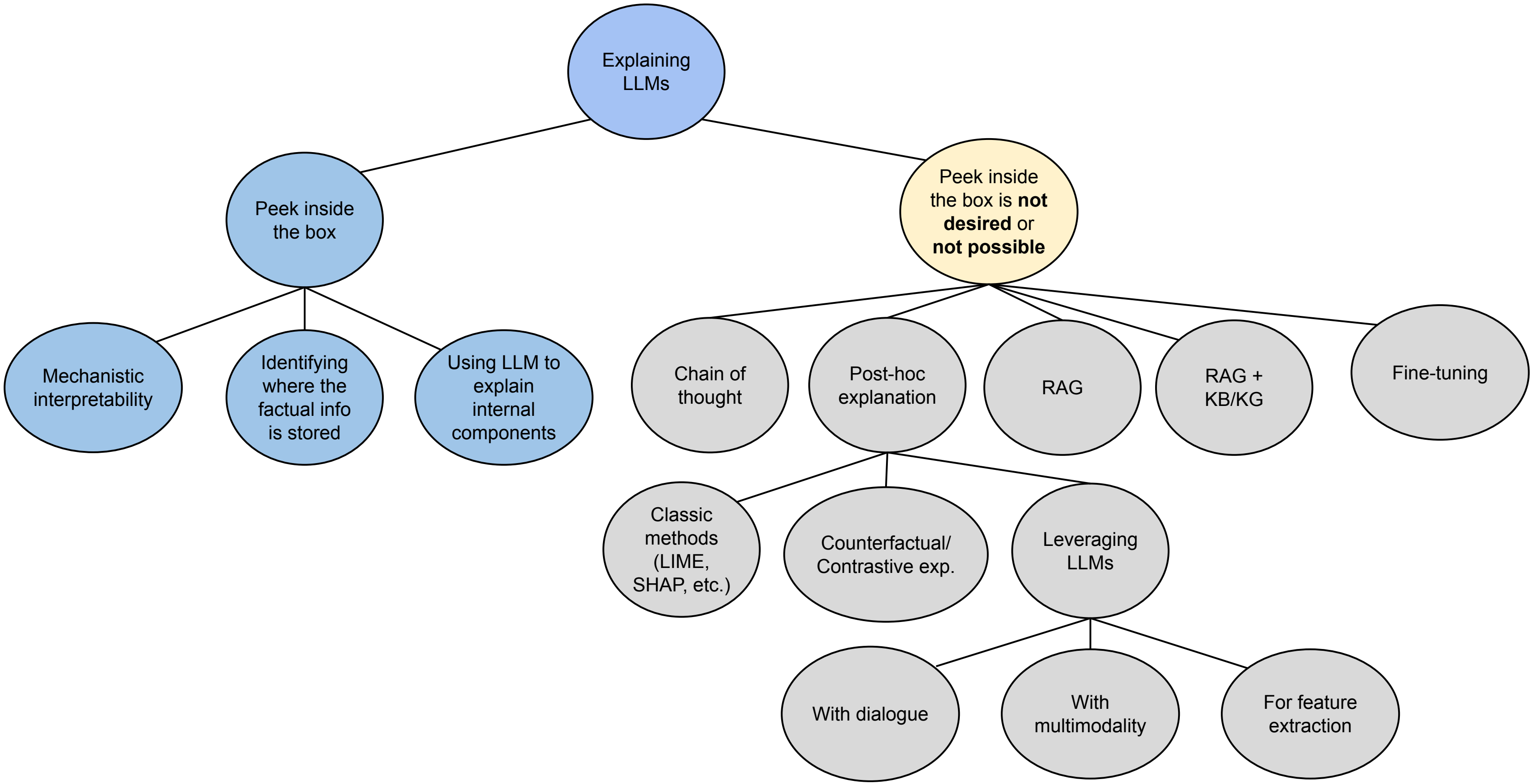
0.337

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., ... & Saunders, W. (2023). Language models can explain neurons in language models.
https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

Anything you would like to add regarding white boxes?



Let's explore black boxes now



Black box – Chain of Thought Prompting

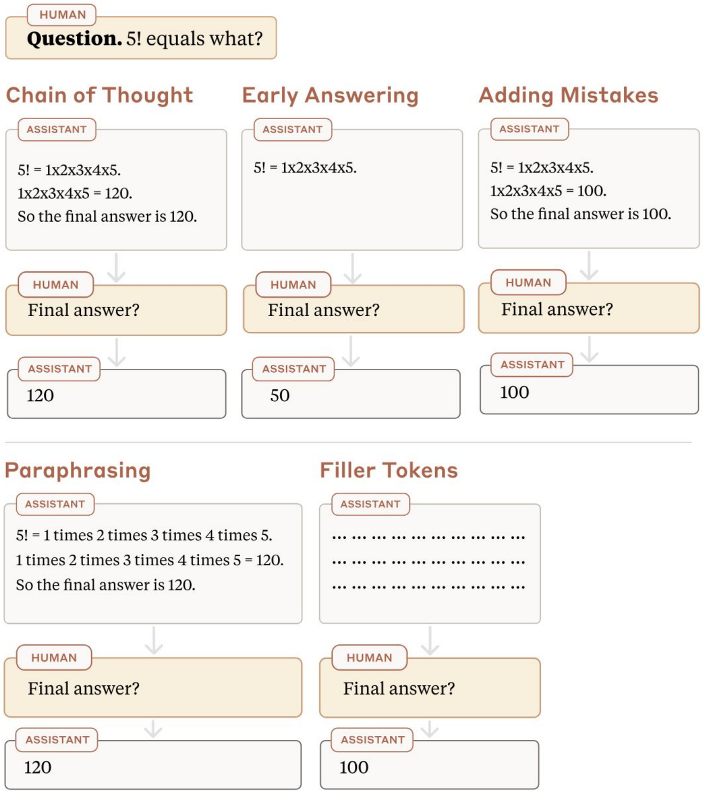
- Local explanations

- Ask the LLM to produce the explanation alongside its final answer

- Mainly used to increase performance,
  but can also be used for explanatory purposes

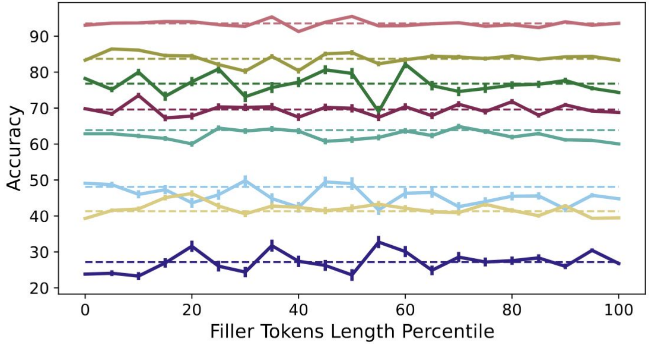
- Examples:
  - Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). **Chain-of-thought prompting elicits reasoning in large language models.** In *Proceedings of NeurIPS*, *35*, 24824-24837.
  - Yoran, O., Wolfson, T., Bogin, B., Katz, U., Deutch, D., & Berant, J. (2023). **Answering questions by meta-reasoning over multiple chains of thought.** In *Proceedings of EMNLP* (pp. 5942-5966).
  - Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., ... & Perez, E. (2023). **Measuring faithfulness in chain-of-thought reasoning.** *arXiv:2307.13702*.

Black box – Chain of Thought Prompting – Example



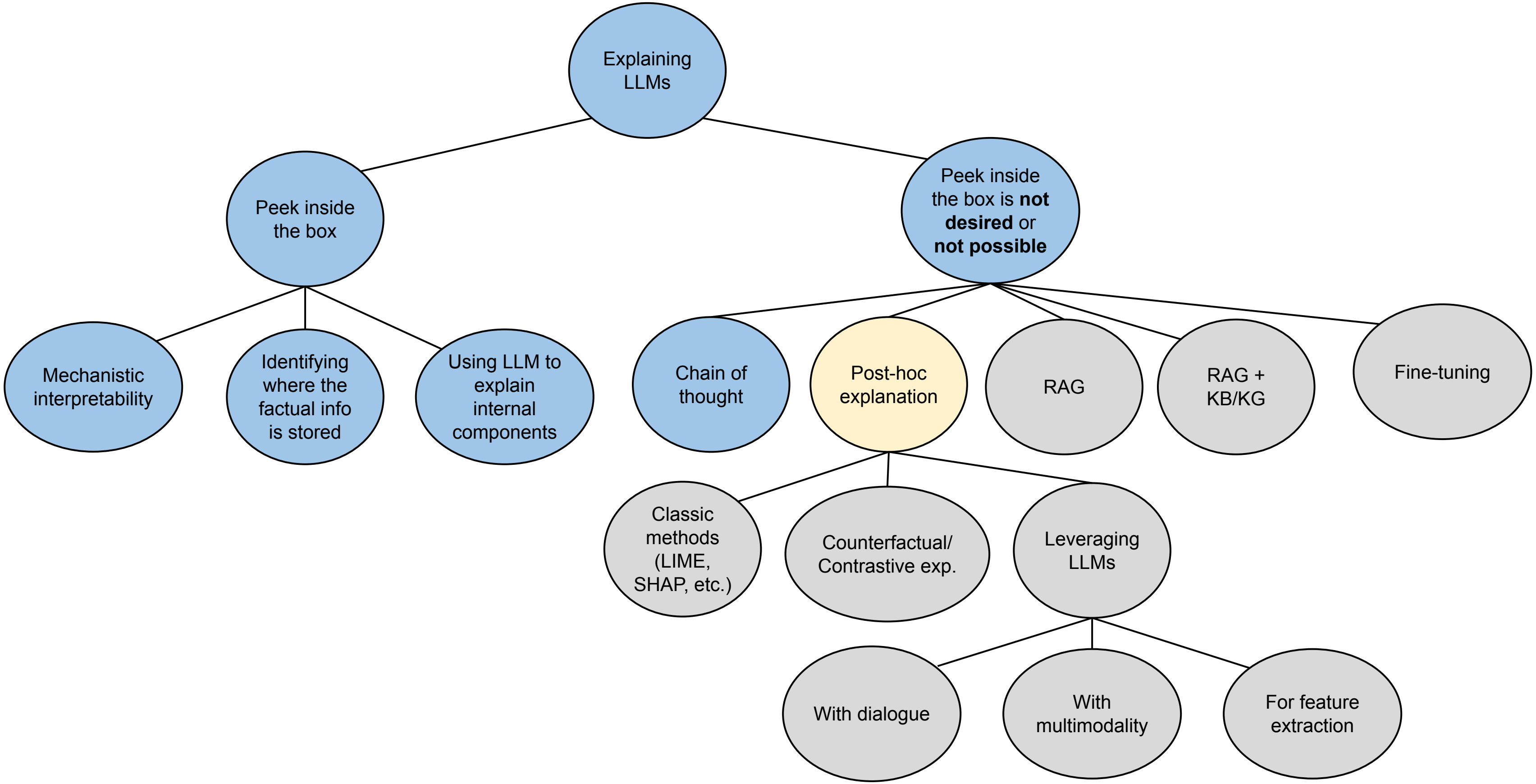
Results:

- Early Answering: Final result depends on the stage of the reasoning process
- Adding Mistakes: Final result relies on the reasoning, even if mistaken
- Paraphrasing: Accuracy is still almost exactly the same
- Filler Tokens: Difficult to know what is happening



Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., ... & Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. *arXiv:2307.13702.*

What about post-hoc explanations?



Post-hoc Explanations in the Era of LLMs

- What post-hoc explanations are?

- Important question because of the "change of paradigm" related to LLMs

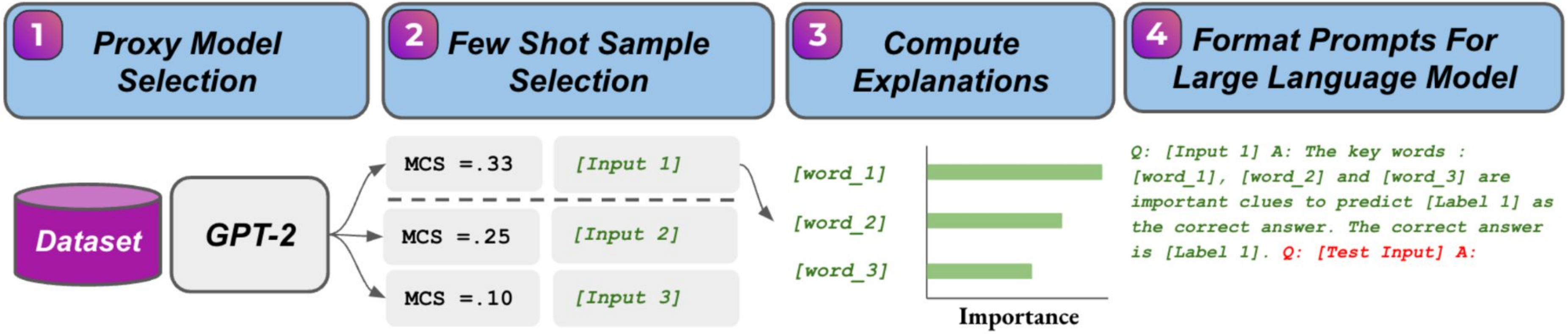
    - Before: training a model = changing its parameters

    - Now: prompt engineering is *kind of* a training process
        - I.e., it's now the prompt that needs to be tuned to get a good classifier

Black box – Post-hoc – Classic methods

- Local explanations
- Use known explanation methods (LIME, SHAP, etc.) in the context of LLMs
- Hypothesis that simple models can estimate complex models locally may not hold
  So need to rethink how to use these classic explanation methods
- Examples:
    - Chen, H., Covert, I. C., Lundberg, S. M., & Lee, S. I. (2023). **Algorithms to estimate Shapley value feature attributions.** *Nature Machine Intelligence*, *5*(6), 590-601.
    - Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., & Lakkaraju, H. (2023). **Post hoc explanations of language models can improve language models.** In *Proceedings of NeurIPS*, *36, 65468-65483*.
    - Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). **Rethinking interpretability in the era of large language models.** *arXiv:2402.01761*.

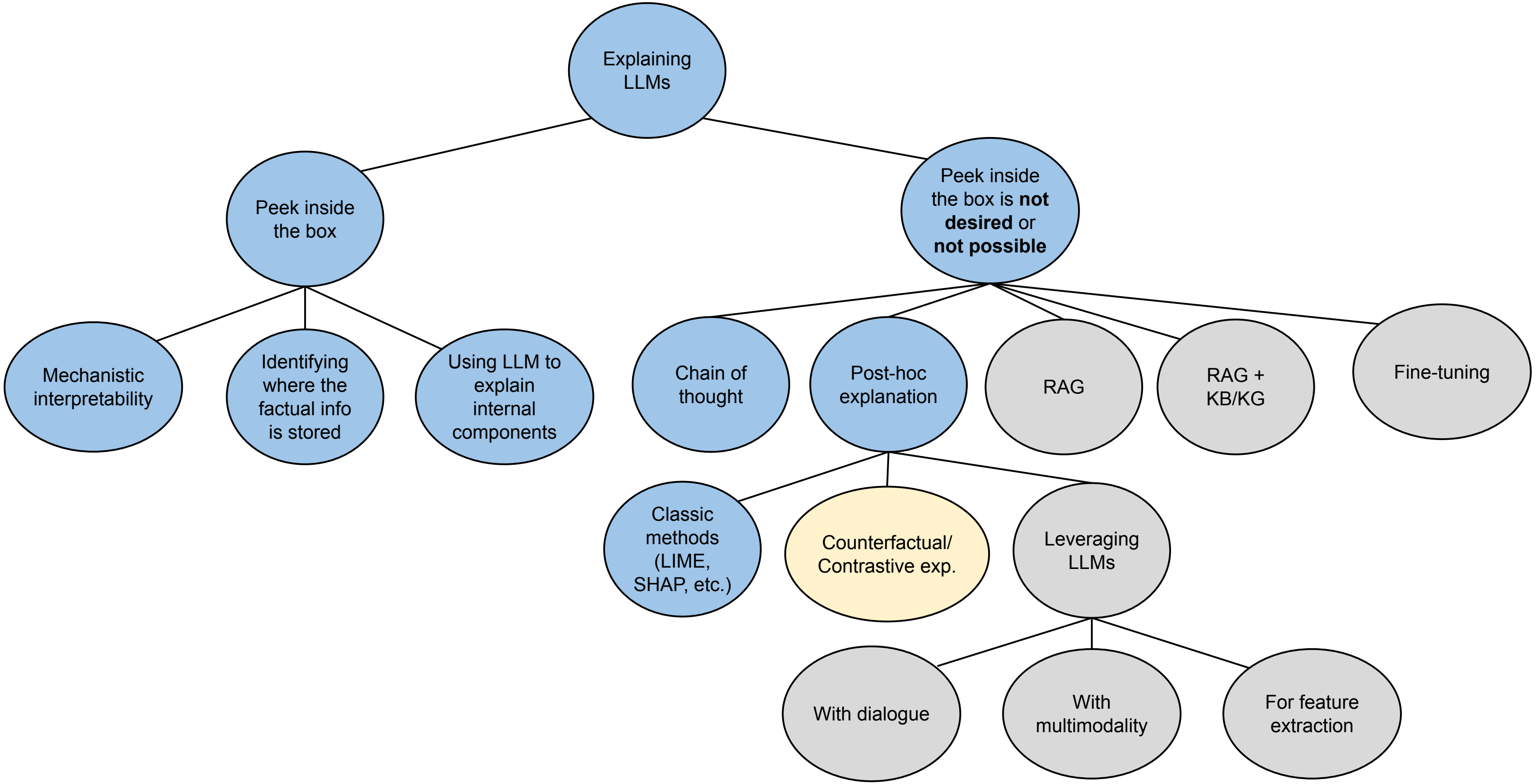
Black box – Post-hoc – Classic methods – Example



Use classic method on GPT-2, then use as the expl. as few-shot in a larger model.

Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., & Lakkaraju, H. (2024). Post hoc explanations of language models can improve language models. In *Proceedings of NeurIPS*, 36, 65468-65483.

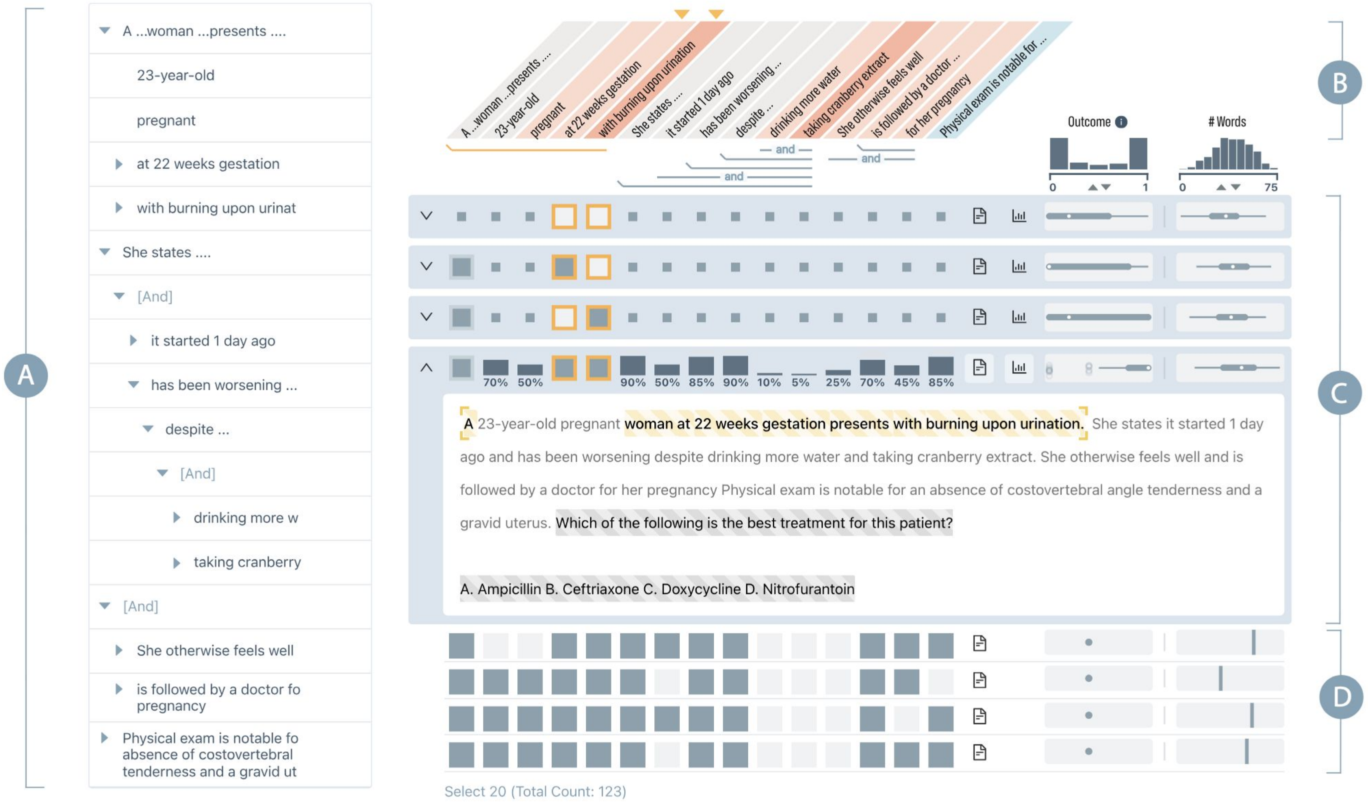
What about counterfactual and contrastive explanations?



Black box – Post-hoc – Counterfactual/contrastive exp.

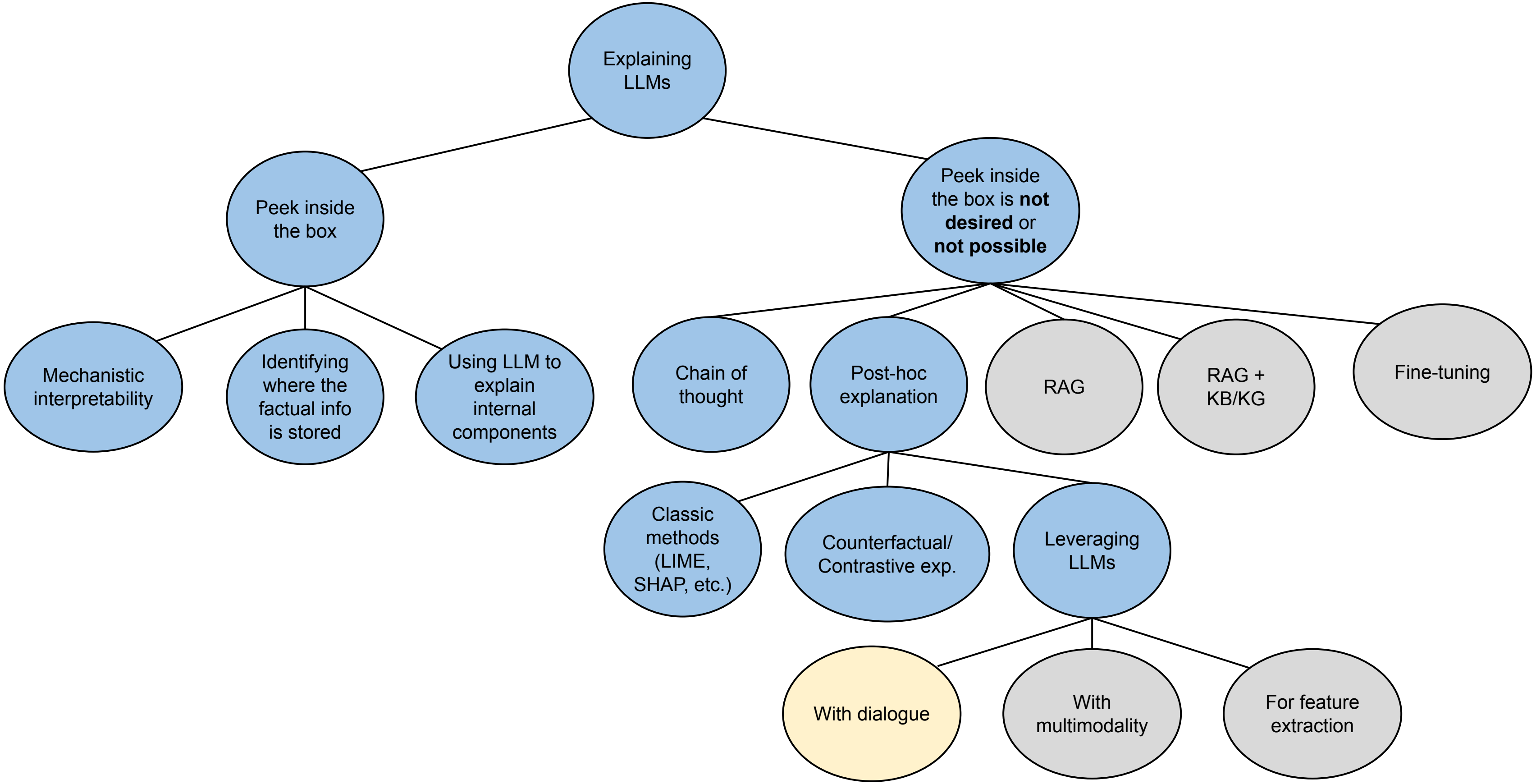
- Local explanations
- Use variations of the input to understand the behavior of the model
- Generally done by modifying the prompt. Examples:
  - Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobelt, H., & El-Assady, M. (2024). **Interactive analysis of LLMs using meaningful counterfactuals.** *arXiv:2405.00708*.
  - Luss, R., Miehling, E., & Dhurandhar, A. (2024). **CELL your model: Contrastive explanation methods for large language models.** *arXiv:2406.11785*.
- But if the black box can be opened, can be done by computing the saliency. Example:
  - Yin, K., & Neubig, G. (2022). Interpreting language models with contrastive explanations. In *Proceedings of EMNLP* (pp. 184-198).

Black box – Post-hoc – Counterfactual/contrastive exp. – Example



Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobelt, H., & El-Assady, M. (2024). Interactive analysis of LLMs using meaningful counterfactuals. arXiv:2405.00708.

What about leveraging an LLM to explain an LLM?



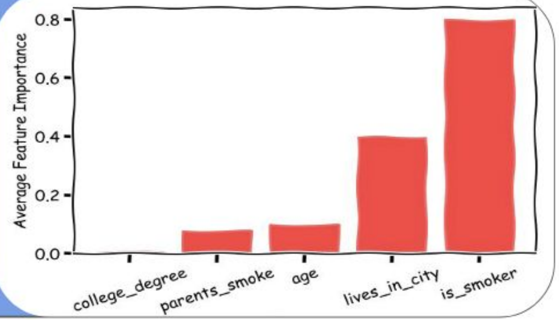
Black box – Leveraging the LLM – With dialogue

- Local explanations
- Chat with the LLM to better understand a certain output
- While seemingly naive, makes it possible
  - to have customized explanations
  - to view the explanations under different viewpoints
  - to have understandable explanations for non-expert users
- Examples:
  - Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). **Rethinking explainability as a dialogue: A practitioner's perspective.** In *Proceedings of NeurIPS Workshop on Human Centered AI*.
  - Wang, Q., Anikina, T., Feldhus, N., van Genabith, J., Hennig, L., & Möller, S. (2024). **LLMCheckup: Conversational examination of large language models via interpretability tools.** *arXiv:2401.12576*.
  - Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., & Zhang, J. (2023). **Chat-REC: Towards interactive and explainable LLMs-augmented recommender system.** *arXiv:2303.14524*.

Black box – Leveraging the LLM – With dialogue

FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

GOOD QUESTION! IT LOOKS LIKE THE MODEL PREDICTED THESE INDIVIDUALS AS HIGH RISK MOSTLY BECAUSE THEY WERE SMOKERS BUT ALSO BECAUSE THEY LIVE IN LARGE CITIES. I'M HIGHLY CONFIDENT THESE ARE THE REASONS BECAUSE THE EXPLANATIONS HAVE HIGH FIDELITY. HERE'S THE AVERAGE FEATURE IMPORTANCE FOR THESE PEOPLE (HIGHER MEANS MORE IMPORTANT).

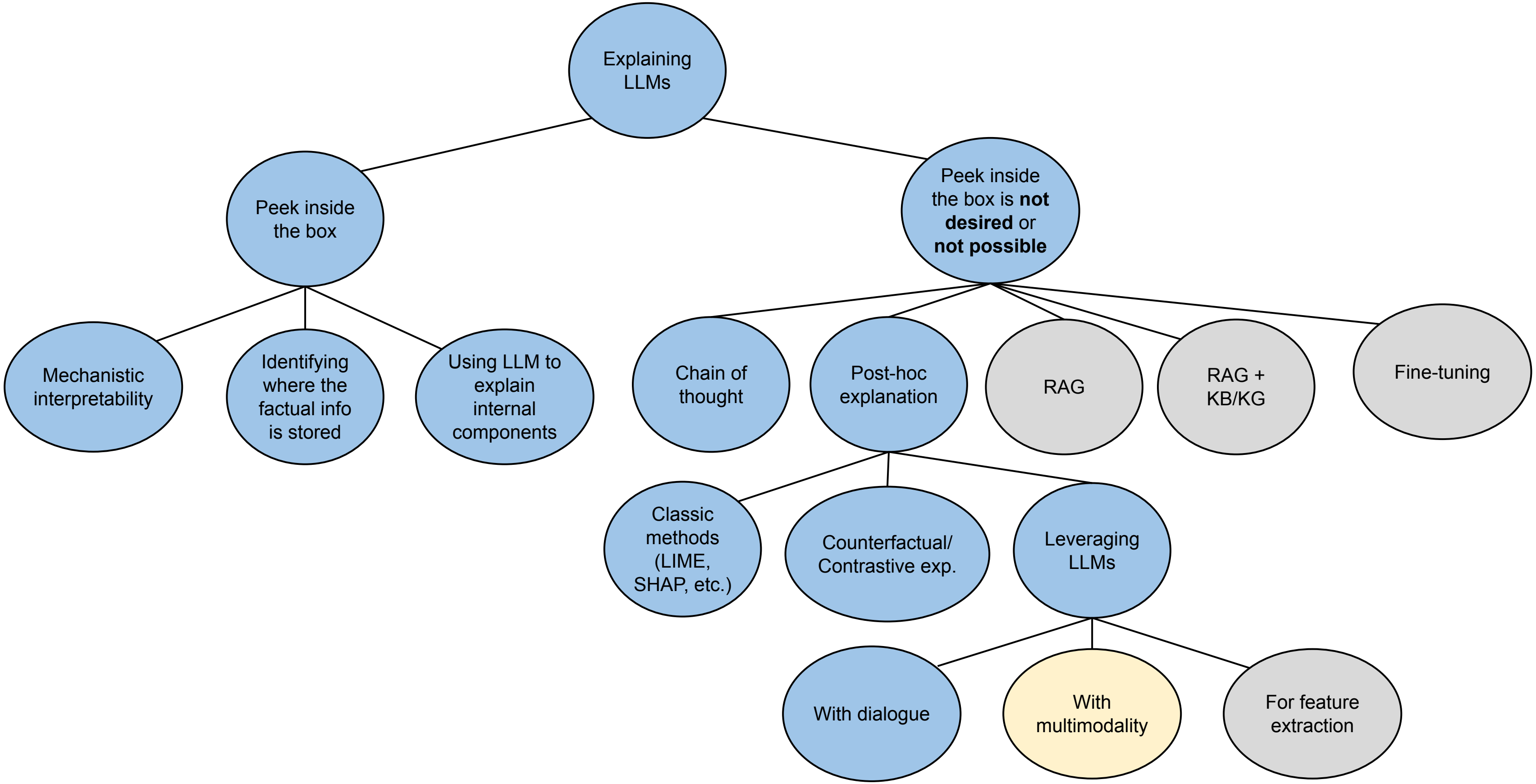


WOW, IT'S SURPRISING THAT WHETHER THE PERSON LIVES IS IN A CITY IS SO IMPORTANT.

YES, LIVES_IN_CITY HAS A SIGNIFICANT EFFECT ON THE PREDICTIONS FOR THESE INDIVIDUALS. PERTURBING THIS FEATURE CAN FLIP THE PREDICTION FOR 4 OF 15 OF THE INSTANCES IN THIS GROUP.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner's perspective. In *Proceedings of NeurIPS Workshop on Human Centered AI*.

What about multimodality?



Black box – Leveraging the LLM – With multimodality

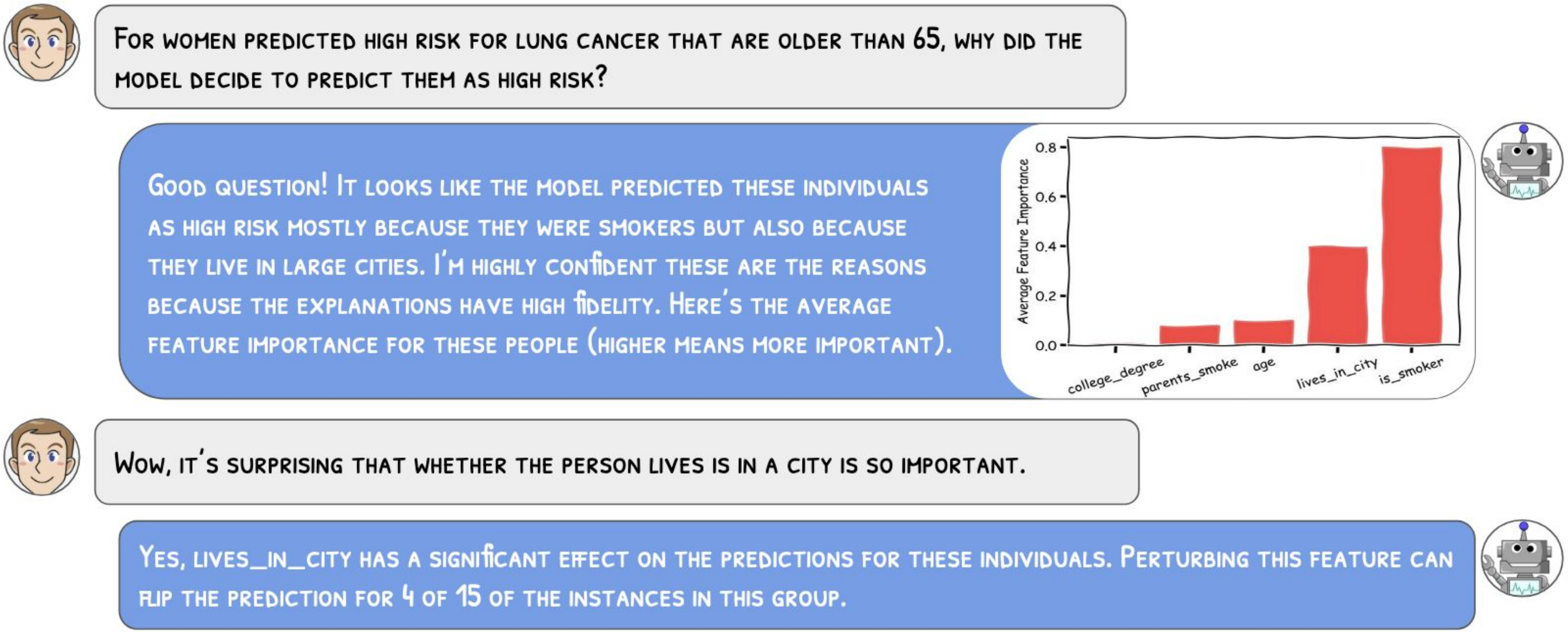
- Local explanations

- Use multimodality to help with the explanation

- Many ways to leverage modality. For instance,
  - Generating an image to explain some input text (or vice-versa), or
  - Explaining things in both an input text and input image

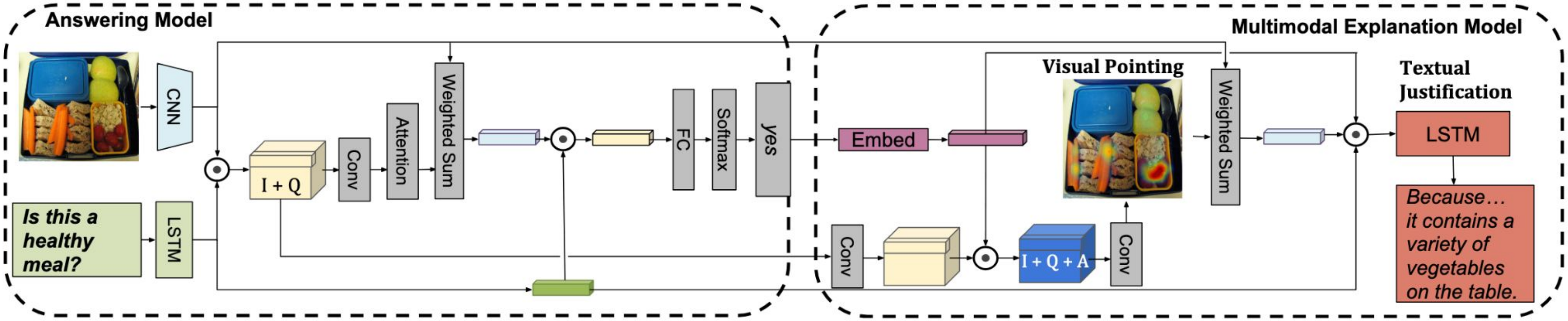
- Examples:
  - Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). **Rethinking explainability as a dialogue: A practitioner's perspective**. In *Proceedings of NeurIPS Workshop on Human Centered AI*.
  - Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). **Multimodal explanations: Justifying decisions and pointing to the evidence.** In *Proceedings of CVPR* (pp. 8779-8788).

Black box – Leveraging the LLM – With multimodality – Example



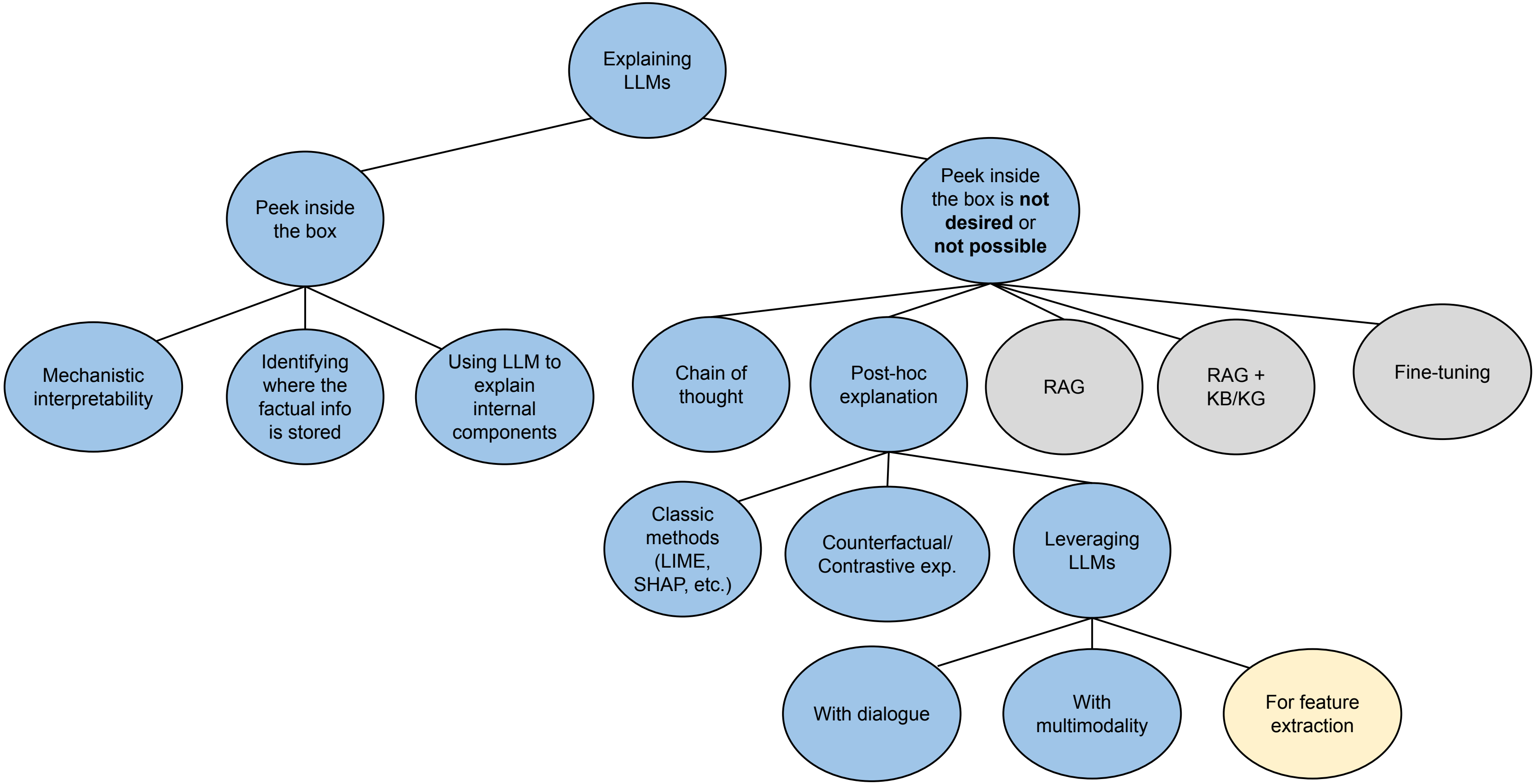
Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner's perspective. In *Proceedings of NeurIPS Workshop on Human Centered AI*.

Black box – Leveraging the LLM – With multimodality – Example

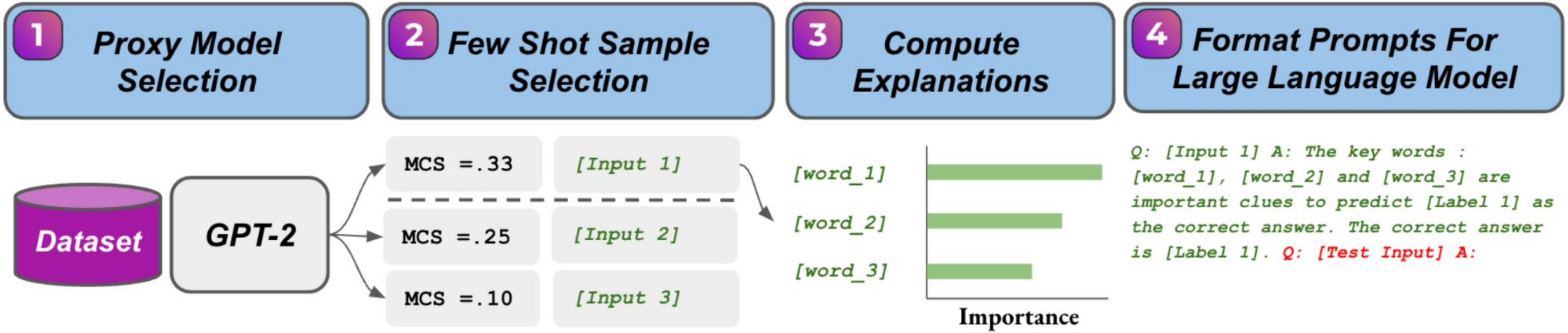


Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of CVPR* (pp. 8779-8788).

What about leveraging an LLM to explain an LLM?



Remember? (Black box – Post-hoc – Classic methods)

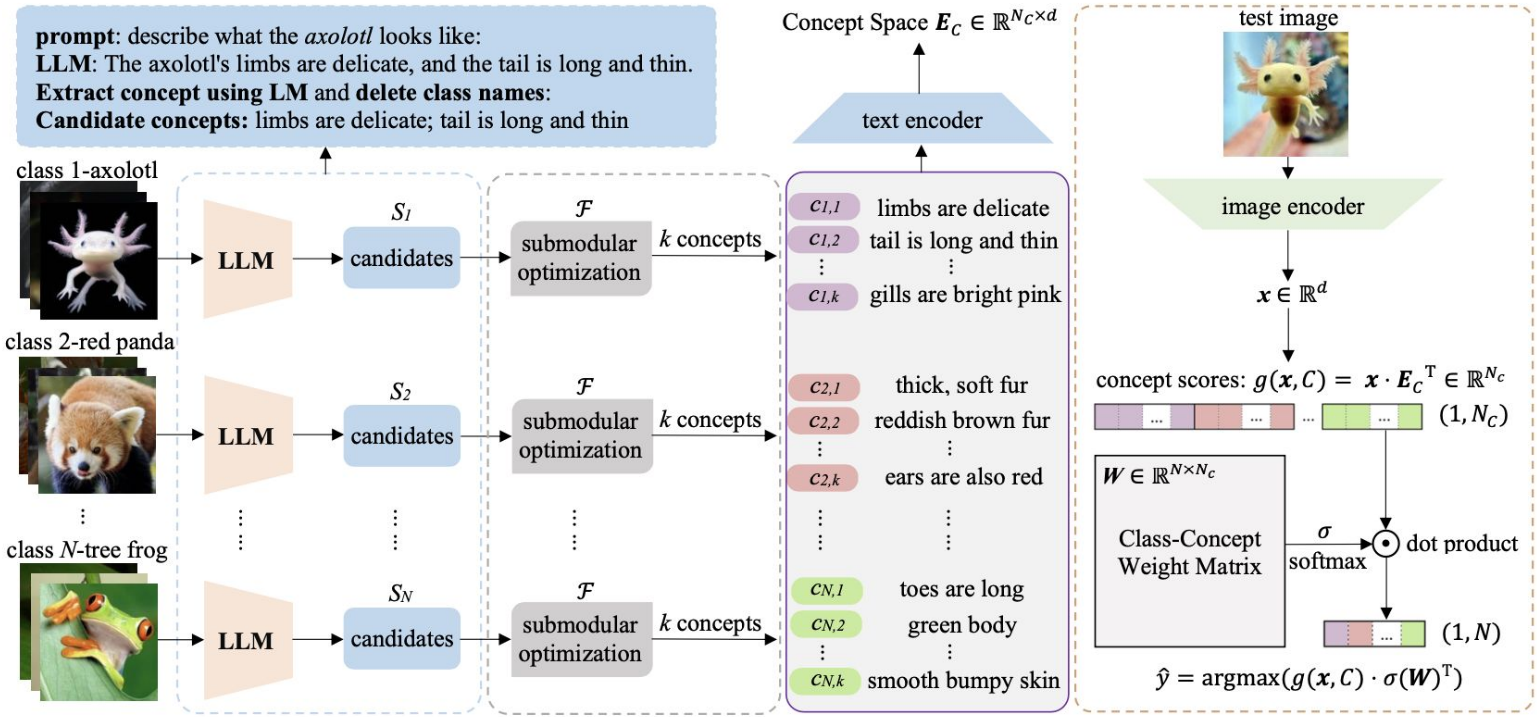


Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., & Lakkaraju, H. (2024). Post hoc explanations of language models can improve language models. In *Proceedings of NeurIPS*, 36, 65468-65483.

Black box – Leveraging the LLM – Feature Extraction

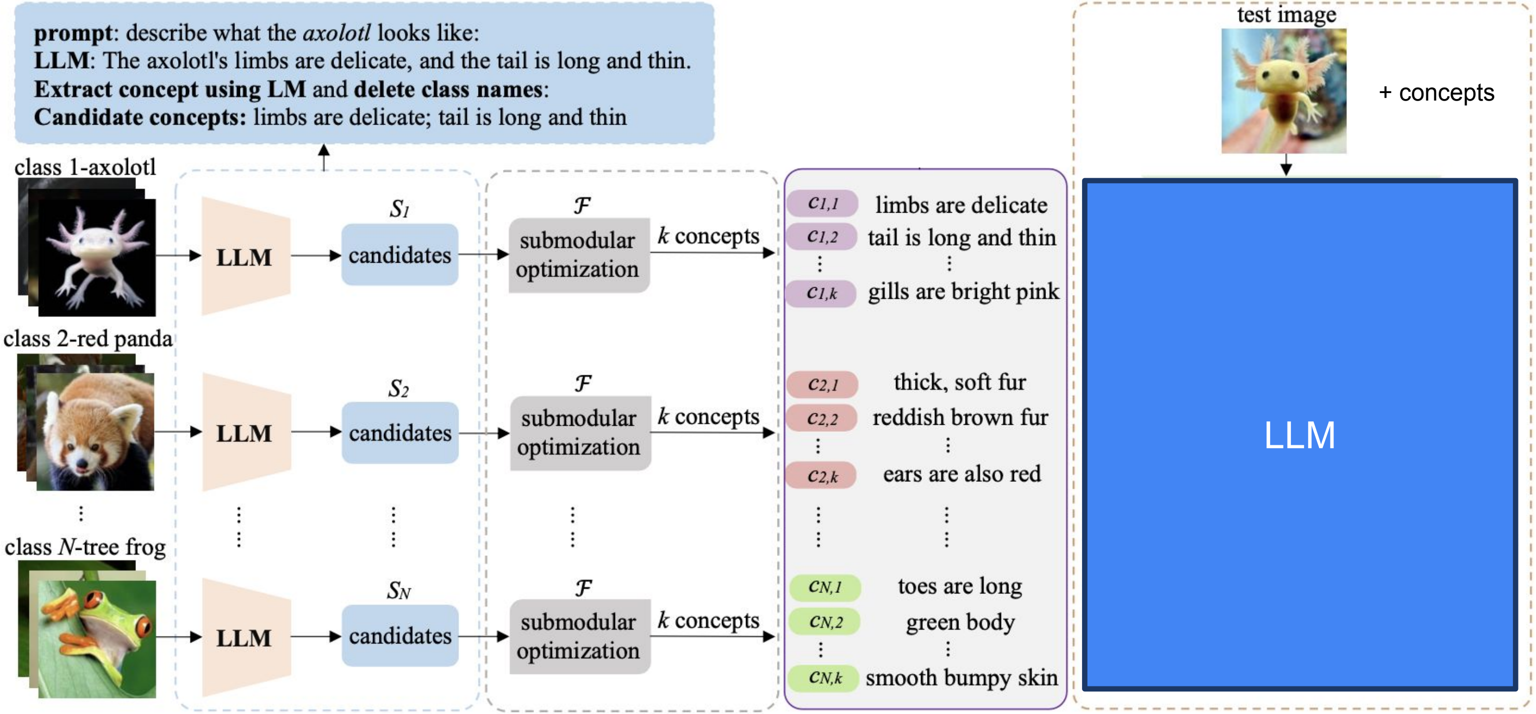
- Local explanations
- Use an LLM
  - to extract complex features (e.g., concepts)
  - then ask a subsequent LLM to use them in the reasoning
- Advantage: same as word importance, but with more meaningful features
- Example: Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). **Language in a bottle: Language model guided concept bottlenecks for interpretable image classification.** In *Proceedings of CVPR* (pp. 19187-19197).

Black box – Leveraging the LLM – Feature Extraction – Example



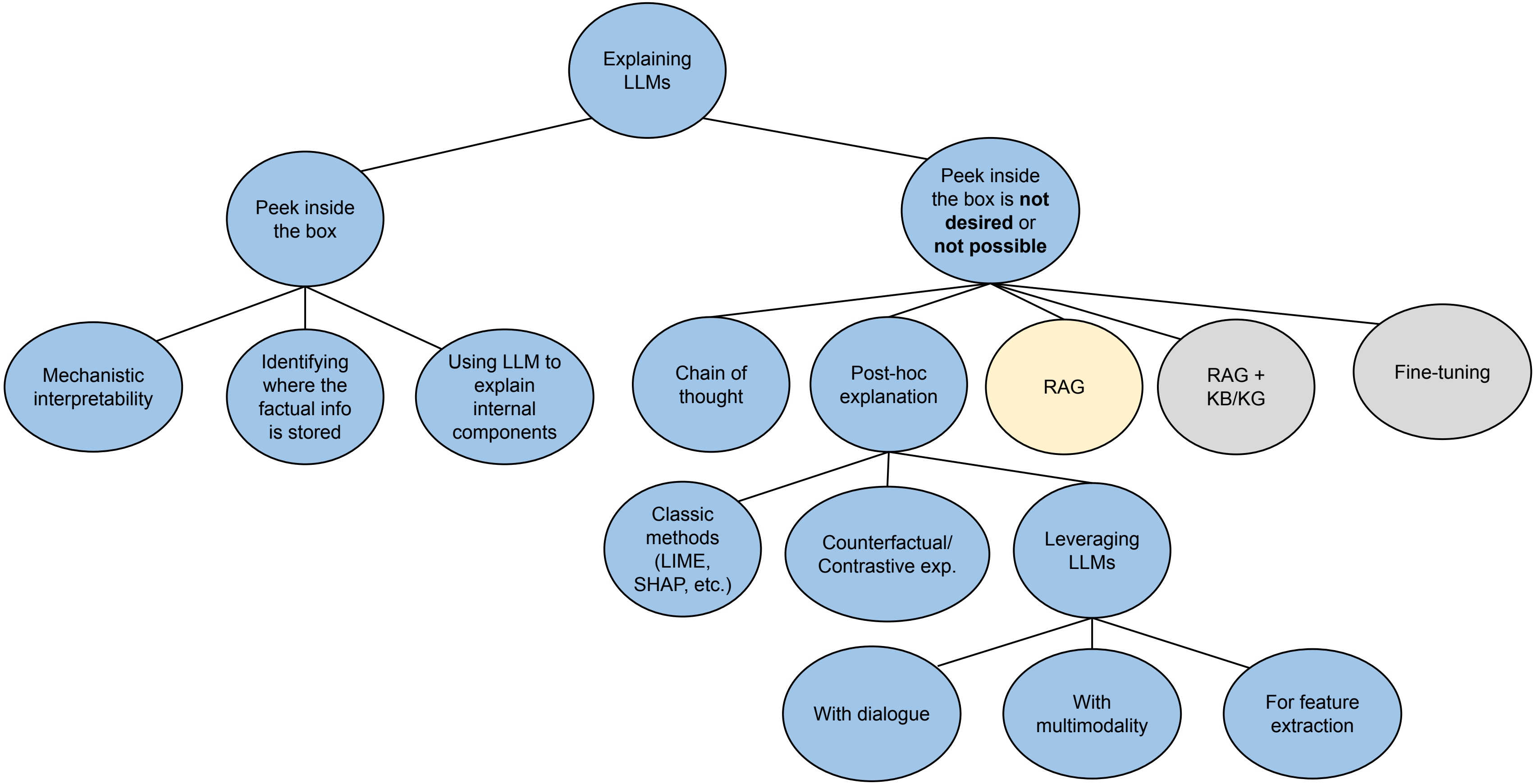
Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of CVPR* (pp. 19187-19197).

Black box – Leveraging the LLM – Feature Extraction – Example



Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., & Yatskar, M. (2023). Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of CVPR* (pp. 19187-19197).

What about Retrieval-Augmented Generation (RAG)?



Black box – RAG

- Local explanations

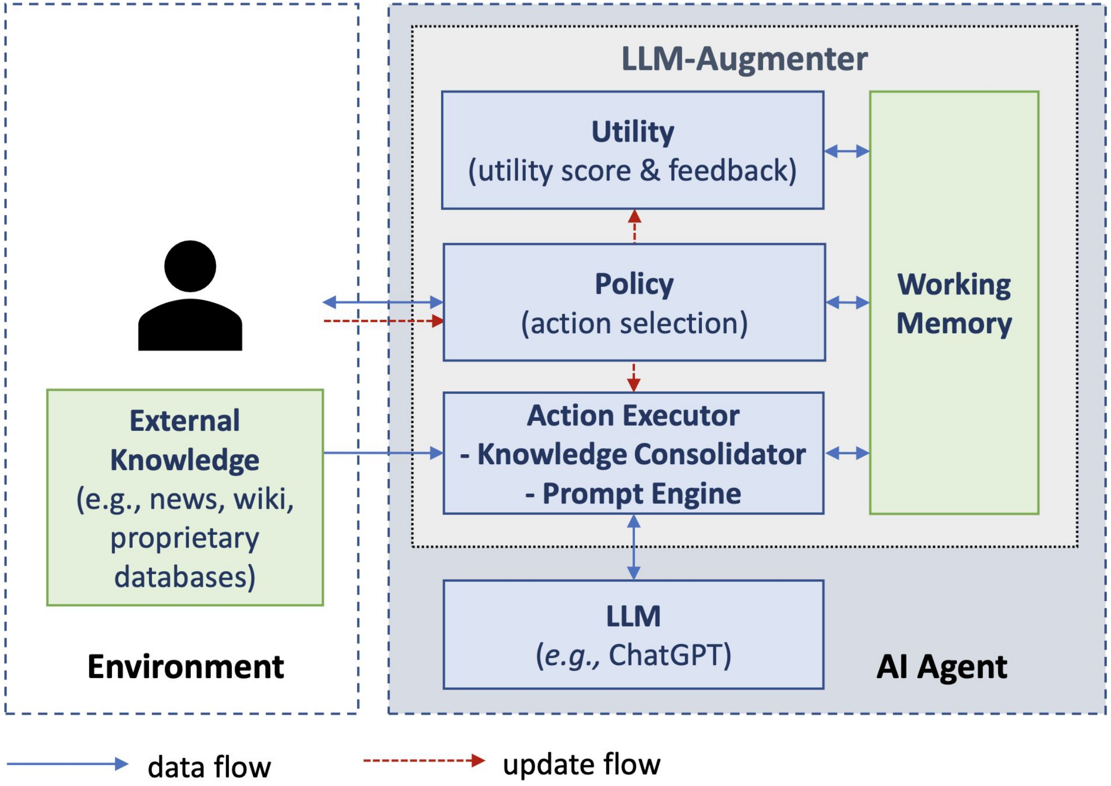
- Use external knowledge to ground explanations in reality

- PRO: RAG reduces hallucinations (Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). **Retrieval augmentation reduces hallucination in conversation**. In *Findings of EMNLP* (pp. 3784-3803)), which also affects explanations

  CONs:
    - Constrained by the external knowledge used
    - Does not nullify the risk of hallucinations

- Examples:
  - Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... & Gao, J. (2023). **Check your facts and try again: Improving large language models with external knowledge and automated feedback.** *arXiv:2302.12813*.
  - Tekkesinoglu, S., & Kunze, L. (2024). **From feature importance to natural language explanations using LLMs with RAG.** *arXiv:2407.20990*.

Black box – RAG – Example



Based on input,

acquire the necessary knowledge,

then loop

- build prompt
- query LLM
- fact check response

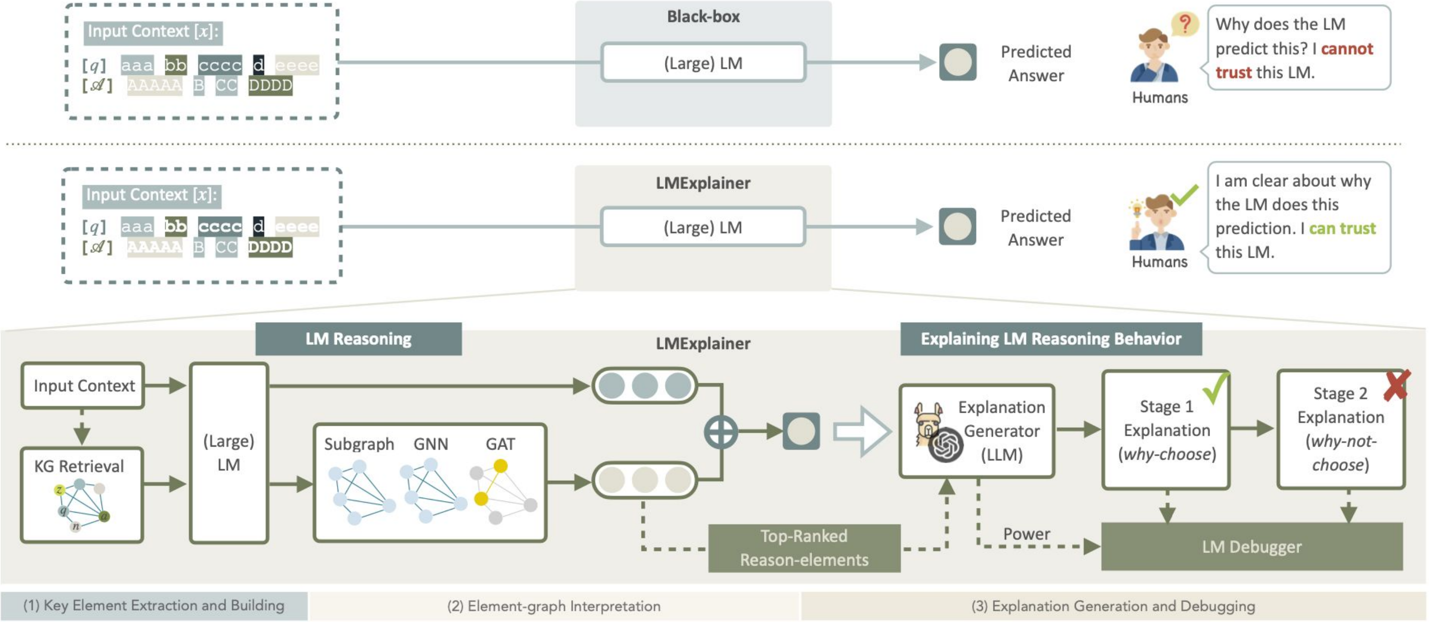
until response is factually correct

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., ... & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv:2302.12813*.

Black box – RAG + Knowledge Graph/Base

- Local explanations
- Use external knowledge to ground explanations in reality… But from KG/BG
- Pro: Easier to retrieve and check information
  Con: The information is constrained by the structure of the KG/BG
- Examples:
  - He, H., Zhang, H., & Roth, D. (2022). **Rethinking with retrieval: Faithful large language model inference.** *arXiv:2301.00303*.
  - Chen, Z., Singh, A. K., & Sra, M. (2023). **LMExplainer: A knowledge-enhanced explainer for language models.** *arXiv:2303.16537*.
  - Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). **Unifying large language models and knowledge graphs: A roadmap.** *IEEE Transactions on Knowledge & Data Engineering*, *36*(07), 3580-3599.

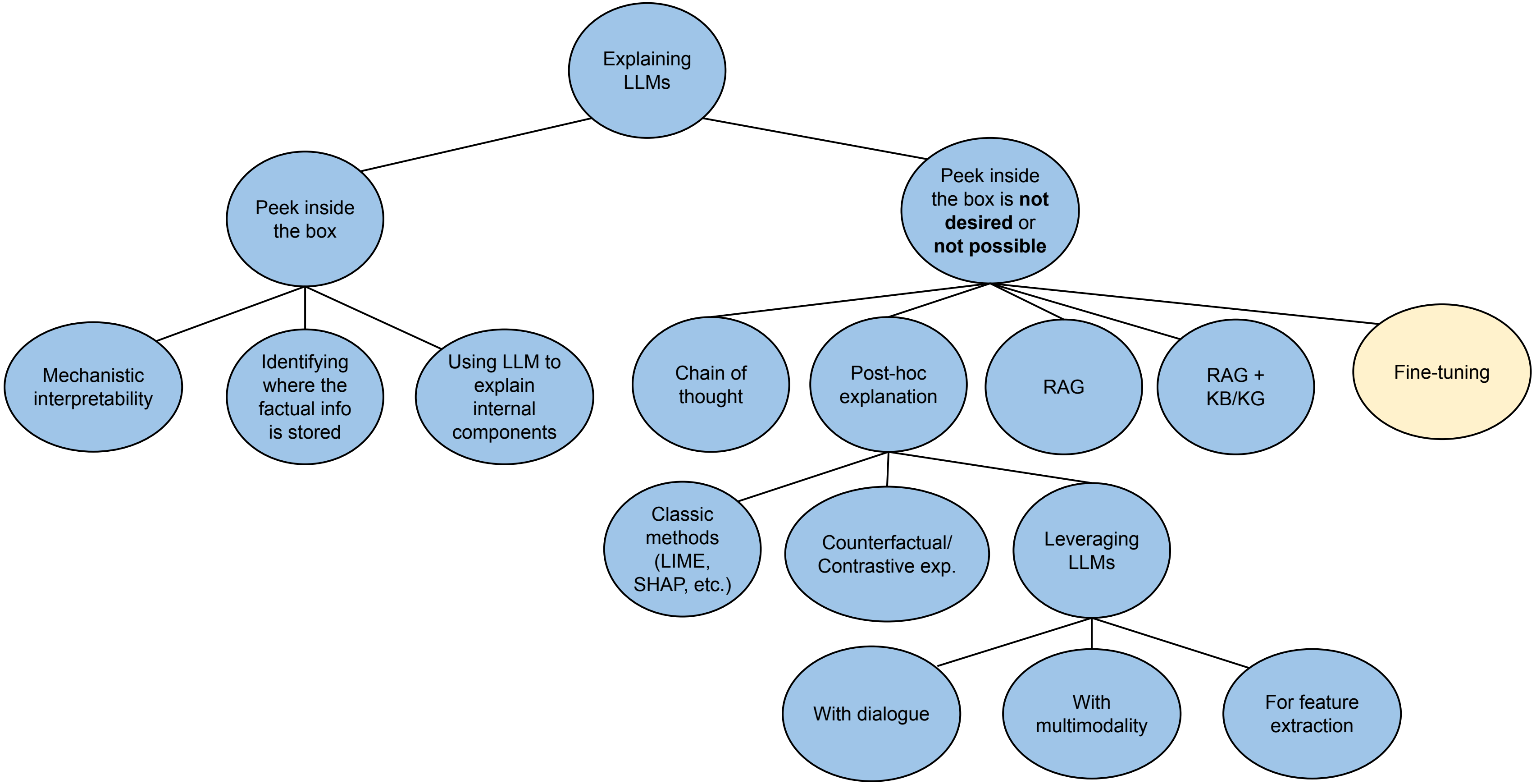
Black box – RAG + Knowledge Graph/Base – Example



Retrieve relevant nodes in KG; provide them to LLM; highlight important nodes for decision; provide answer and highlighted nodes to another LLM for explanation

Chen, Z., Singh, A. K., & Sra, M. (2023). **LMExplainer: A knowledge-enhanced explainer for language models.** *arXiv:2303.16537*.

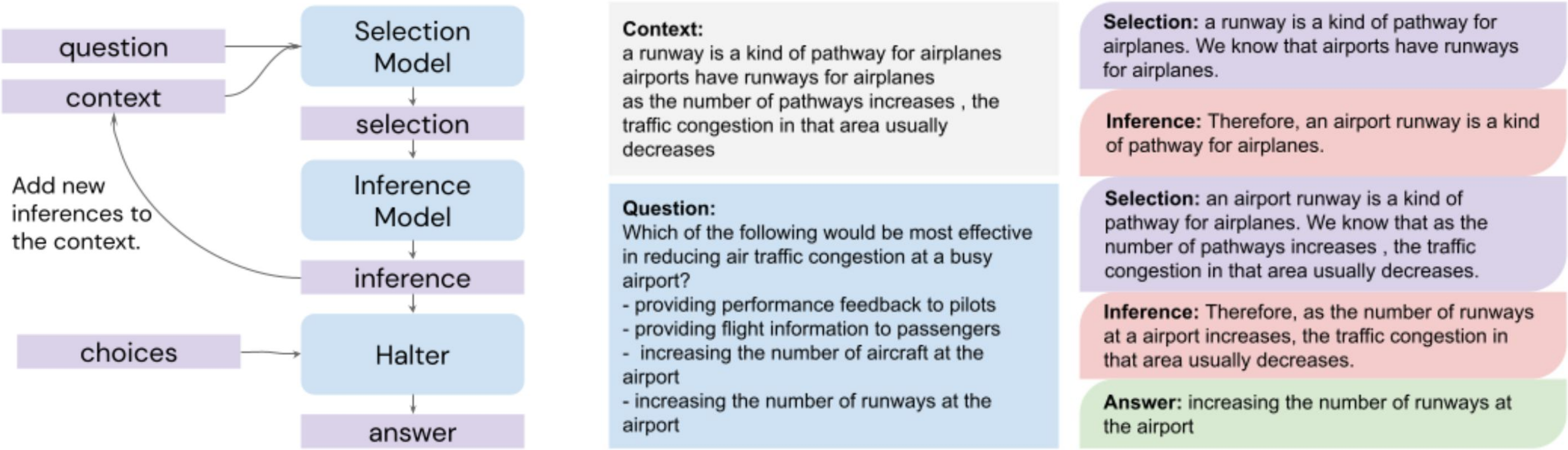
What about Fine-tuning?



Black box – Fine-tuning to Explain

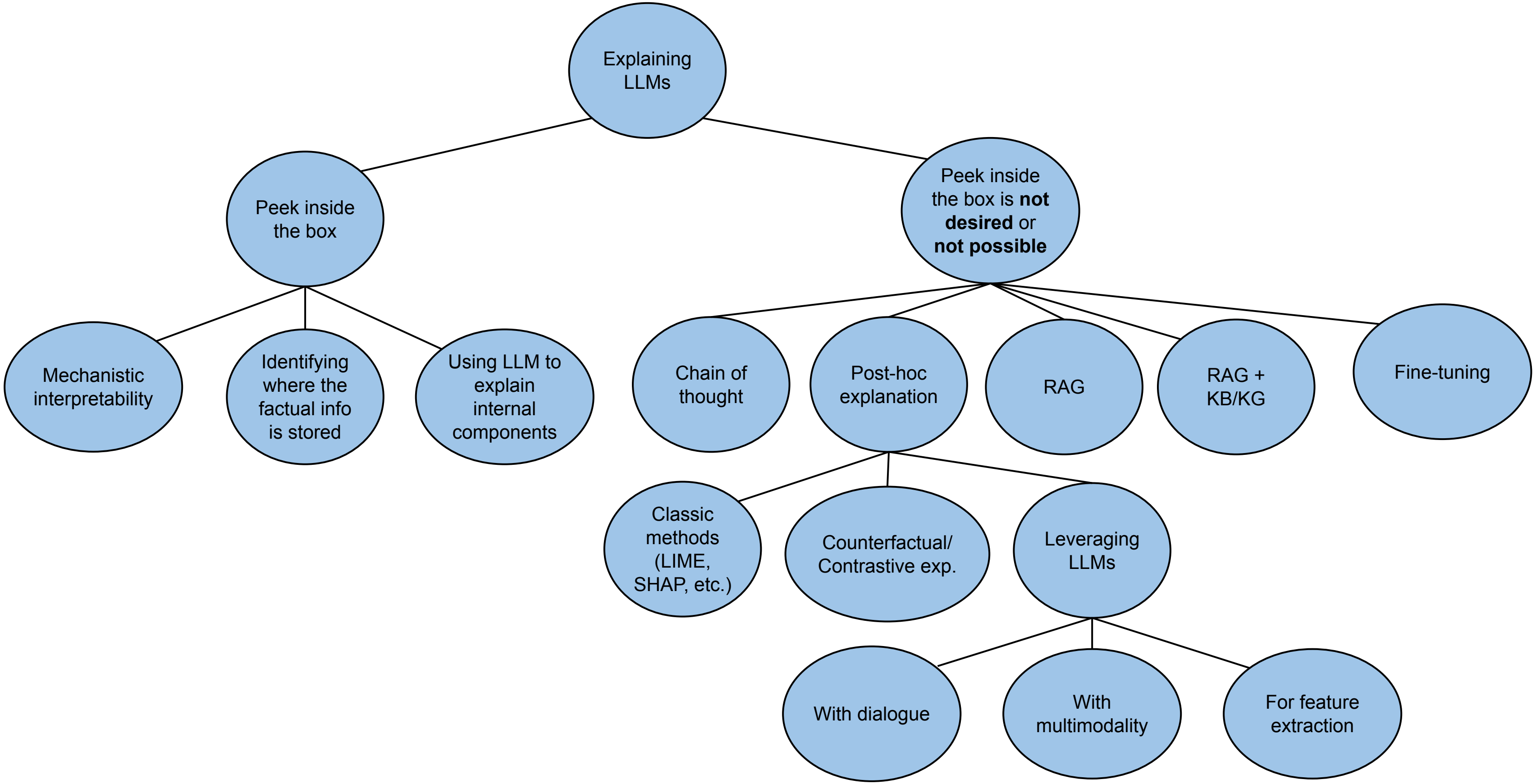
- Local explanation
- Training a model (or a set of models) to produce explanations
- The explanatory component is built in
- Example: Creswell, A., & Shanahan, M. (2022). **Faithful reasoning using large language models.** *arXiv:2208.14271*.

Black box – Fine-tuning to Explain – Example

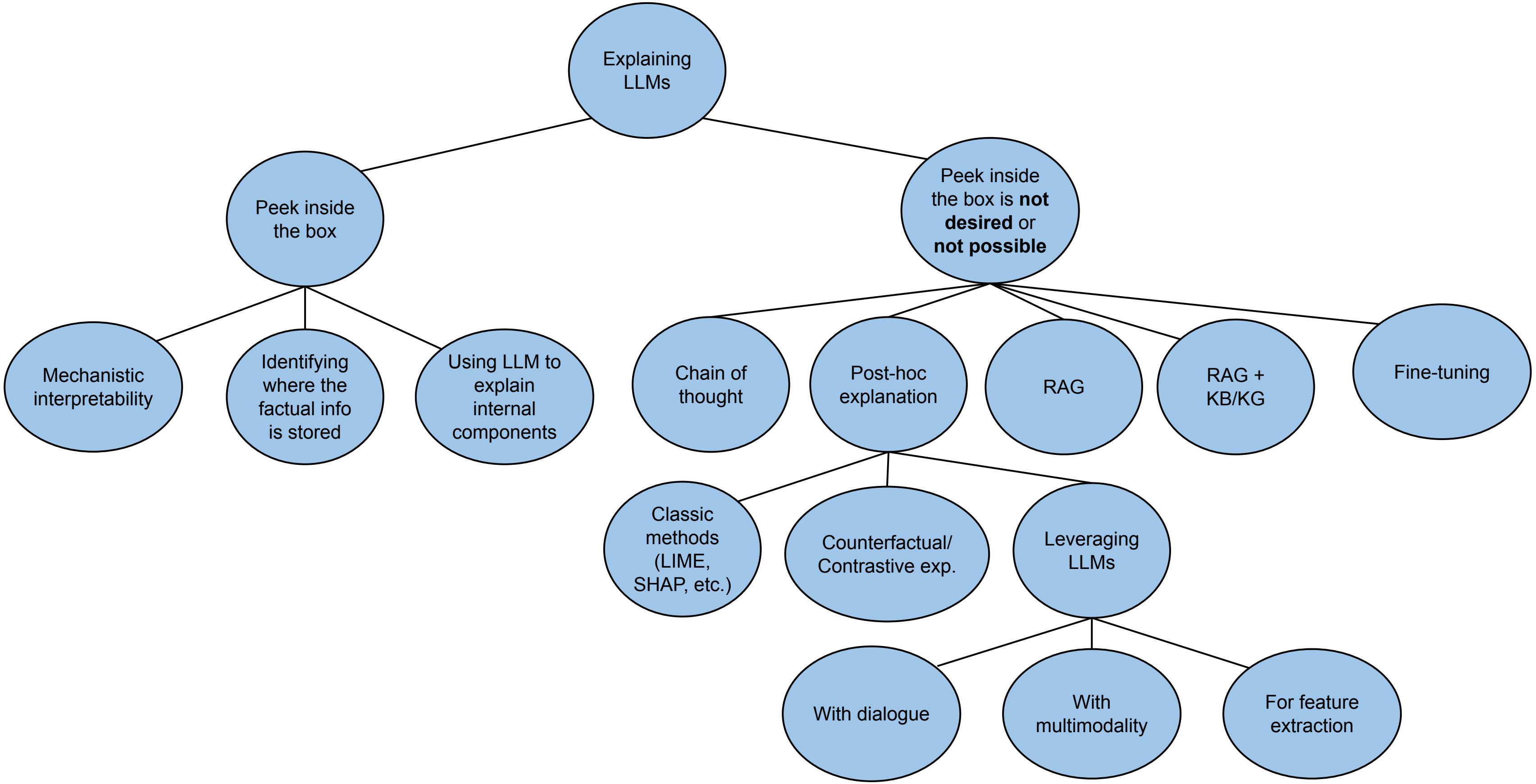


Creswell, A., & Shanahan, M. (2022). Faithful reasoning using large language models. *arXiv:2208.14271*.

Conclusion of Part 2: Explainability of LLMs



Anything you would like to add?



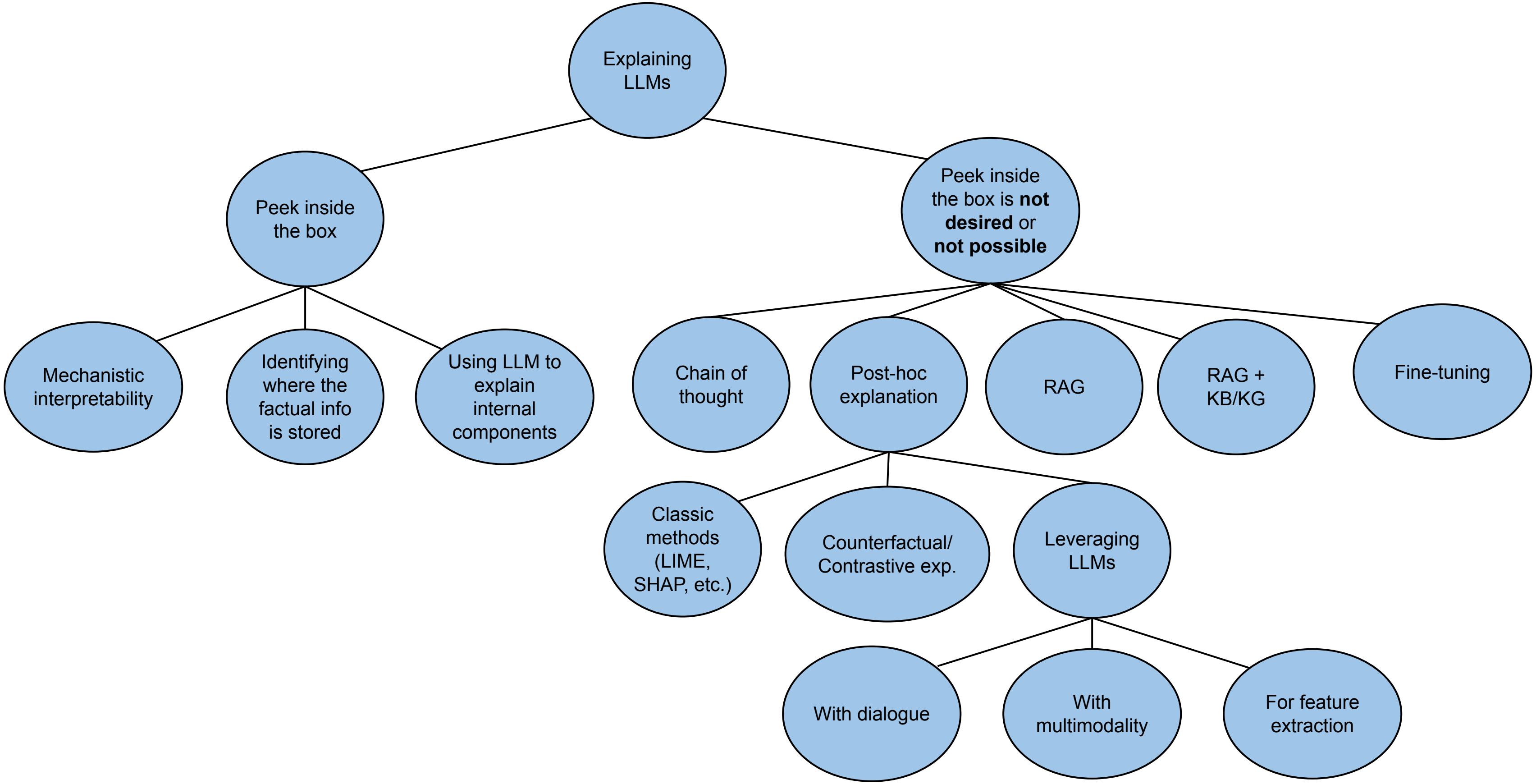
Suggested Reading

- Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). **Rethinking interpretability in the era of large language models.** arXiv:2402.01761.

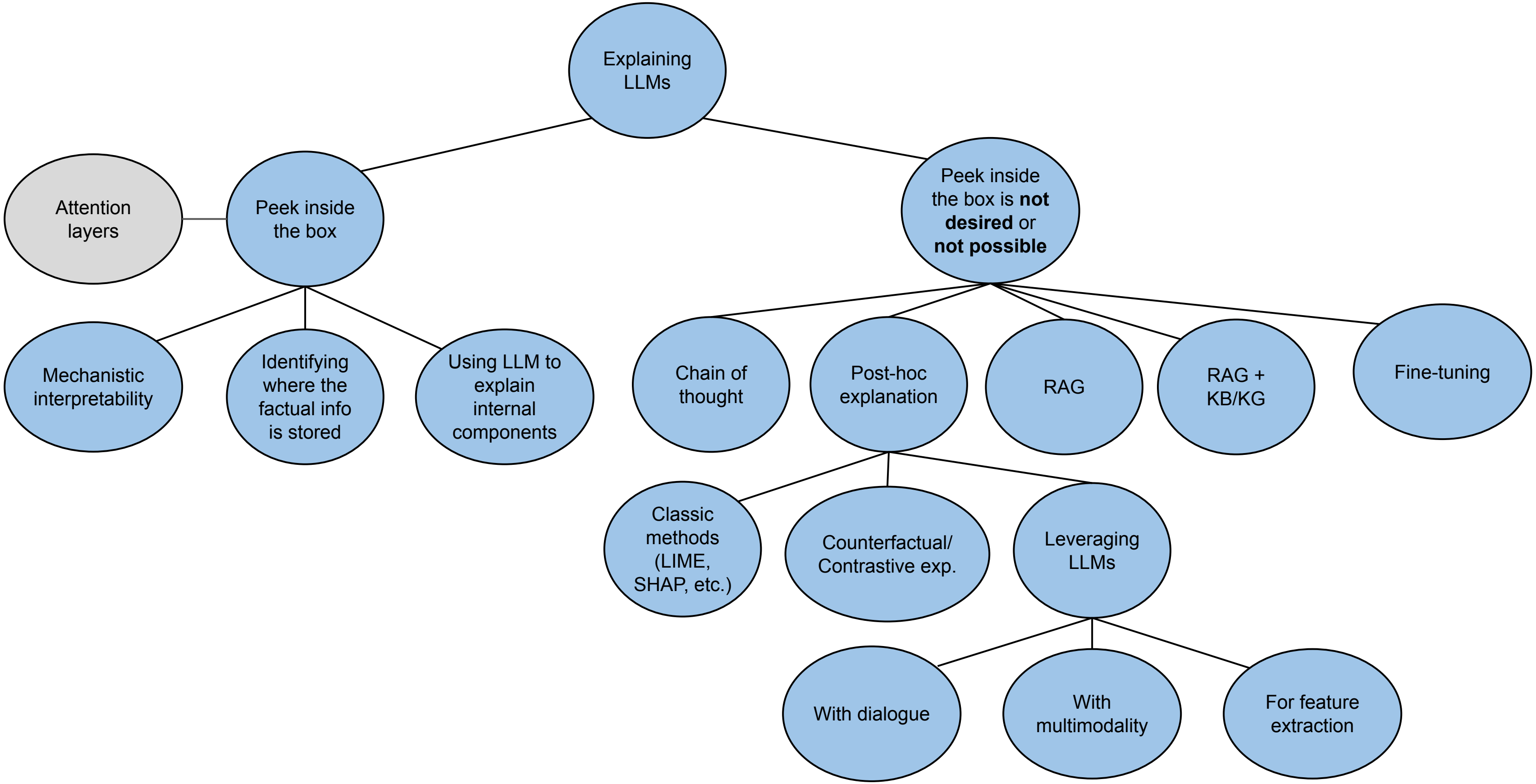
- Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Liu, T., ... & Liu, N. (2024). **Usable XAI: 10 strategies towards exploiting explainability in the LLM era.** arXiv:2403.08946.

Interesting structure

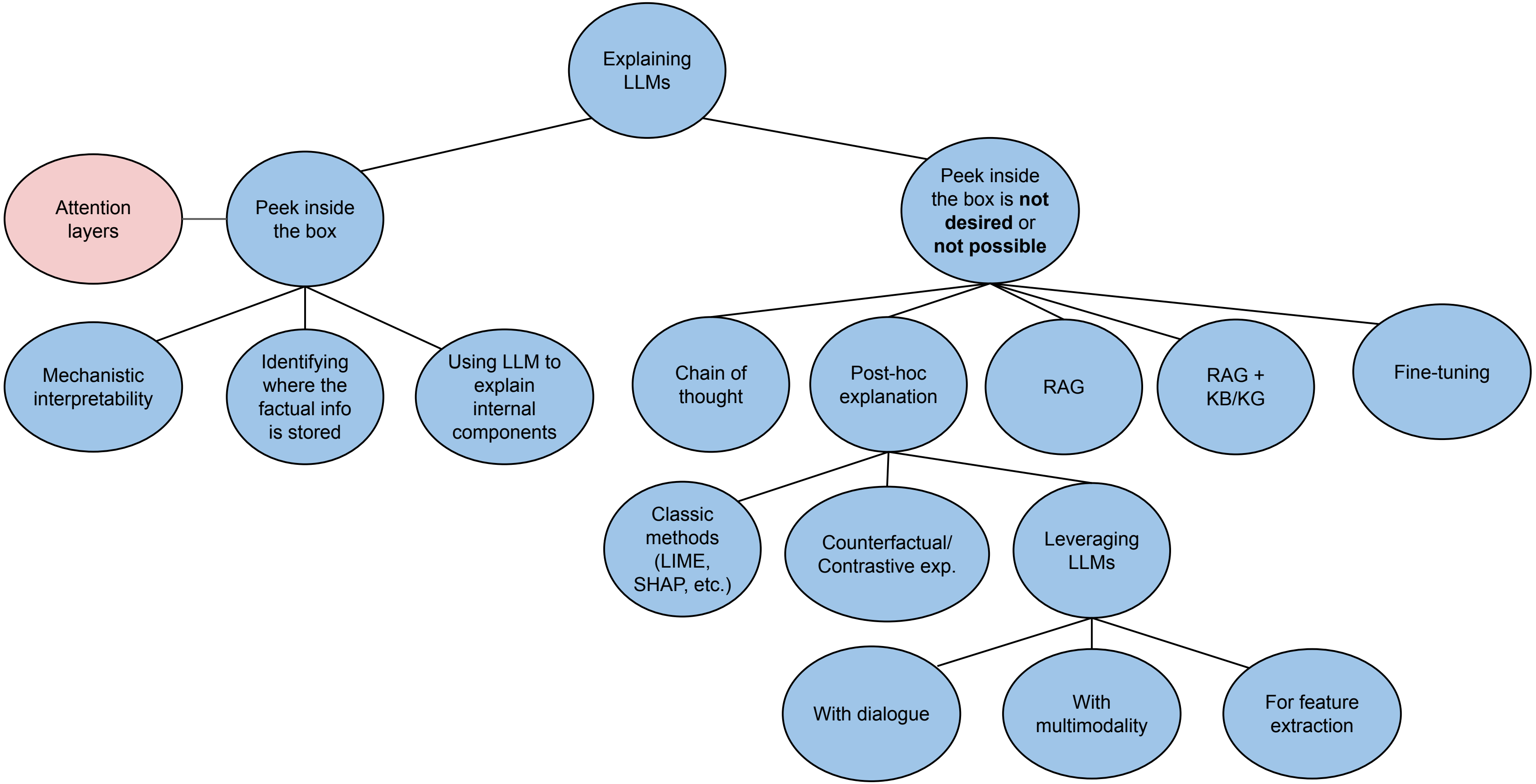
Conclusion of this Tutorial and Take Home Messages



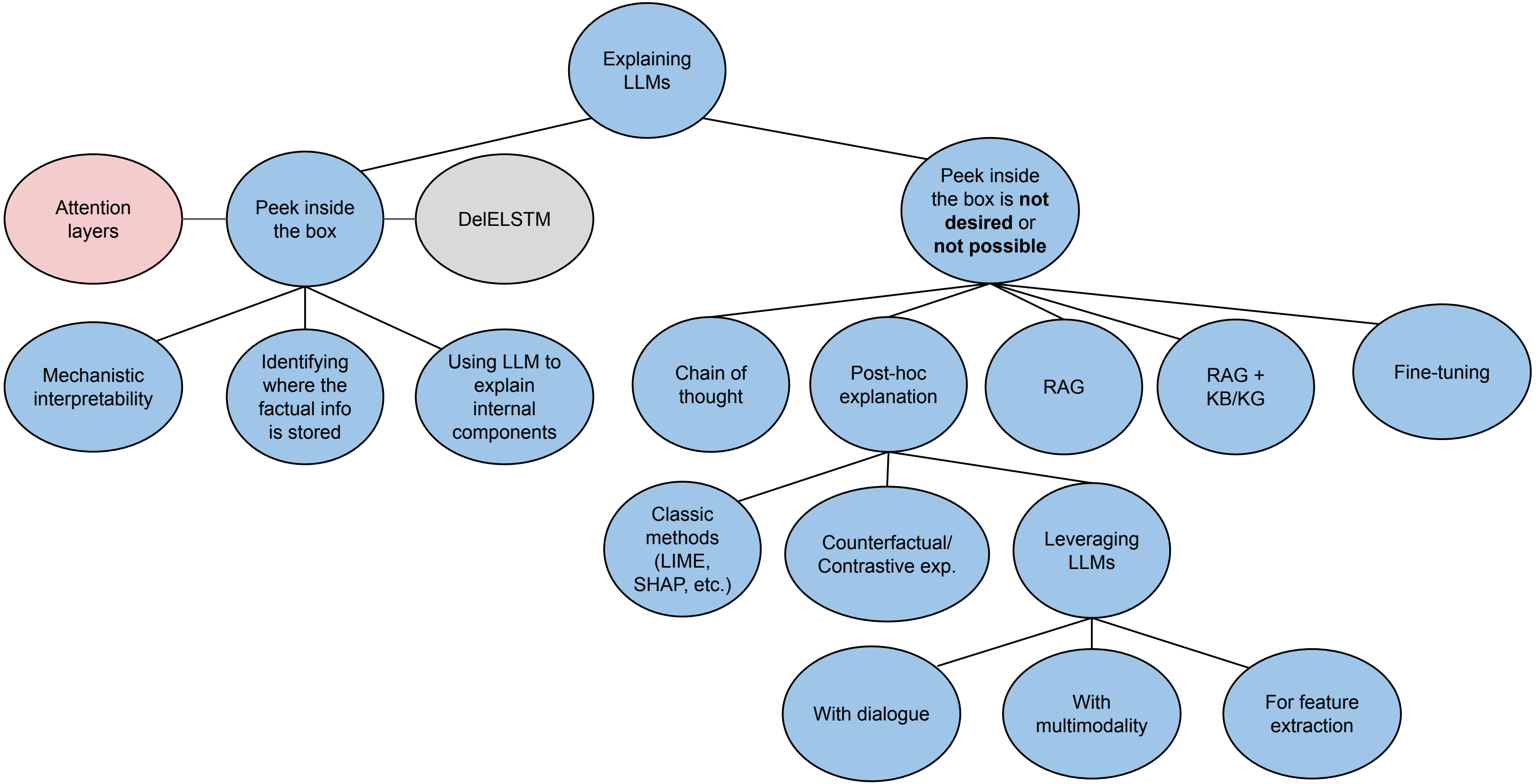
Conclusion of this Tutorial and Take Home Messages



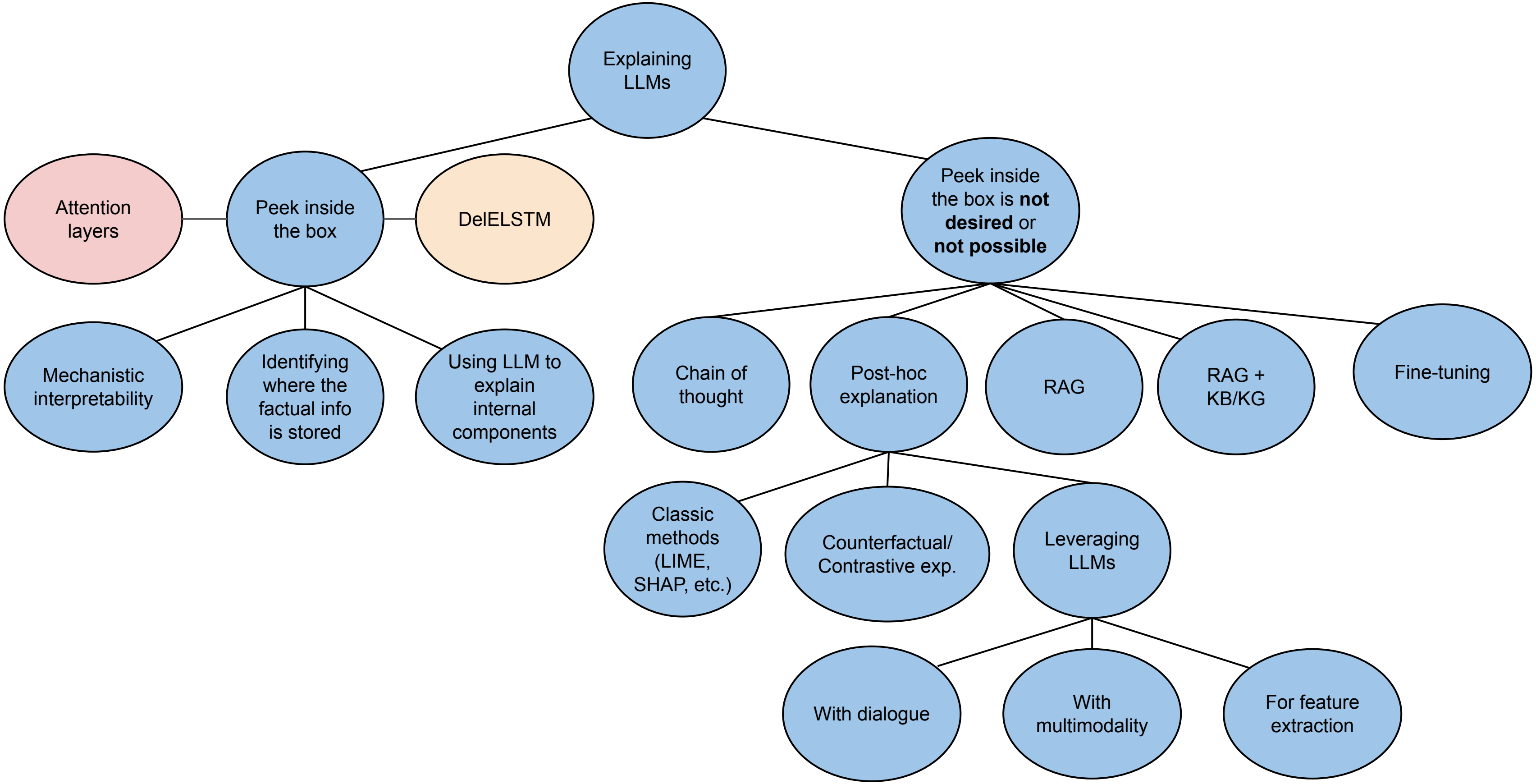
Conclusion of this Tutorial and Take Home Messages



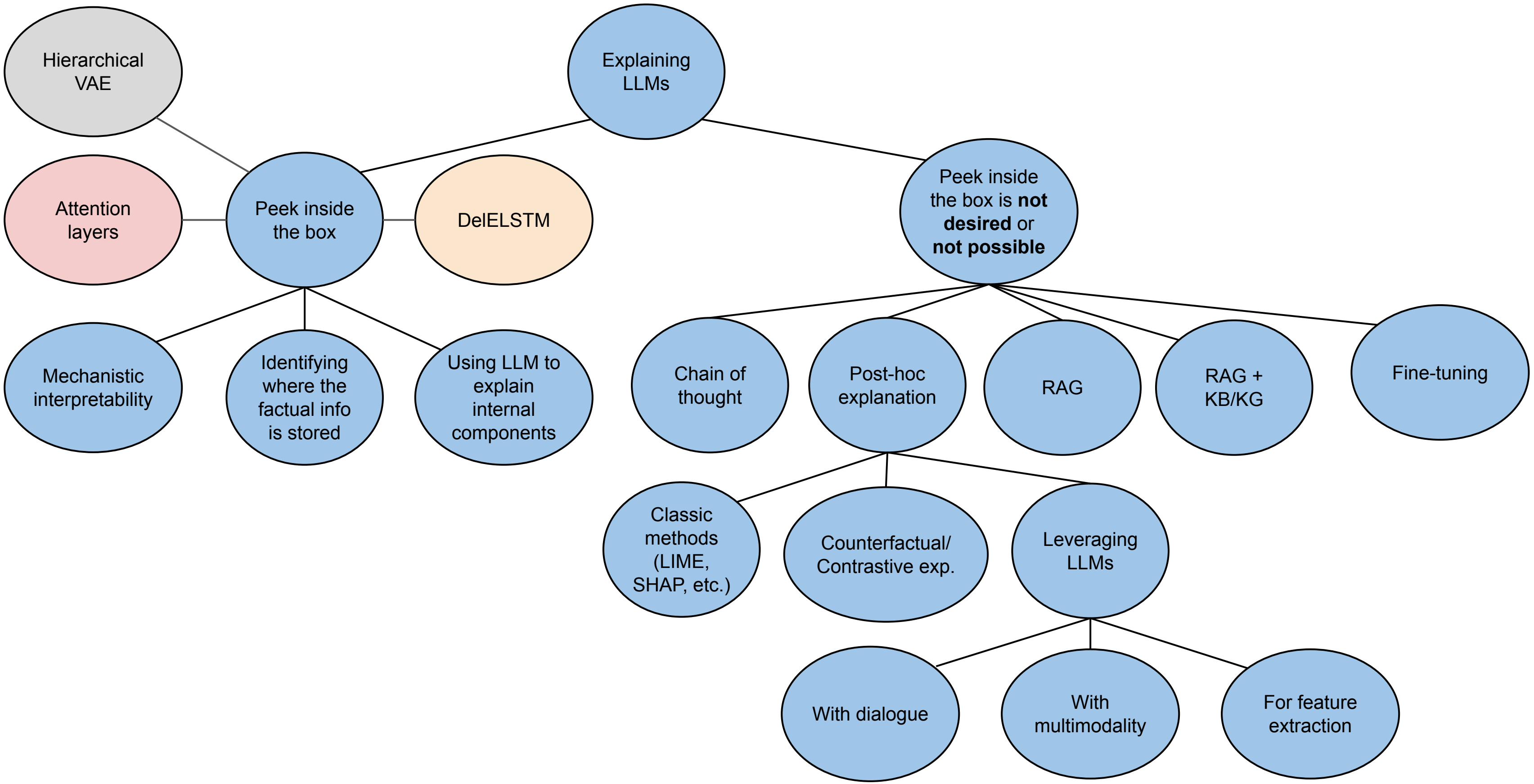
Conclusion of this Tutorial and Take Home Messages



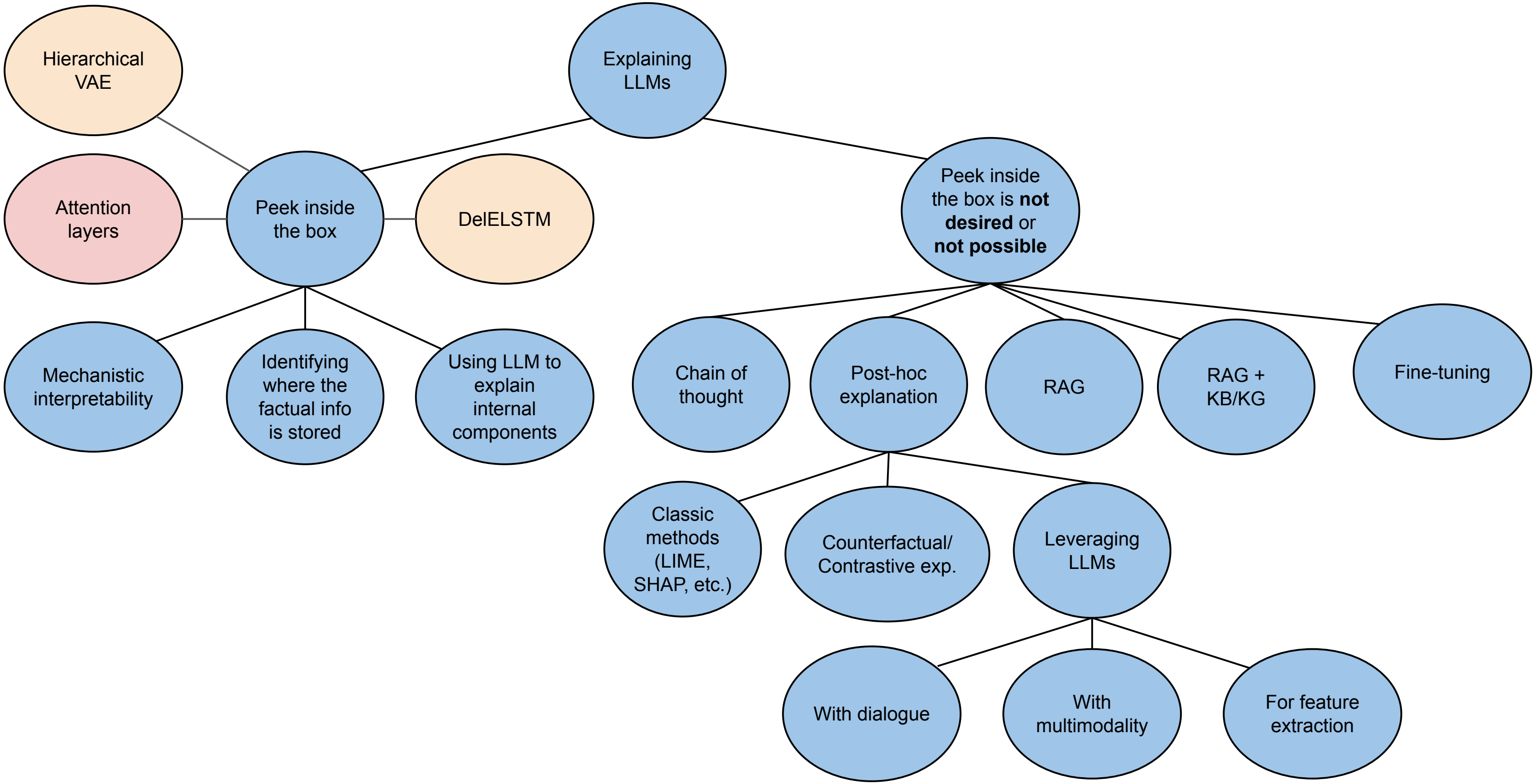
Conclusion of this Tutorial and Take Home Messages



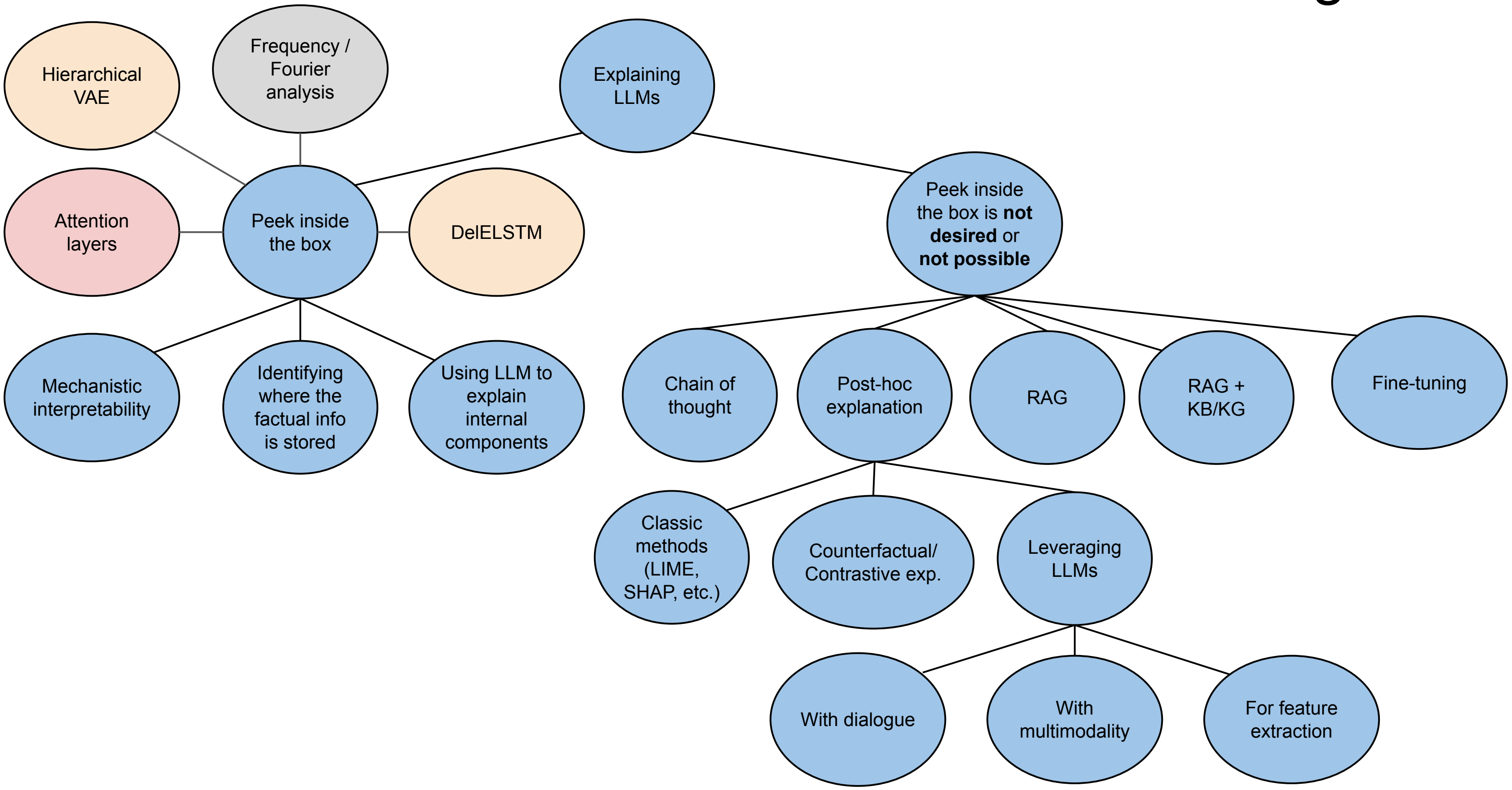
Conclusion of this Tutorial and Take Home Messages



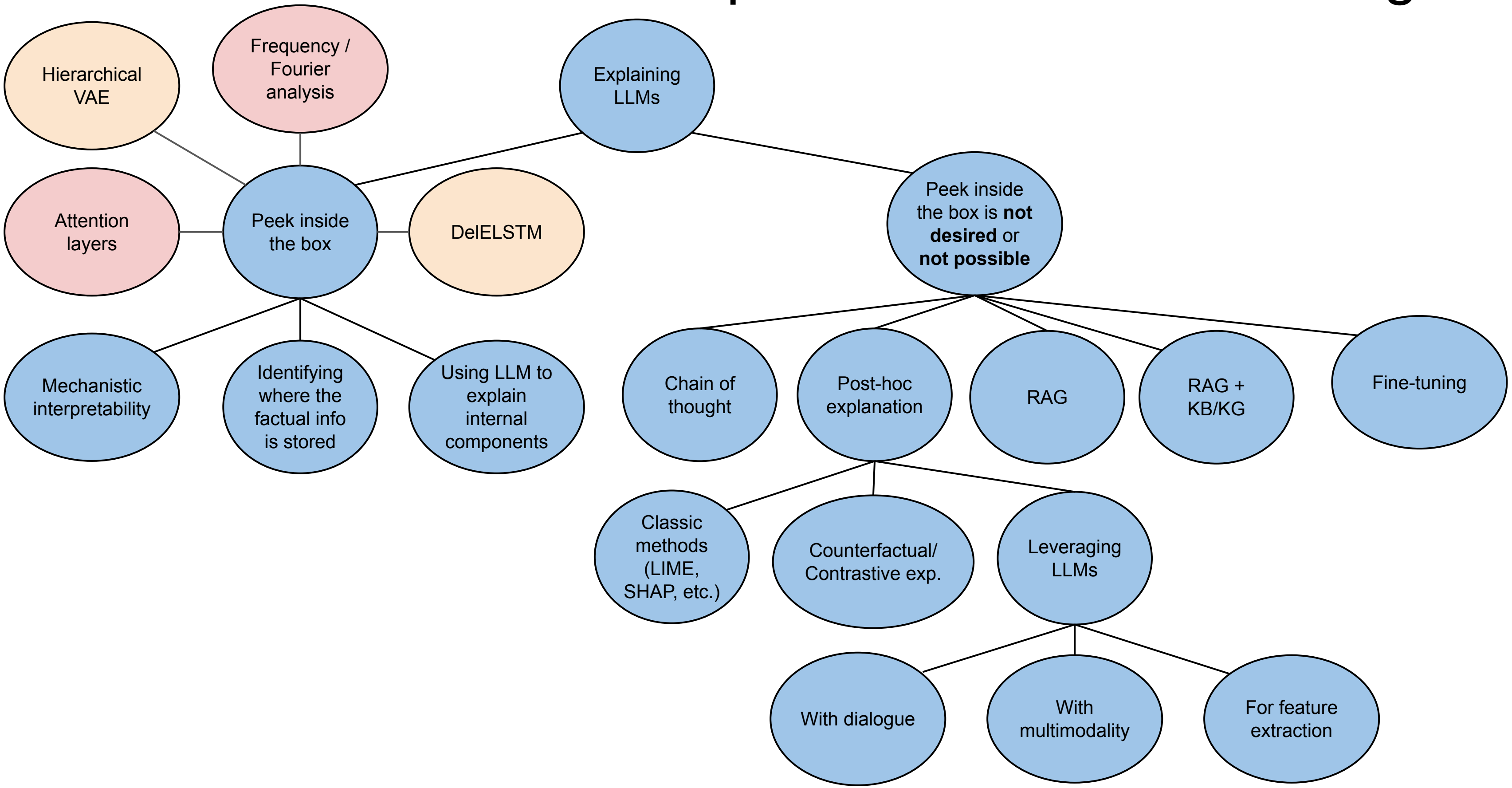
Conclusion of this Tutorial and Take Home Messages



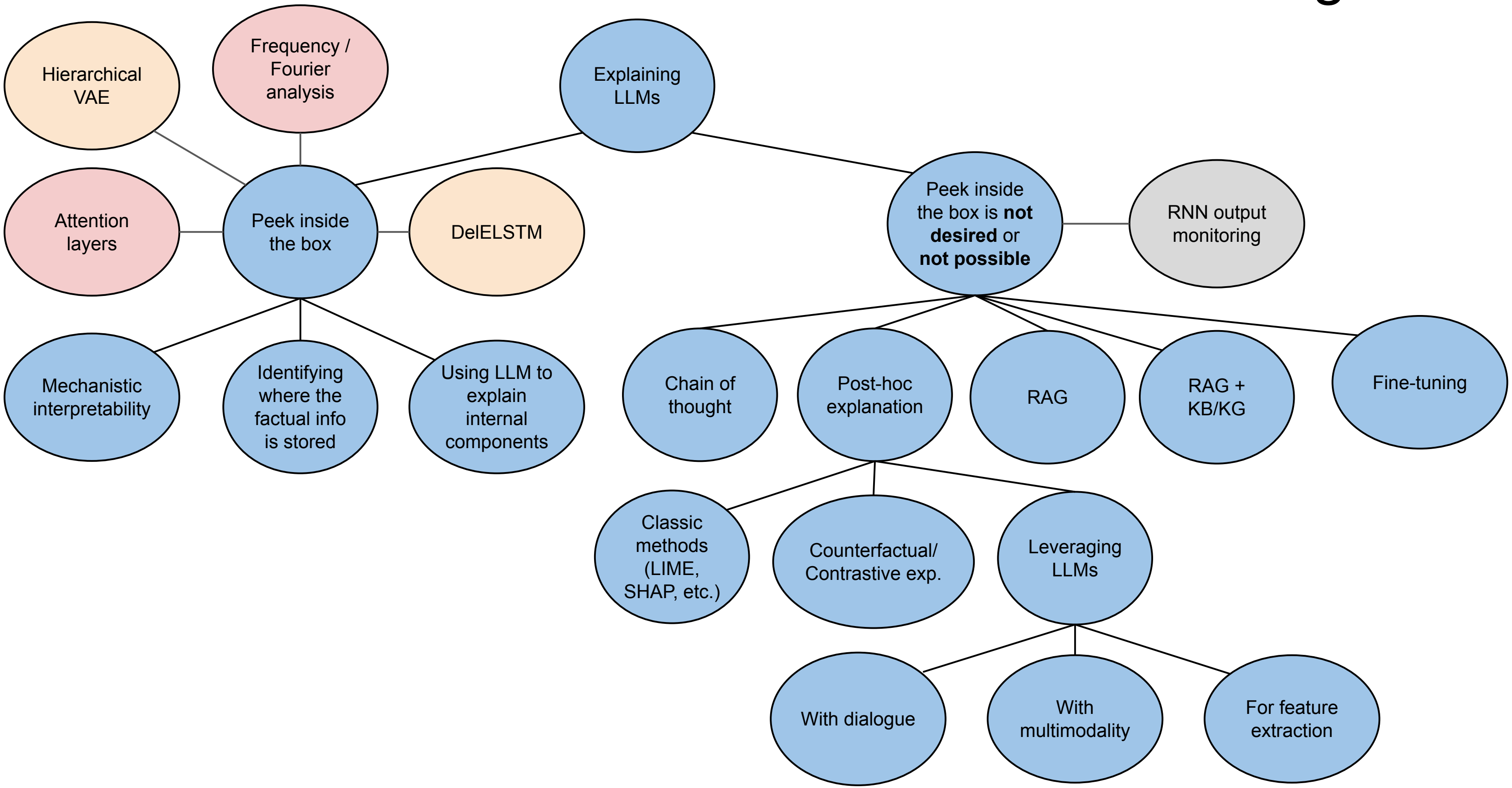
Conclusion of this Tutorial and Take Home Messages



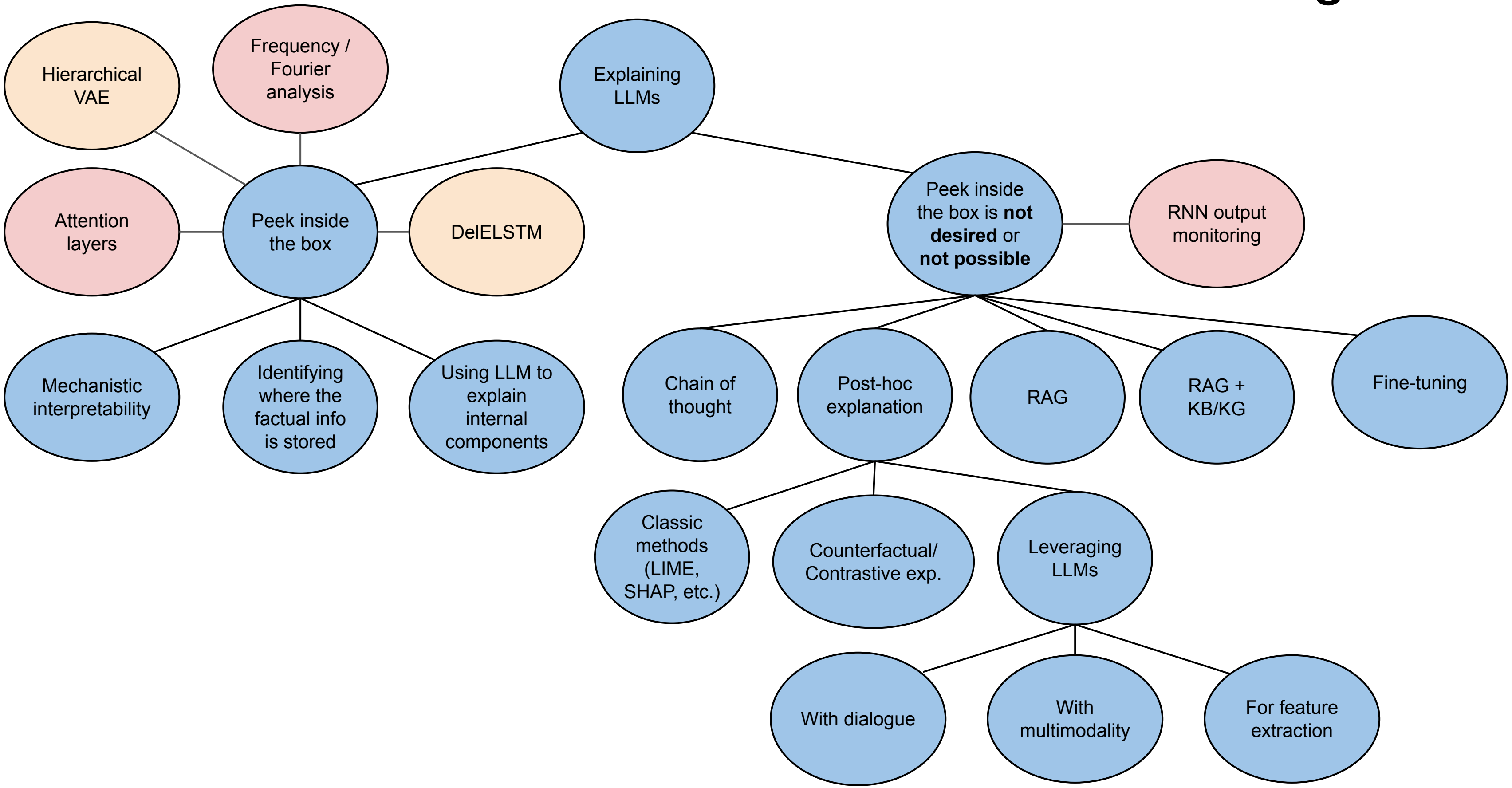
Conclusion of this Workshop and Take Home Messages



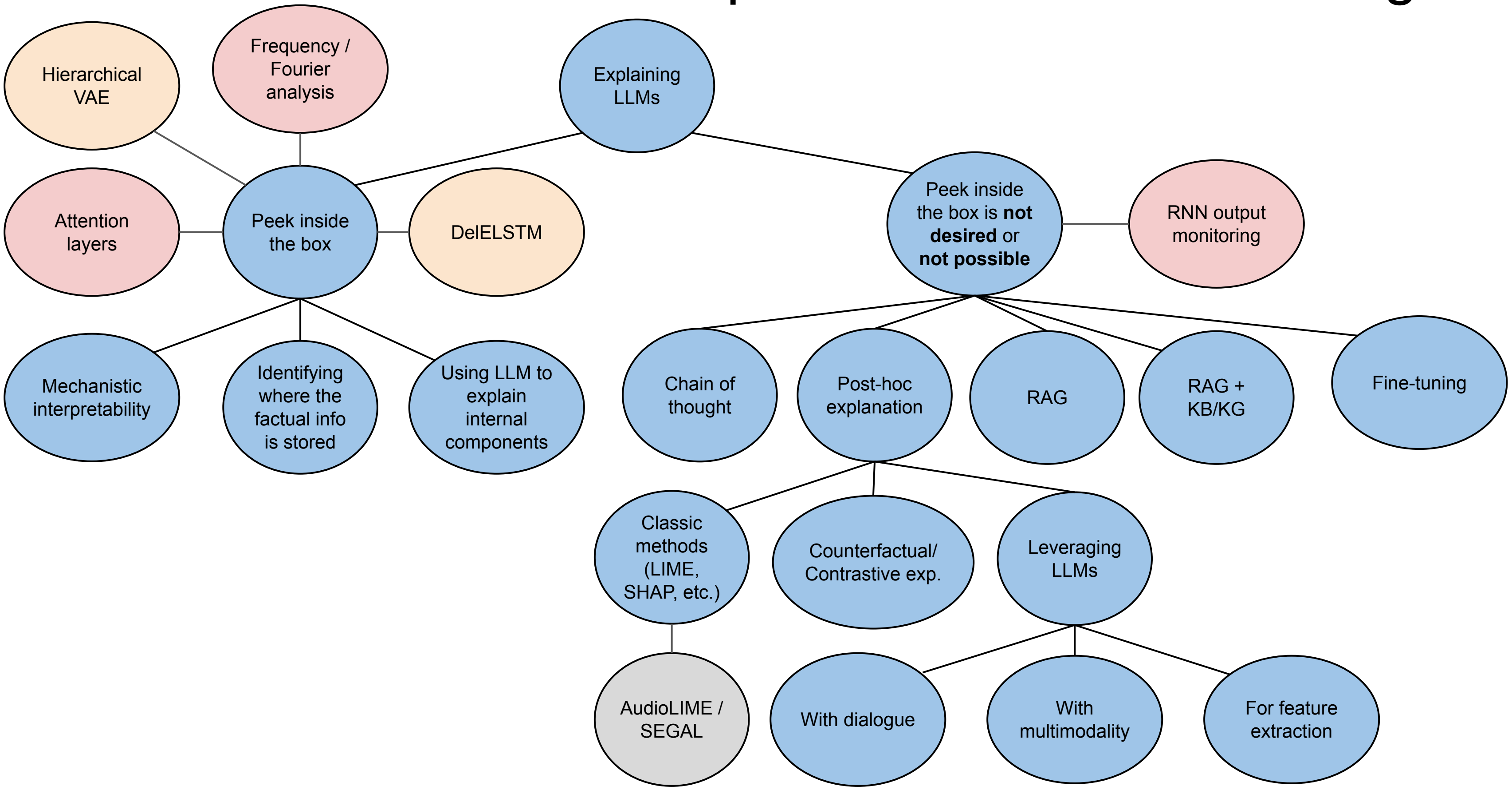
Conclusion of this Tutorial and Take Home Messages



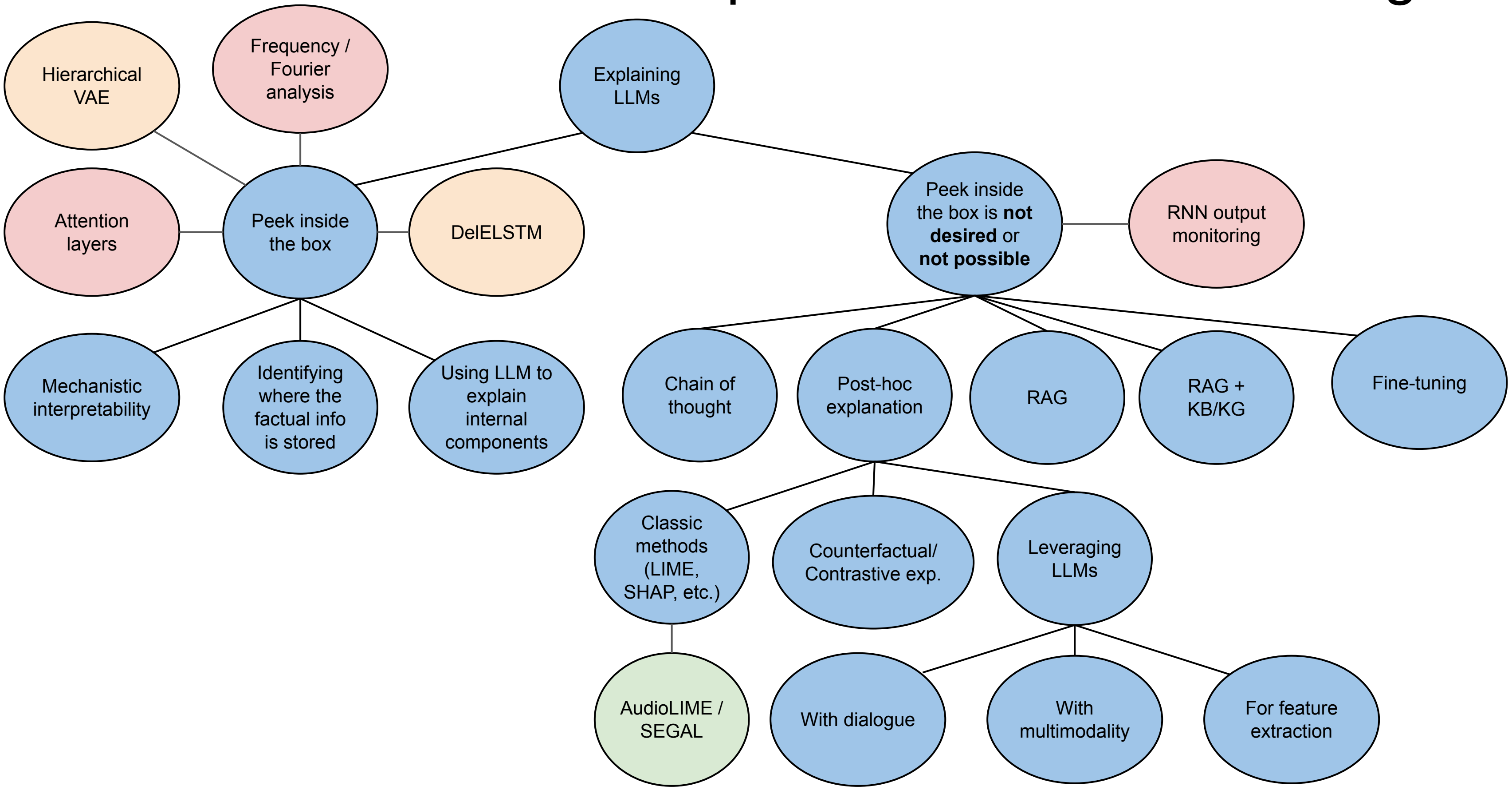
Conclusion of this Tutorial and Take Home Messages



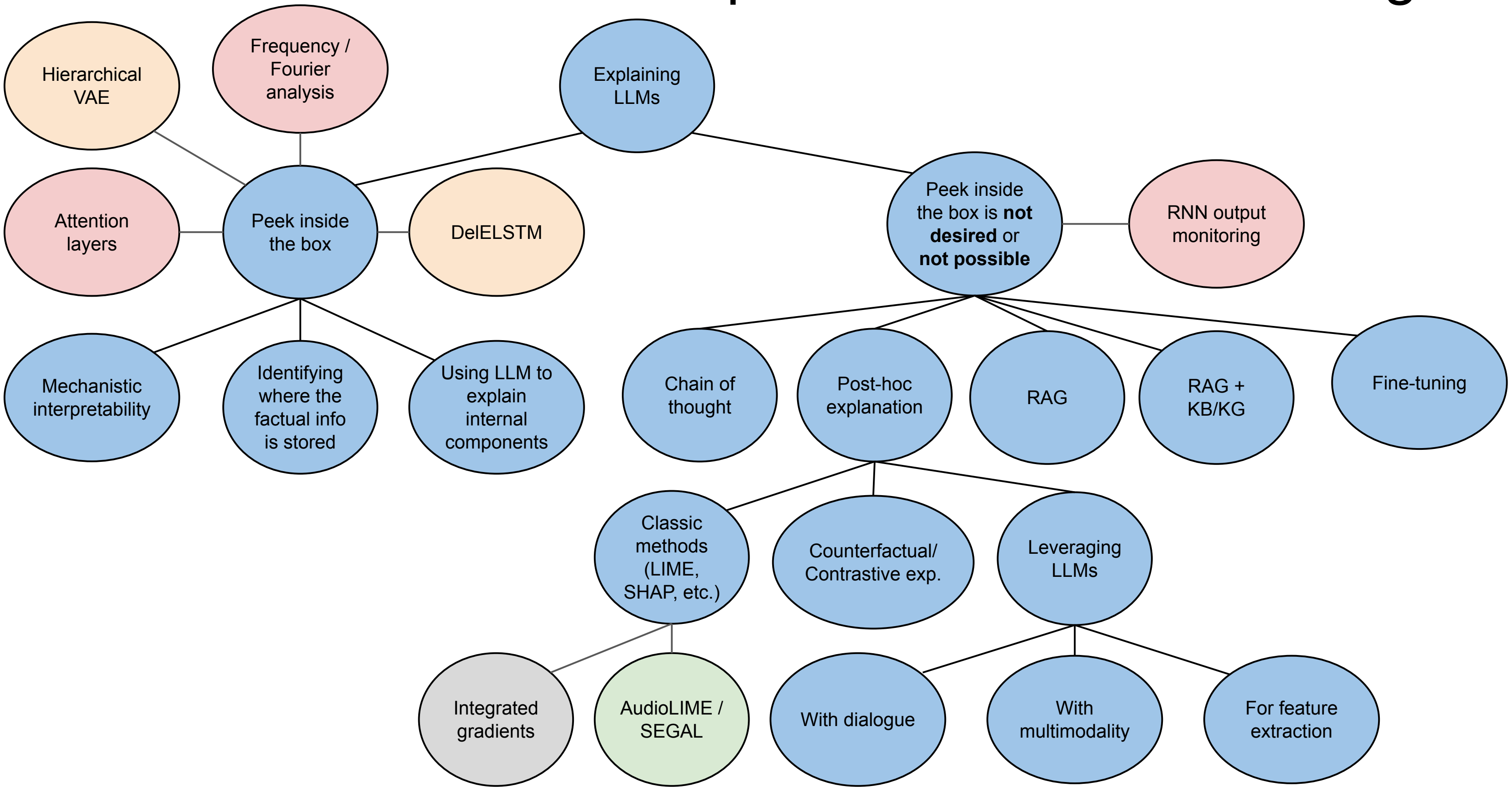
Conclusion of this Workshop and Take Home Messages



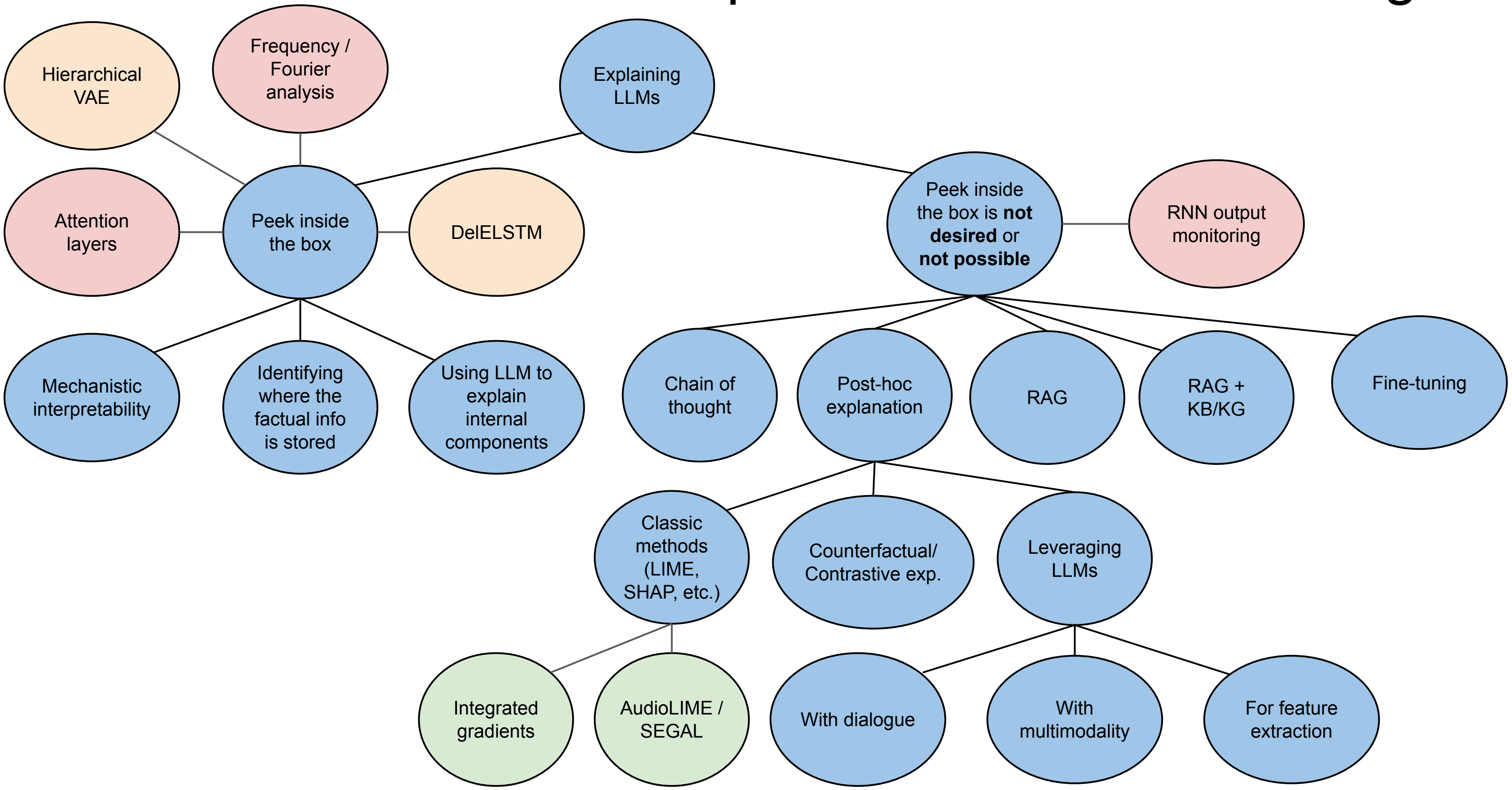
Conclusion of this Workshop and Take Home Messages



Conclusion of this Workshop and Take Home Messages



Conclusion of this Workshop and Take Home Messages



Let's discuss!