Working with Large Language Models

Thiemo Fetzer

University of Warwick & University of Bonn & CEPR & LSE & AEAI

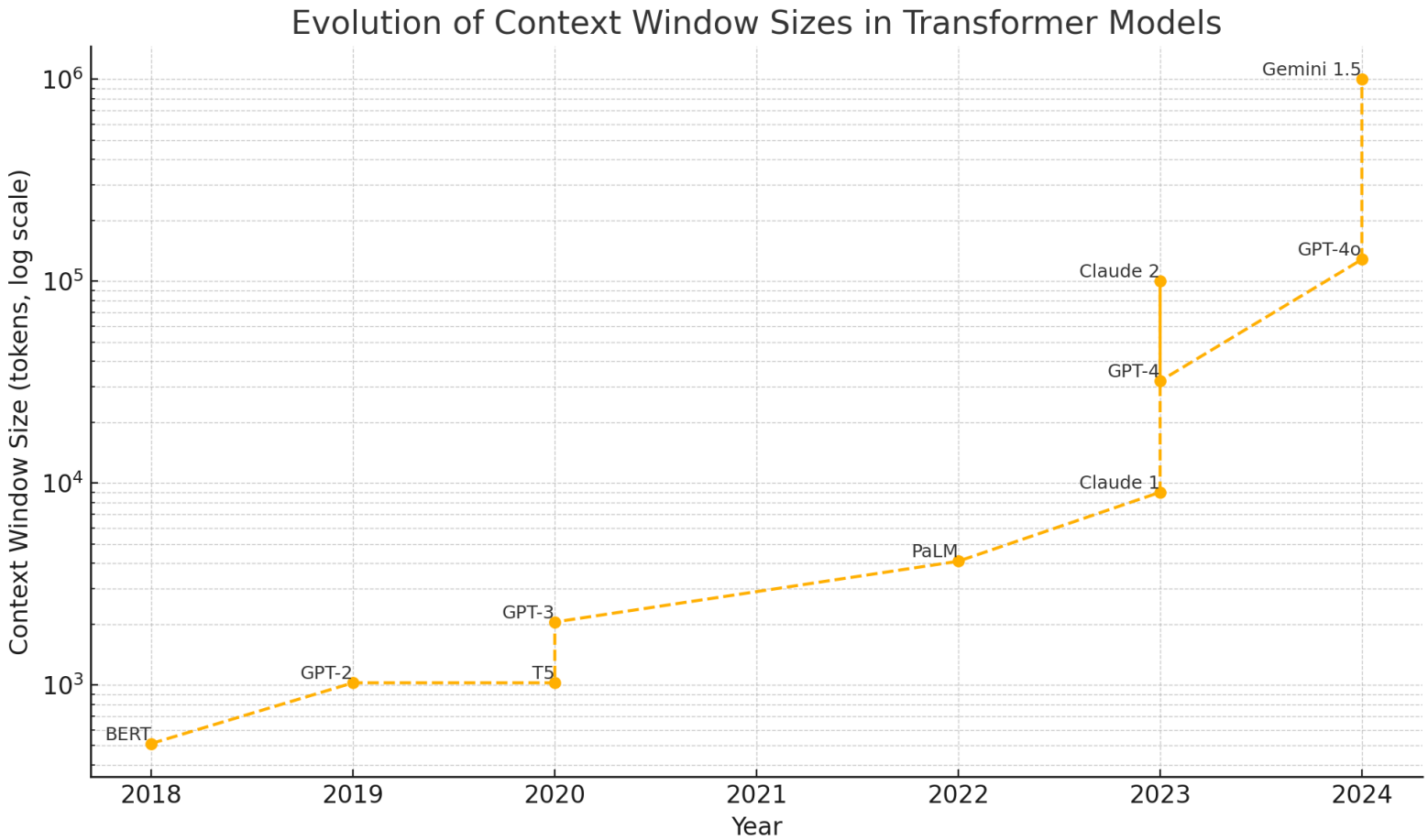
March 11, 2026

Where are we?

- We have learned the basic building blocs and training methods for large language models
- But how should we think of the interplay of the blocks in day to day use
- How do the tools work in what context and for what purpose?
- Cover today some common pipelines or architectures for LLM deployment

$\rightarrow$ evolution of LLMs has followed scaling patterns with ever expanding context windows, increase in number of parameters and development of reasoning capabilities

Context Window Evolution



$\rightarrow$ exponential evolution of context window with frontier models achieving up to 1 million tokens

A human life in context?

Let us go back to the amount of information that a human life generates or can transmit through speech

- We said its around `0.5 MB` per day
- Average human life has several vocabulary expansions, but assume we live 80 years in *speaking mode*

  `365 x 80 x 0.5 MB = 14.6 GB`

  In textual encoding approximately **3.76 billion** tokens are contained in 14 GB of plain text, assuming an average of 4 characters per token.
- Present LLMs are far from being able to replicate a full human as represented in textual output produced through context
- But we have seen massive and non-linear growth in capability

Plan

Working with LLMs

Causal Claims as an LLM Application

Working with embeddings

Developing scalable classifiers

Graph Representations and Neural Networks

Common use cases

- **Data categorization or classification** – applying of *nomenclatures*

  Example: classification of gender or origins of names
- **Information retrieval applications** – retrieving a *needle from a haystack*

  Example: semantic search – memory expansion
- **Dimensionality reduction** – summarization of *large datasets*

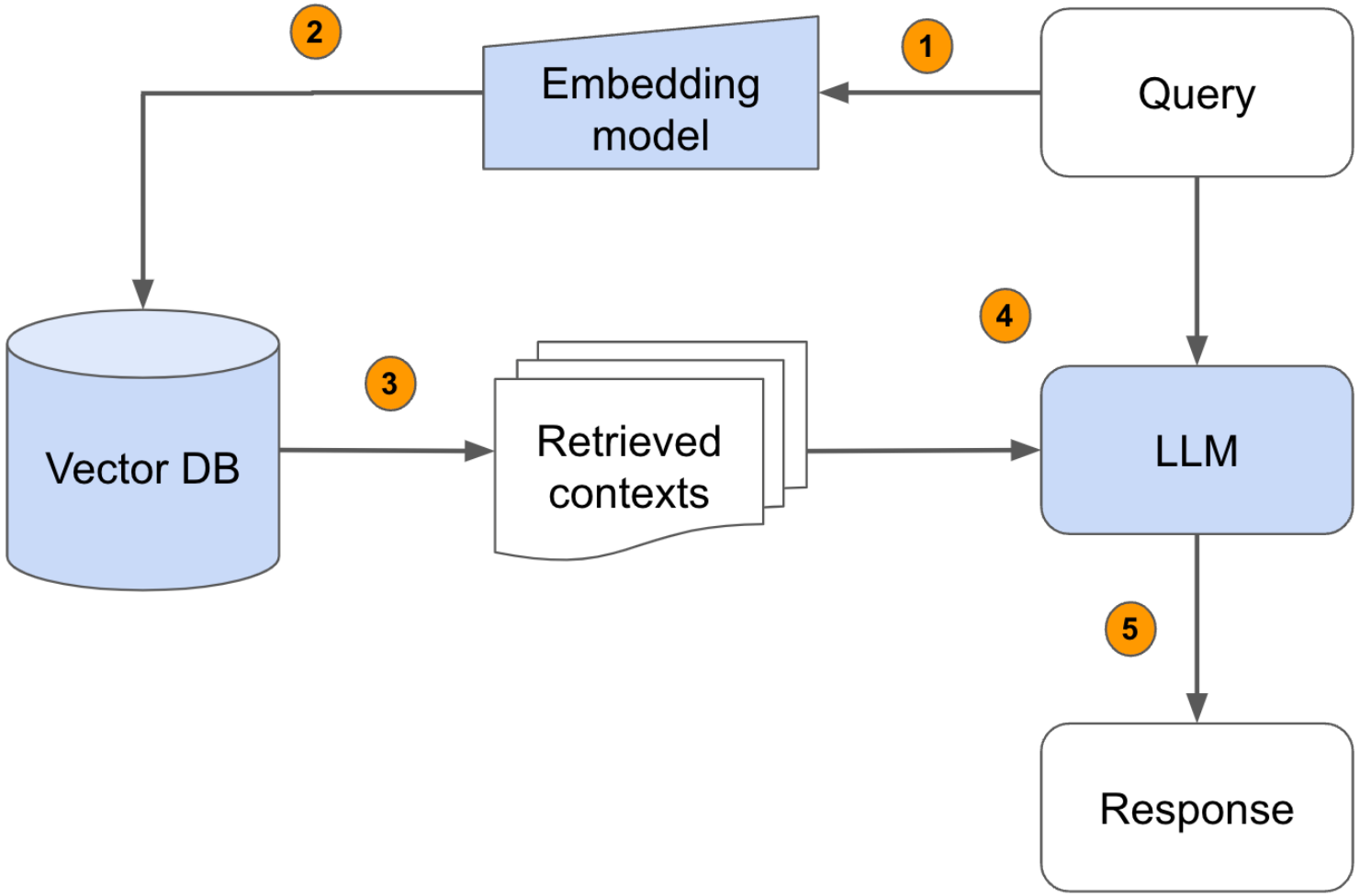
  Example: document summaries
- **Dataset construction** – leaning on stock of embodied knowledge in LLM

  Example: production networks
- **Retrieval augmented dataset construction** – prompt and context into context

  Example: causal claims in economics
- **Reasoning tasks** – complex bipartite matching tasks with incomplete information

  Example: harmonisation of opinion polling datasets

Retrieval augmented generation



Plan

Working with LLMs

Causal Claims as an LLM Application

Working with embeddings

Developing scalable classifiers

Graph Representations and Neural Networks

Case Study: Causal Claims in Economics

Causal Claims in Economics*

Prashant Garg Thiemo Fetzer †

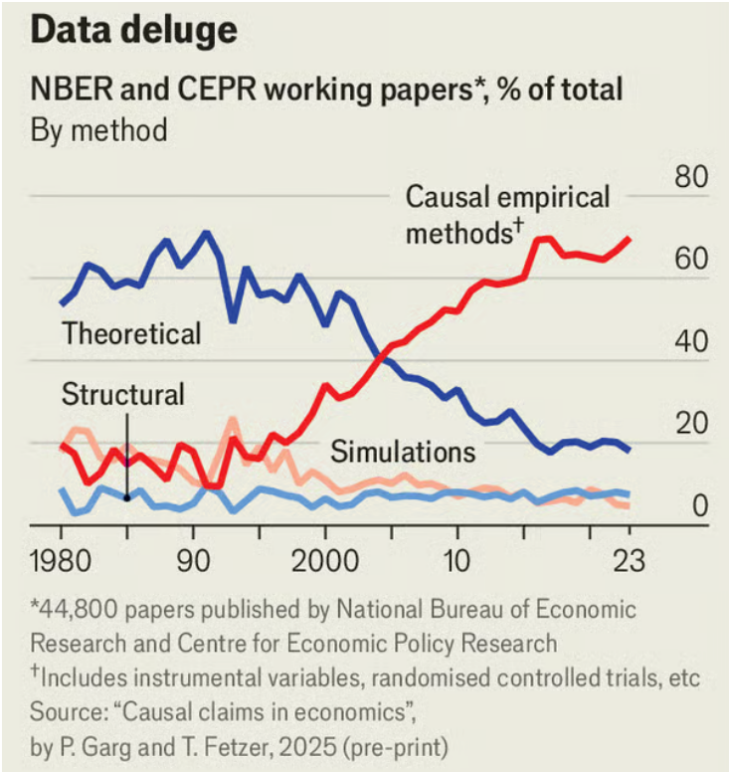
January 14, 2025

Click here for the latest version

Abstract

We analyze over 44,000 NBER and CEPR working papers from 1980–2023 using a custom language model to construct knowledge graphs that map economic concepts and their relationships. We distinguishes between general claims and those documented via causal inference methods (e.g., DiD, IV, RDD, RCTs). We document a substantial rise in the share of causal claims—from roughly 4% in 1990 to nearly 28% in 2020—reflecting the growing influence of the "credibility revolution." We find that causal narrative complexity (e.g., the depth of causal chains) strongly predicts both publication in top-5 journals and higher citation counts, whereas non-causal complexity tends to be uncorrelated or negatively associated with these outcomes. Novelty is also pivotal for top-5 publication, but only when grounded in credible causal methods: introducing genuinely new causal edges or paths markedly increases both the likelihood of acceptance at leading outlets and long-run citations, while non-causal novelty exhibits weak or even negative effects. Papers engaging with central, widely recognized concepts tend to attract more citations, highlighting a divergence between factors driving publication success and long-term academic impact. Finally, bridging under-explored concept pairs is rewarded primarily when grounded in causal methods, yet such gap filling exhibits no consistent link with future citations. Overall, our findings suggest that methodological rigor and causal innovation are key drivers of academic recognition, but sustained impact may require balancing novel contributions with conceptual integration into established economic discourse.

Keywords: Knowledge Graph, Credibility Revolution, Causal Inference, Narrative Complexity, Novelty, Large Language Models
JEL Classification: A10, B41, C18, C80, D83



- Use LLM-supported retrieval and parsing to map causal concepts across 44,000 NBER and CEPR working papers.
- The application is not only summarization: it turns papers into queryable graphs of methods, mechanisms and claims.
- Stylized fact: causal empirical methods become much more prominent after the early 2000s.

Retrieval Links Prose to Evidence Objects

Timing In the main reduced form specification, the news measure enters with a one month lag, n_{hdt-1}. In Figure 2, we explore different leads and lags. This highlights that there is a strong and immediate negative effect which becomes stronger when we add longer lags. This is not surprising given that tourists book travel in advance and will not book travel to a place with negative reporting at the time of booking. Importantly, we also find no pre-trends in card activity before the news events that we study.

(Figure 2)

However, Figure 2 also highlights that a simple reduced form regression analysis as

- Economists often refer to evidence indirectly: "Figure 2", "one month lag", "no pre-trends".
- A RAG system retrieves the paragraph and the referenced object together.
- This lets the model reason about identification checks rather than only surface wording.

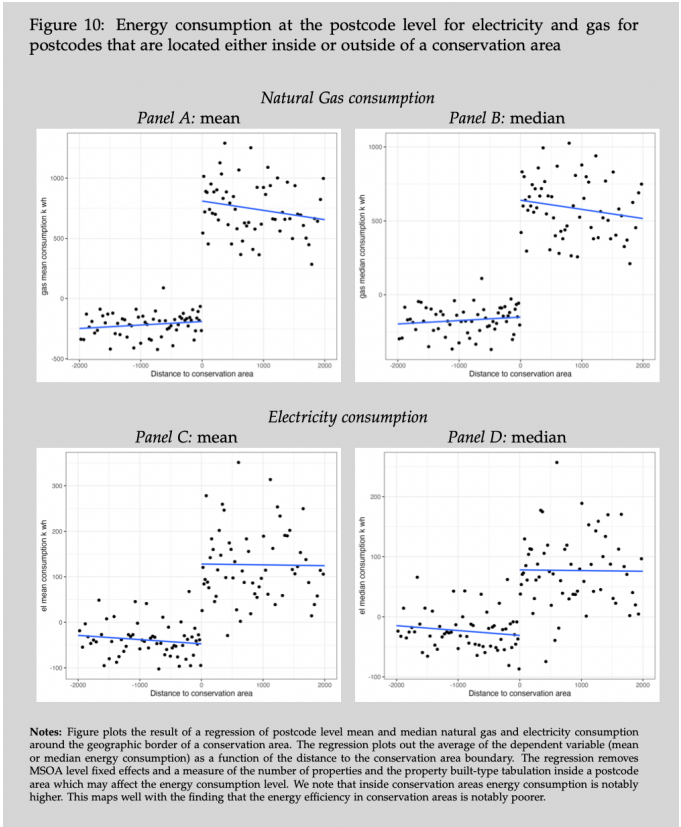
$\rightarrow$ retrieval is the bridge from narrative text to the specific figure, table or appendix object carrying the empirical claim

Figures Become Structured Evidence

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%]
\begin{figure}[h!]
\caption{Energy consumption at the postcode level for electricity and gas for postcodes that are located either inside or outside of a conservation area \label{fig:rdd-design-energyuse}}
\begin{center}
$
\begin{array}{ll}
  \multicolumn{2}{c}{\emph{Natural Gas consumption} } \\

\multicolumn{1}{c}{\text{\emph{Panel A:} mean}} &  \multicolumn{1}{c}{\text{\emph{Panel B:}  median}} \\
\includegraphics[width=.5\columnwidth]{figures/postcode-rdd-plot-gas_mean_consumption_k_wh.png} &
\includegraphics[width=.5\columnwidth]{figures/postcode-rdd-plot-gas_median_consumption_k_wh.png} \\
 \\
  \multicolumn{2}{c}{\emph{Electricity consumption} } \\
\multicolumn{1}{c}{\text{\emph{Panel C:} mean }} &  \multicolumn{1}{c}{\text{\emph{Panel D:}  median}} \\
\includegraphics[width=.5\columnwidth]{figures/postcode-rdd-plot-el_mean_consumption_k_wh.png} &
\includegraphics[width=.5\columnwidth]{figures/postcode-rdd-plot-el_median_consumption_k_wh.png} \\

    \end{array}$
\end{center}
\scriptsize{\textbf{Notes:} Figure plots the result of a regression of postcode level mean and median natural gas and electricity consumption around the geographic border of a conservation area. The regression plots out the average of the dependent variable (mean or median energy consumption) as a function of the distance to the conservation area boundary. The regression removes  MSOA level fixed effects and a measure of the number of properties and the property built-type tabulation inside a postcode area which may affect the energy consumption level. We note that inside conservation areas energy consumption is notably higher. This maps well with the finding that the energy efficiency in conservation areas is notably poorer.}

\end{figure}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

figure caption label is meaningful

Figures have meaningful names and are loaded in presentation grid via latex



- Parse captions, labels, panels and variable names into a machine-readable object.
- Once recovered, figures can be indexed across papers by outcome, design and mechanism.

Tables Become Structured Evidence Too

Similar with tables

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 \begin{table}[h]
\centering
\scalebox{0.75}{
 \begin{threeparttable}
 \caption{Studying energy consumption and the energy efficiency gap \emph{between properties within 500m of a conservation area
border}\label{table:border-rdd-500}}
%
\begin{tabular}{lcccccccc}
 \hline


\addlinespace
      &\multicolumn{1}{c}{(1)}  &\multicolumn{1}{c}{(2)}  &\multicolumn{1}{c}{(3)}  &\multicolumn{1}{c}{(4)}  &\multicolumn{1}{c}{(5)}
&\multicolumn{1}{c}{(6)}  \\
\hline
\addlinespace
\addlinespace

\multicolumn{6}{l}{\emph{Panel A:} EPC Estimated Energy consumption (kWh)} \\
\input{tables/fragment-rdd-energy_consumption_current-distcutoff-500.tex} \\
\multicolumn{4}{l}{\emph{Panel B:}  EPC Estimated Energy efficiency gap (kWh) }\\
\input{tables/fragment-rdd-energy_efficiency_gap-distcutoff-500.tex}\\

\multicolumn{4}{l}{\emph{Panel C:} EPC Estimated CO2 emissions (tonnes CO2) }\\
\input{tables/fragment-rdd-co2_emissions_current-distcutoff-500.tex}\\
\multicolumn{4}{l}{\emph{Panel D:} EPC Estimated CO2 emissions gap (tonnes CO2) }\\
\input{tables/fragment-rdd-emissions_gap-distcutoff-500.tex}\\

\textbf{Regression specification}:  \\
Certificate Year&  X & X & X & X &  X  & X \\
Continuous Property Characteristics &  X & X & X & X &  X  & X \\
Categorical Property Characteristics &  Additive & Interactive & Interactive & Interactive &  Interactive  & Interactive \\
Census Tract  FE &  &  & X &  X  &  X  & X  \\
District x Council Tax Band FE &  &  &  & X  & X  & X  \\
Property Value &  &  &   &  X  & X  \\
Nearest Conservation Area FE &  &  &  &   &   & X  \\
 \hline
 \end{tabular}
```

Table 2: Studying energy consumption and the energy efficiency gap *between properties within 500m of a conservation area border*

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: EPC Estimated Energy consumption (kWh)						
Inside Conservation Area	543.0***	514.5***	681.3***	657.2***	653.2***	654.1***
	(78.90)	(62.86)	(45.84)	(43.44)	(45.17)	(46.07)
Dependent variable mean	22,226.0	22,226.0	22,226.0	22,226.0	23,892.9	23,892.9
R^2	0.69405	0.71744	0.75033	0.75215	0.75134	0.75262
Observations	1,770,660	1,770,660	1,770,660	1,770,660	1,103,509	1,103,509
Panel B: EPC Estimated Energy efficiency gap (kWh)						
Inside Conservation Area	403.6***	346.7***	352.3***	347.7***	424.2***	426.5***
	(34.86)	(32.39)	(30.41)	(29.58)	(38.27)	(38.78)
Dependent variable mean	9,575.4	9,575.4	9,575.4	9,575.4	10,478.7	10,478.7
R^2	0.54017	0.56546	0.60085	0.60279	0.59496	0.59694
Observations	1,770,652	1,770,652	1,770,652	1,770,652	1,103,503	1,103,503
Panel C: EPC Estimated CO2 emissions (tonnes CO2)						
Inside Conservation Area	0.1070***	0.1057***	0.1381***	0.1303***	0.1346***	0.1355***
	(0.0144)	(0.0115)	(0.0088)	(0.0084)	(0.0087)	(0.0089)
Dependent variable mean	4.0111	4.0111	4.0111	4.0111	4.3182	4.3182
R^2	0.69602	0.71952	0.75186	0.75382	0.74889	0.75031
Observations	1,770,660	1,770,660	1,770,660	1,770,660	1,103,509	1,103,509
Panel D: EPC Estimated CO2 emissions gap (tonnes CO2)						
Inside Conservation Area	0.0711***	0.0636***	0.0664***	0.0648***	0.0805***	0.0814***
	(0.0062)	(0.0059)	(0.0055)	(0.0054)	(0.0070)	(0.0071)
Dependent variable mean	1.7160	1.7160	1.7160	1.7160	1.8761	1.8761
R^2	0.54427	0.57166	0.60658	0.60849	0.59893	0.60121
Observations	1,770,660	1,770,660	1,770,660	1,770,660	1,103,509	1,103,509
Regression specification:						
Certificate Year	X	X	X	X	X	X
Continuous Property Characteristics	X	X	X	X	X	X
Categorical Property Characteristics	Additive	Interactive	Interactive	Interactive	Interactive	Interactive
Census Tract FE			X	X	X	X
District x Council Tax Band FE				X	X	X
Property Value					X	X
Nearest Conservation Area FE						X

Notes: Table presents results from a border regression discontinuity design. Each observation refers to a unique property. The dependent variable measured at the property level is indicated in the panel heading from A-D. The regressions control for increasingly more demanding sets of control variables as we move across the columns. Not all control variables are available at the property level for all properties resulting in the sample size to shrink. Continuous property characteristics include the number of habitable rooms and the floor area in square meters. Categorical property characteristics include the property age, the property type, the built form, the tenure and the main heating fuel and whether the property is (in a) listed building. The features are included either additively or in interactive fashion. Census tracts refer to output areas which are statistical geographies with, on average, 100 households. Council tax bands refer to the level of local taxation for local public goods that are levied following tax bands based on property values. The underlying tax burden is specific to the council. Property values refer to the most recent price that was paid for a property as per HMRC's Price Paid dataset. In addition to the log value of the price paid in pounds the regression also controls for the year of the transaction as a fixed effect. Lastly, the fixed effect of the nearest conservation area is included. Standard errors provided in parentheses are clustered at the district level with stars indicating *** p <0.01, ** p <0.05, and * p <0.1.

- Tables reveal outcomes, samples, controls, fixed effects and significance patterns.
- Structured extraction turns a PDF archive into a dataset searchable by identification strategy, variable and result.

Plan

Working with LLMs

Causal Claims as an LLM Application

Working with embeddings

Developing scalable classifiers

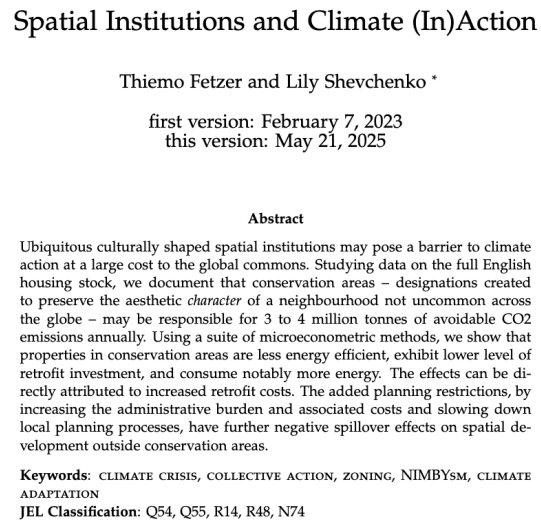
Graph Representations and Neural Networks

Embeddings and Semantic Search

- Embeddings are the workhorse to most RAG applications

  Example: `planning applications` for local councils across the UK
- Expanding work with Lily Shevchenko on *Spatial Institutions and Climate (In)Action*, we are building a large database of planning applications

  Study how individual spatial institutions constrain e.g. housing development



Spatial Institutions and Climate (In)Action

Thiemo Fetzer and Lily Shevchenko *

first version: February 7, 2023
this version: May 21, 2025

Abstract

Ubiquitous culturally shaped spatial institutions may pose a barrier to climate action at a large cost to the global commons. Studying data on the full English housing stock, we document that conservation areas – designations created to preserve the aesthetic *character* of a neighbourhood not uncommon across the globe – may be responsible for 3 to 4 million tonnes of avoidable CO2 emissions annually. Using a suite of microeconometric methods, we show that properties in conservation areas are less energy efficient, exhibit lower level of retrofit investment, and consume notably more energy. The effects can be directly attributed to increased retrofit costs. The added planning restrictions, by increasing the administrative burden and associated costs and slowing down local planning processes, have further negative spillover effects on spatial development outside conservation areas.

Keywords: CLIMATE CRISIS, COLLECTIVE ACTION, ZONING, NIMBYSM, CLIMATE ADAPTATION
JEL Classification: Q54, Q55, R14, R48, N74

Example of planning applications

Planning – Application Summary

Help with this page

PA/19/02266/NC | Proposed installation of 6 no antennas, 1 no 600mm dish, 1 no 300mm dish, 2 no flatpack frames, 1 no power cabinet and 1 no meter cabinet at ground level together with ancillary development. | Lariat Court, 34 Nellie Cressall Way, London E3 4RQ

Save search | Refine search | Track | Print

Details | Comments (1) | Documents (10) | Map | Related Cases (0)

Summary | Further Information | Important Dates

Reference	PA/19/02266/NC
Application Validated	Fri 25 Oct 2019
Address	Lariat Court, 34 Nellie Cressall Way, London E3 4RQ
Proposal	Proposed installation of 6 no antennas, 1 no 600mm dish, 1 no 300mm dish, 2 no flatpack frames, 1 no power cabinet and 1 no meter cabinet at ground level together with ancillary development.
Status	Decided
Decision	Permit
Decision Issued Date	Wed 18 Dec 2019
Appeal Decision	Not Available

put together database of 10 million + planning applications, together with their council discussions and minutes to understand NIMBYism.

Embeddings and semantic search

- A sample of around 9.5 million planning applications each with short description
- Use semantic search and matching to identify *identical* planning applications subject to different *constraints*
- This requires creating creating pairwise comparisons of the application in embedding space
- `9.5 million documents x 1024 embedding dimensionality x 4 bytes for 32-bit float precision = 35 GB`
- Quite often, building embeddings on the fly is not feasible for applications but will get back to this later

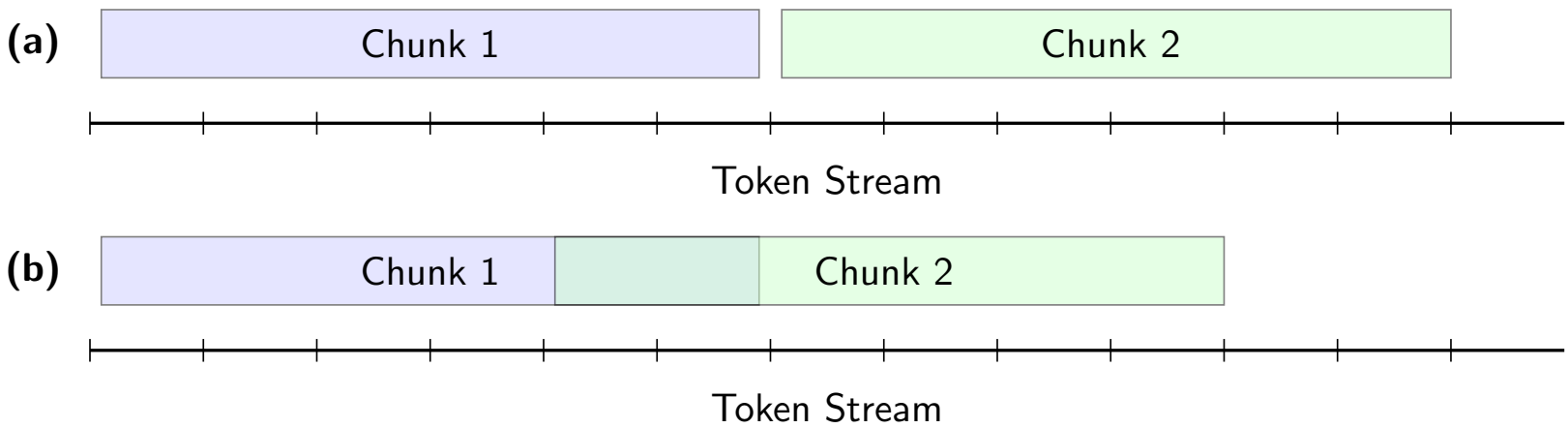
Storage Options for Embeddings

- **Local Storage Options for Embeddings**
  - **Pickle (.pckl)**: easy, native Python, but not optimized for search.
  - **HDF5 / Parquet**: columnar binary formats, efficient for large I/O.
  - **SQLite / DuckDB**: lightweight databases for local embedding storage with metadata.
- **Search-Optimized Vector Databases**
  - **Faiss (Meta)**: fast approximate nearest neighbors, supports flat and quantized indexes (essentially adds a clustering step
  - **Qdrant / Milvus (Rust / C++)**: local server setups, support filtering and persistence.
  - **PgVector**: plugin for PostgreSQL, integrates embedding search into SQL workflows.

Working with Text Embeddings: Chunking Strategy

Given limitations to context windows and often large datasets it still is necessary to work with embeddings. Use **text chunking** to split into manageable sizes.

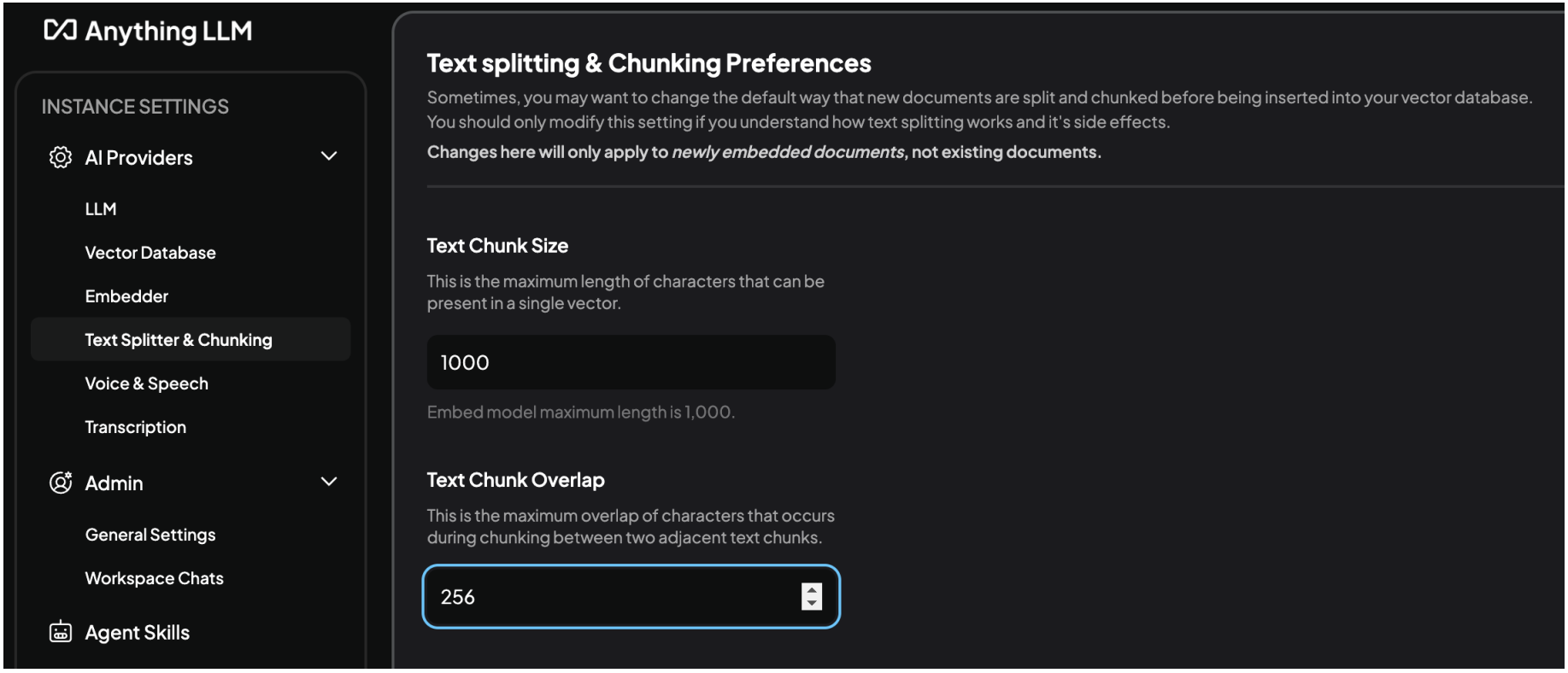
- `chunk_size` (e.g., 300–1000 tokens)
- `chunk_overlap` (e.g., 50–200 tokens)



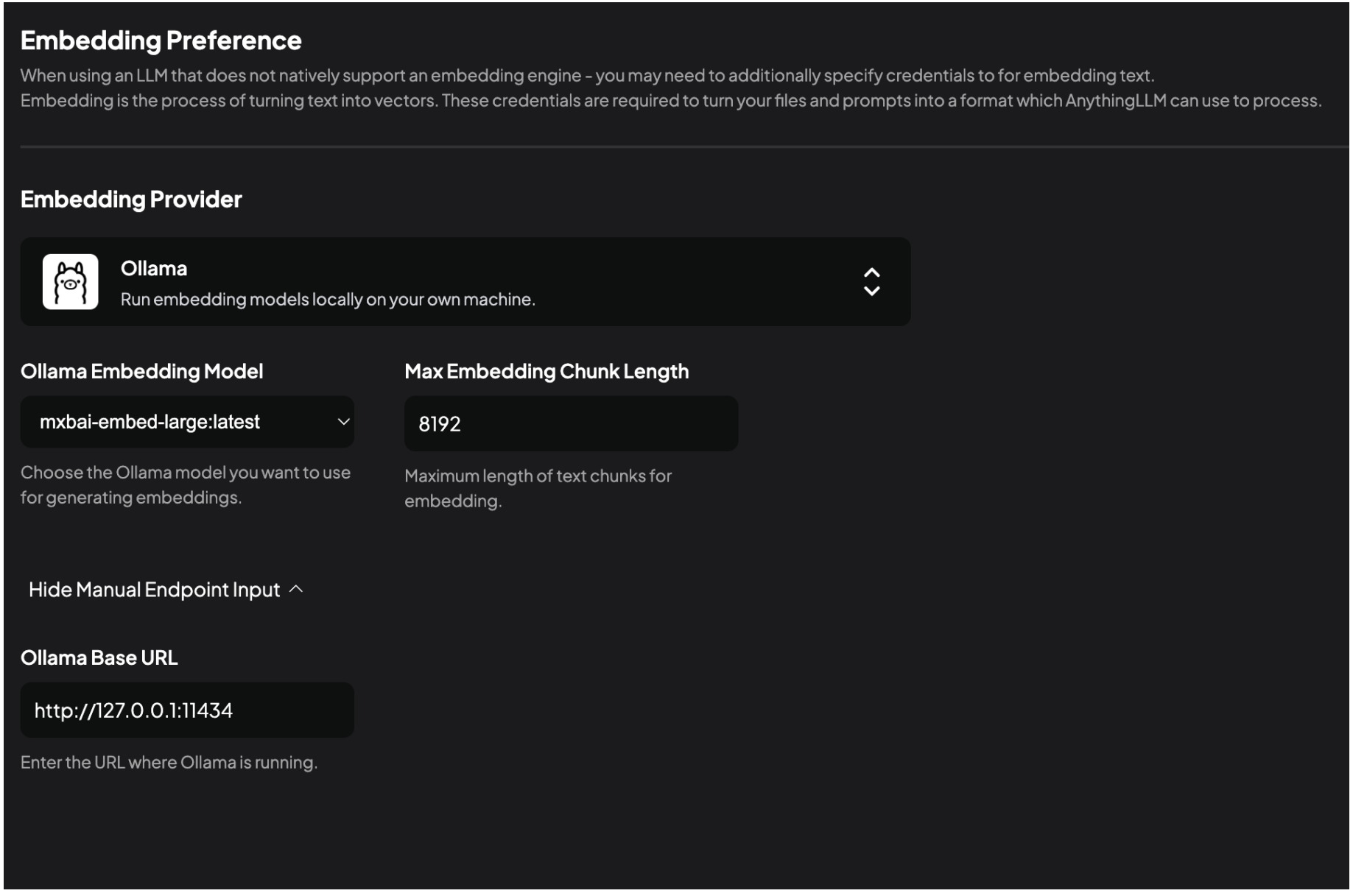
Typical guidance

- **Short documents**: one vector per document.
- **Medium documents (1–2k tokens)**: chunk size 500–800, overlap 100–200.
- **Long documents (more than 2k tokens)**: use a sliding window with overlap and aggregate chunk embeddings.

Settings for chunking & embedding in AnythingLLM



Settings for chunking & embedding in AnythingLLM



Chunking and Embedding using OpenAI API

```
def chunk_text(text, size=512, overlap=100):
    chunks = []
    for i in range(0, len(text), size - overlap):
        chunk = text[i:i+size]
        chunks.append(chunk)
    return chunks

embeddings = [
  openai.Embedding.create(input=chunk, model="text-
     embedding-3-small")
  for chunk in chunk_text(document)
]
```

Estimating Embedding Dataset Size

Formula:

$$\text{Size (bytes)} = N \times D \times B$$

Where:

- N = Number of vectors (e.g. documents or chunks)
- D = Embedding dimension (e.g. 768, 1024)
- B = Bytes per float (4 for float32, 2 for float16, 1 for quantized)

Convert to storage units:

$$\text{Size (MB)} = \frac{N \times D \times B}{1024^2}$$

$$\text{Size (GB)} = \frac{N \times D \times B}{1024^3}$$

Example:

- 9M documents, 1024-dim, float32 $\Rightarrow 9 \times 10^6 \times 1024 \times 4 = 34.3$ GB

Plan

Working with LLMs

Causal Claims as an LLM Application

Working with embeddings

Developing scalable classifiers

Graph Representations and Neural Networks

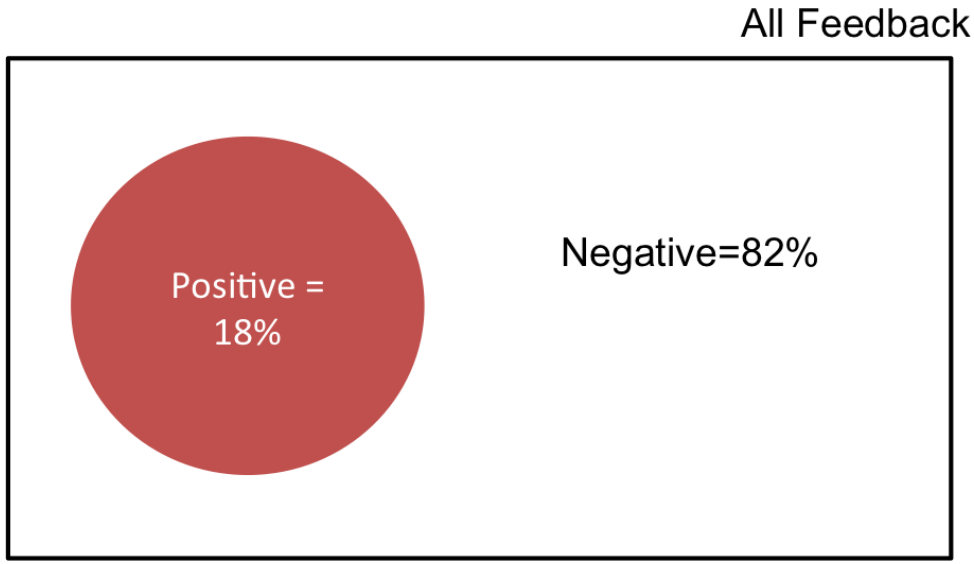
Classification is a most common application

- Working with unstructured data or even just for data filtering may require deployment of classifiers
- Human brain is constantly in process of filtering information or classifying – small kids, object sorting etc
- Building classification systems and nomenclatures enables functional diplomacy
- Traditionally classification was done with a suite of techniques typically distinguishing **generative** versus **discriminative** classifiers
  - Typically dist
  - Naive Bayes classification (we touched on this)
  - Logistic regression or multinomial logistic regression
  - Random forests, trees etc
  - Support Vector machines

$\rightarrow$ LLMs can be used to replace traditional machine-learning approaches or can be used to supercharge

Introduction to classification

- Suppose you want to categorize some *teaching feedback* into two categories.
- In our example we have $Y \in \{\text{Positive}, \text{Negative}\}$.
- Suppose you know what the *priors* are, i.e. you know what the share of positive and negative feedbacks are *in the population*.



Best Classification Decision Without Additional Information

- Suppose you now receive a new feedback $\tilde{Y}$, and you want to categorize it *without any additional information*
- What decision / assignment rule would minimize the missclassification error?

Clearly this should be

$$\tilde{Y} = \begin{cases} + & \text{if } P(+) > P(-) \\ - & \text{if } P(+) < P(-) \end{cases}$$

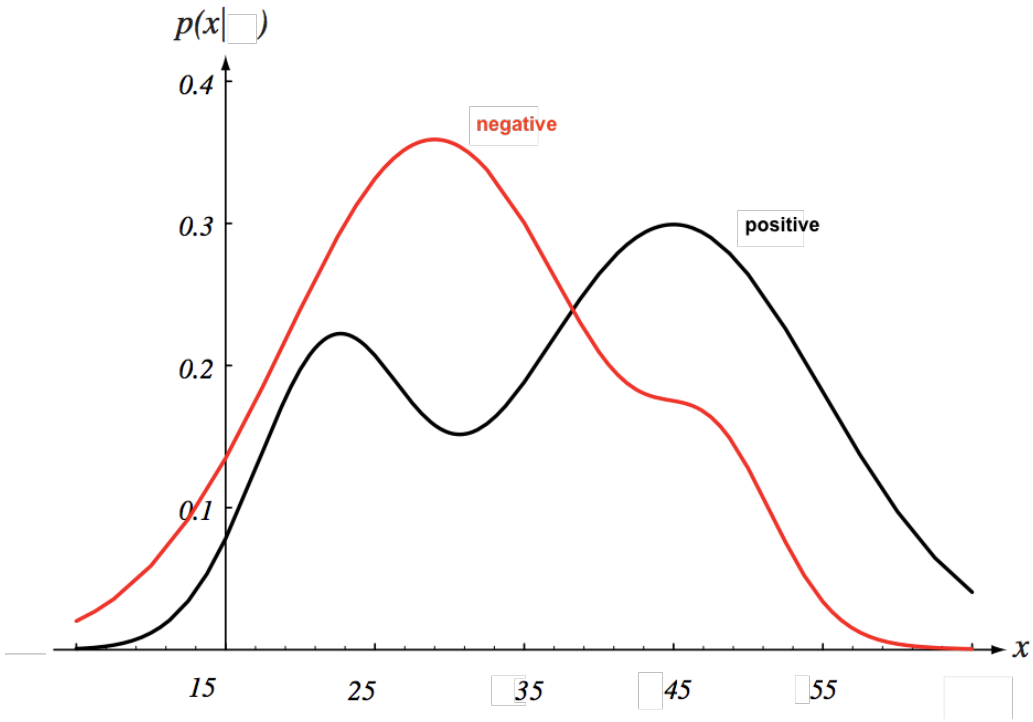
What is the error that we make with this *decision rule*?

$$P(error) = \min\{P(+), P(-)\}$$

So here, we would assign any feedback coming in as negative $-$; the maximum error we can make is 18%.

Best Decision With Additional Information

- Suppose you now have additional information X, let that information be the number of characters contained in a feedback.
- Suppose that the distribution of the number of characters is smooth in the population.
- Plot the density function $p(x|+)$ and $p(x|-)$.



Using Bayes Rule

- For every value of x, we can compute the *posterior* probability using Bayes rule.

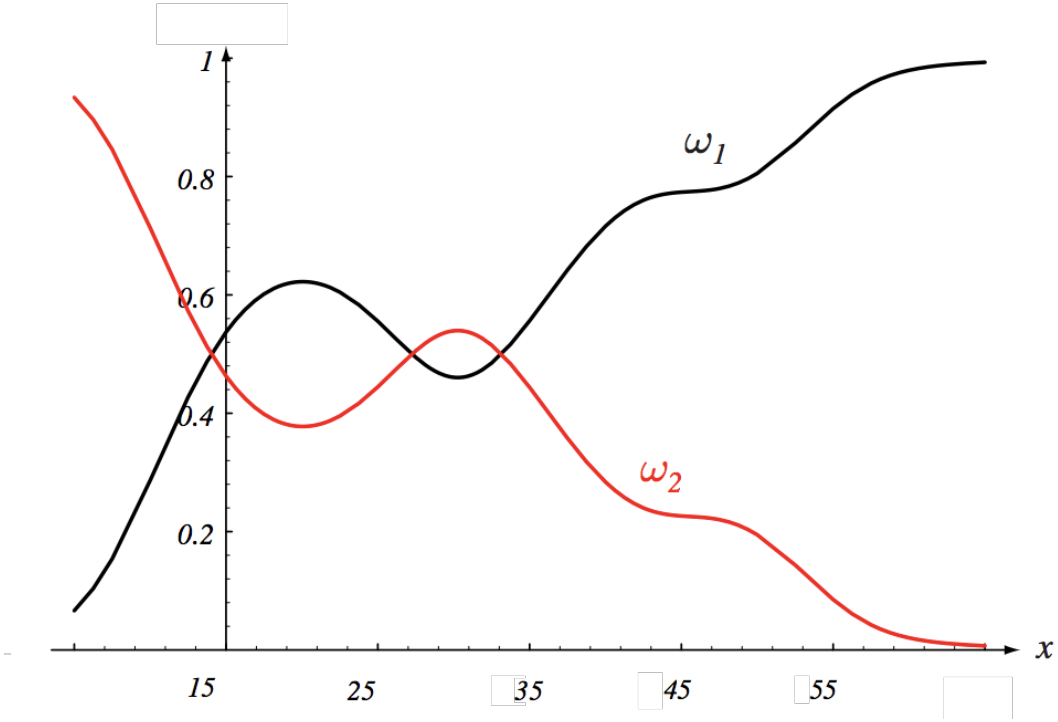
$$\underbrace{P(+|x)}_{posterior} = \frac{\overbrace{p(x|+)}^{likelihood} \times \overbrace{P(+)}^{prior}}{\underbrace{p(x)}_{marginal\ likelihood}}$$

where

$$p(x) = p(x|+)P(+) + p(x|-)P(-)$$

Using Bayes Rule

- Plot the posterior densities $p(+|x)$ and $p(-|x)$.



- How would you now decide on which label to assign? I.e. what is the decision rule that minimzes test error?

MAP Decision Rule

- It seems intuitive, that you should assign

$$\tilde{Y} = \begin{cases} + & \text{if } P(+|x) > P(-|x) \\ - & \text{if } P(+|x) < P(-|x) \end{cases}$$

- This is called the **Maximum A Posteriori** decision rule.

$$P(+|x) > P(-|x)$$

- This boils down to (after plugging in)

$$\frac{p(x|+) \times P(+)}{p(x)} > \frac{p(x|-) \times P(-)}{p(x)}$$

$$\frac{p(x|+)}{p(x|-)} > \frac{P(-)}{P(+)}$$

- Here, we have just two values that Y can take, so the decision rule is

$$P(+|x) > P(-|x) \Rightarrow P(+|x) > 1 - P(+|x) \Rightarrow P(+|x) > 1/2$$

MAP Decision Rule Minimizes Test Error

- It turns out that the MAP decision rule is *Bayes optimal*, it minimizes overall test error. We do not provide a formal proof, but the intuition is given
- What is the error that we make incorporating the information contained in x?

$$P(error|x) = \min\{P(+|x), P(-|x)\}$$

- The total error over all x would be:

$$P(error) = \int_0^\infty \overbrace{p(error, x)}^{\text{joint}} dx = \int_0^\infty P(error|x)p(x)dx$$

- MAP decision rule minimizes this error, because for every value of x, we choose

$$\min\{P(+|x), P(-|x)\}$$

The General Classification Problem

- A qualitative variable takes a value in set $\mathcal{C}$, for example Teaching Feedback $\in \{\text{Positive, Negative, Neutral}\}$, or, binary, Email $\in \{\text{Spam, No Spam}\}$.
- We will denote $|\mathcal{C}| = c$ the number of different categories.
- We still observe a feature vector X and a responsive variable Y that takes on values from the set $\mathcal{C}$
- The MAP Decision rule says that we should assign

$$\hat{Y} = argmax_{y \in \mathcal{C}} \hat{P}(Y = y|X)$$

- As we saw, in the binary case, where $c = 2$, this boils down to assigning an observation to $Y = 1$ if $\hat{P}(Y = 1|X) > .5$

Getting a $\hat{P}(Y = y|X)$

There are different ways to arrive at $\hat{P}(Y = y|X)$.

- In this section we will present three methods for obtaining a $\hat{P}(Y = y|X)$.
- They belong to two distinct classes of estimators
  1. Discriminative – we treat it as an engineering problem
  2. Generative – we have a model of the data generating process (think: a probabilistic language model for text) and then estimate its underlying fundamentals

Classification Example: Land Use Patterns

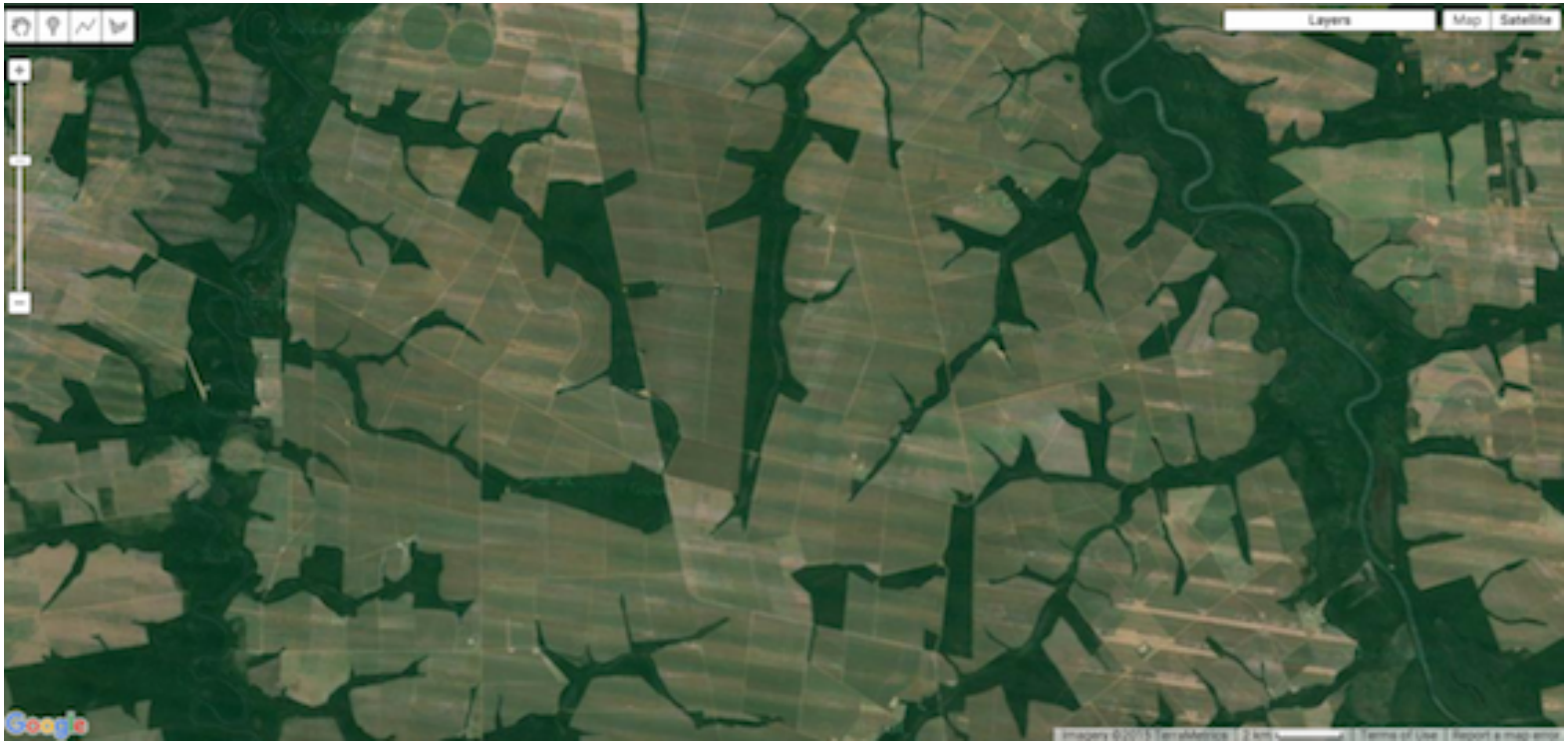


Figure: Classifiying pixels from satellite imagery into common land use clusters: Cropland, Forest and Shrubland as used in Fetzer and Marden (2015).

Classification Example: Land Use Patterns

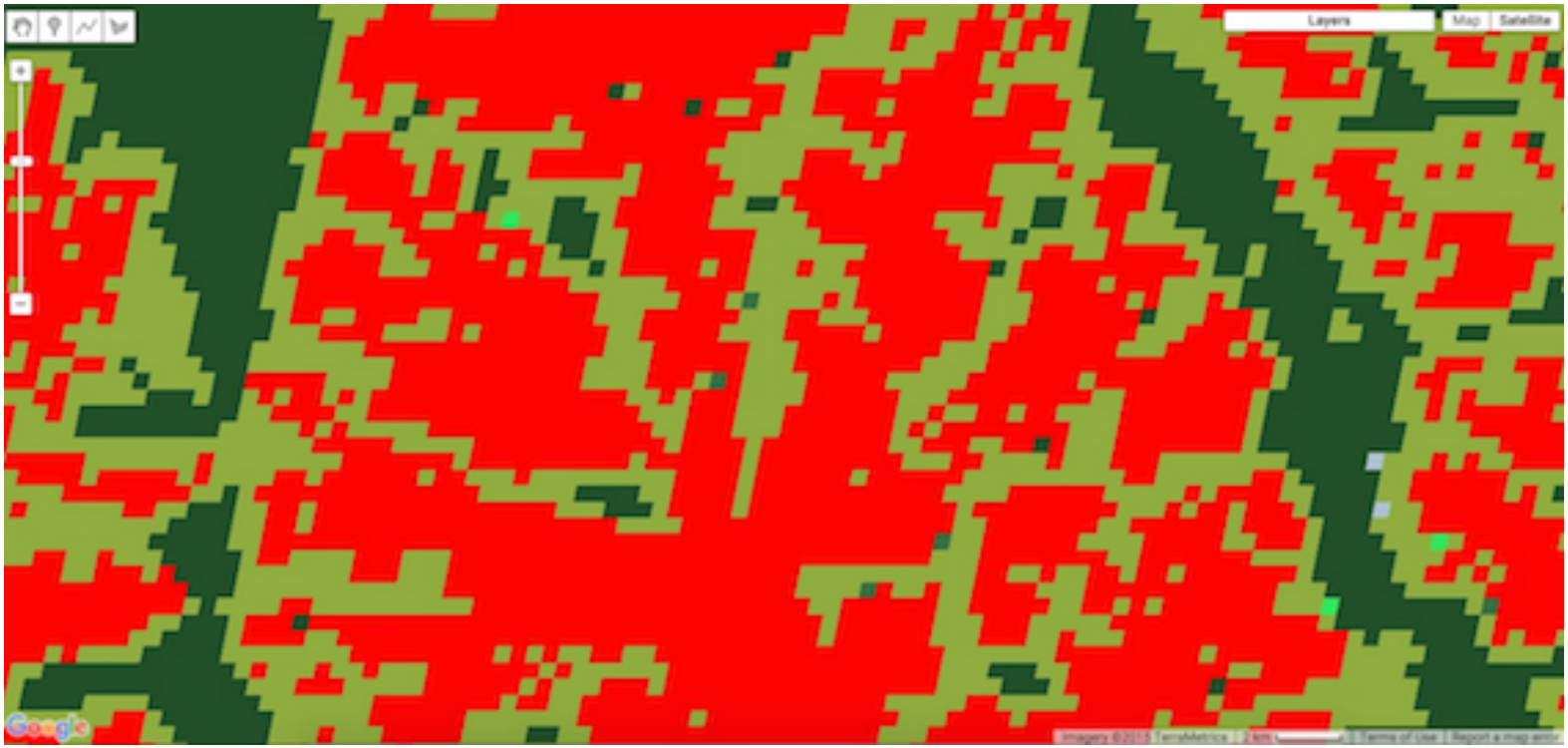


Figure: Classifiying pixels from satellite imagery into common land use clusters: Cropland, Forest and Shrubland as used in Fetzer and Marden (2015).

MAP Decision Rule and Error Types

- MAP minimizes expected classification error.
- Error depends on the threshold $\bar{c}$ used for decision.
- Trade-offs between Type I and Type II errors.

Classifier Evaluation Metrics

- **Confusion matrix**:

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

- **Metrics**:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN},$$

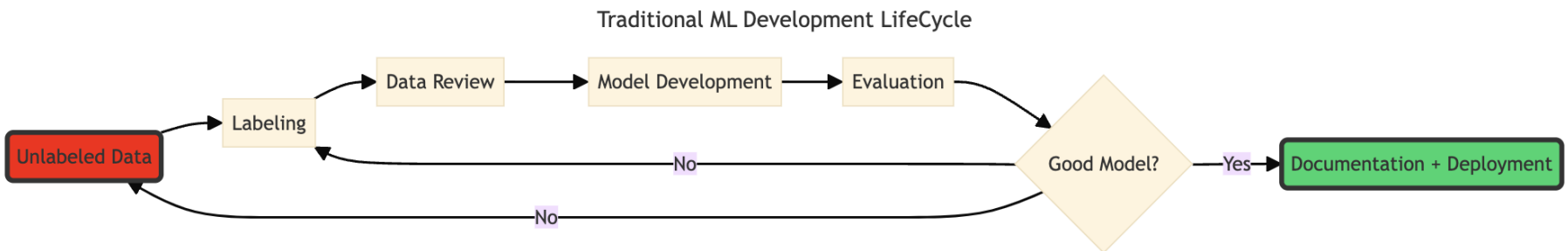
$$\text{Specificity} = \frac{TN}{TN + FP}, \qquad \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision vs Recall: A Trade-off

- Increasing the threshold $\bar{c}$ improves precision but may reduce recall.
- Lowering $\bar{c}$ improves recall but increases false positives.
- Choice depends on application:
  - Spam detection: false positives are costly $\rightarrow$ high precision.
  - Disease screening: false negatives are costly $\rightarrow$ high recall.

Traditional machine learning cycle

Traditional Machine Learning cycle



with many iterative steps going from model learning to evaluation in a circular fashion is overhead heavy and slow.

Image from `https://github.com/kenhktsui/anyclassifier`

With LLMs we can jump this process

- Use well curated **prompt** you can instruct an LLM to perform a classification task
- Prompt may include **examples** in the context window
- Typically main trade-off is over what and how an LLM is being used

  Research use: often times limited in focus to inference

  Production use: focus on prediction and real time performance

$\rightarrow$ lets look at some examples

Model Parameters Parameters and Model Behavior

The following parameters shape how the transformer decodes output during inference:

- **Response format**: If specified (e.g., JSON schema), the model conditions output structure – affects token probability masking.
- **Temperature**: Controls softmax sharpness
- **Top P** : Nucleus sampling truncates token distribution to smallest set whose cumulative probability $\geq p$. Helps limit long tail of unlikely completions.
- **Frequency penalty** : Penalizes tokens already generated — implemented via logit bias before softmax.
- **Presence penalty** : Penalizes any prior appearance of a token — encourages introducing new topics.

Where Temperature and Sampling Fit in Generation

Inference Pipeline Steps:

1. Transformer outputs logits $z_t(w)$ for all tokens.
2. Logits are adjusted:
   - Apply temperature: $z_t(w)/T$
   - Apply presence/frequency penalties: specific $z_t(w)$ adjusted
3. Apply softmax:

$$P(w_t \mid w_{<t}) = \frac{\exp(z_t(w)/T)}{\sum_{w'} \exp(z_t(w')/T)}$$

4. Apply Top-P Sampling:
   - Sort tokens by $P(w)$
   - Select smallest set where cumulative $\sum P(w) \geq p$
   - Sample from this truncated set

Note: This sequence governs how language is sampled at each step of generation.

Temperature and the Softmax Distribution

In a text generation task, we model:

$$P(w_t \mid w_1, w_2, \ldots, w_{t-1}) = \frac{\exp\left(\frac{z_t(w_t)}{T}\right)}{\sum_{w'} \exp\left(\frac{z_t(w')}{T}\right)}$$

- w_t: next token to be predicted
- $z_t(w)$: unnormalized logit for token w at position t

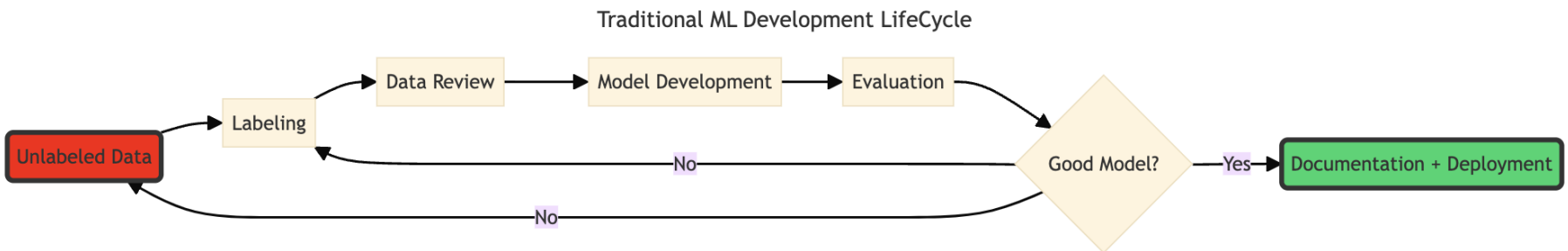
Temperature inflates or deflates the transformer derived scores

- $T = 1$: standard softmax
- $T < 1$: sharper distribution - high-probability tokens become even more likely (low entropy)
- $T > 1$: flatter distribution - encourages exploration (higher entropy)

Temperature scaling is applied *after* logits are produced by the transformer but *before* sampling of words.

Traditional machine learning cycle

Traditional Machine Learning cycle



with many iterative steps going from model learning to evaluation in a circular fashion is overhead heavy and slow.

Image from `https://github.com/kenhktsui/anyclassifier`

Problems with LLM approach

- Some data sources extremely high volume high frequency data

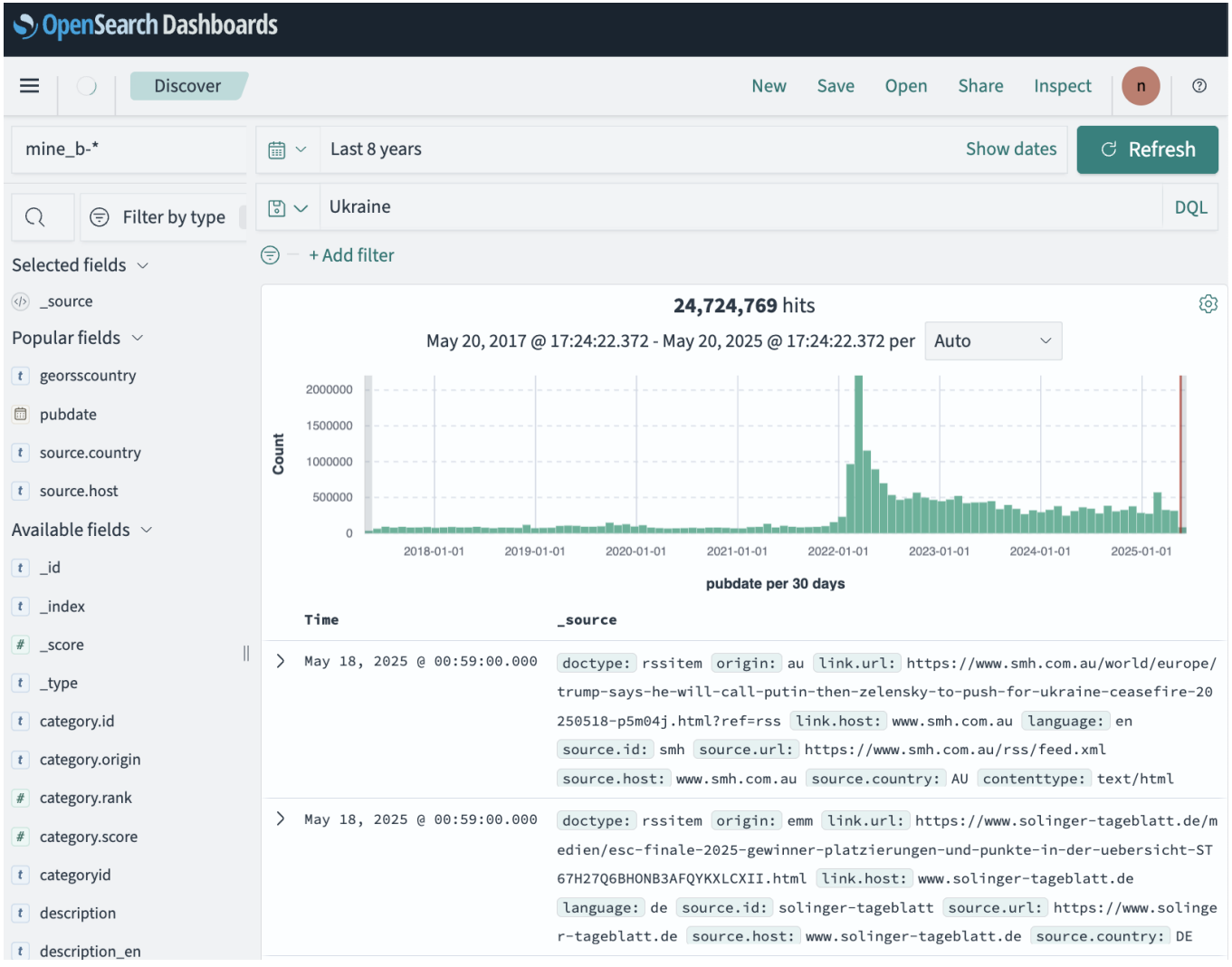
  Think: social media, online streaming platforms etc. or even some text
- If every human were to produce `0.5 MB` of spoken text a day this would still be

  8 billion humans producing text all day amounts to `8 billion x 0.5 MB = 3,204,346 GB`
- Most (signal) intelligence agencies engage in *social listening*

  Need MUCH faster processing of information and quick setup of social listening pipelines
- This is why there is a battle over control of two dimensions

  `digital identity` + `digital infrastructure`

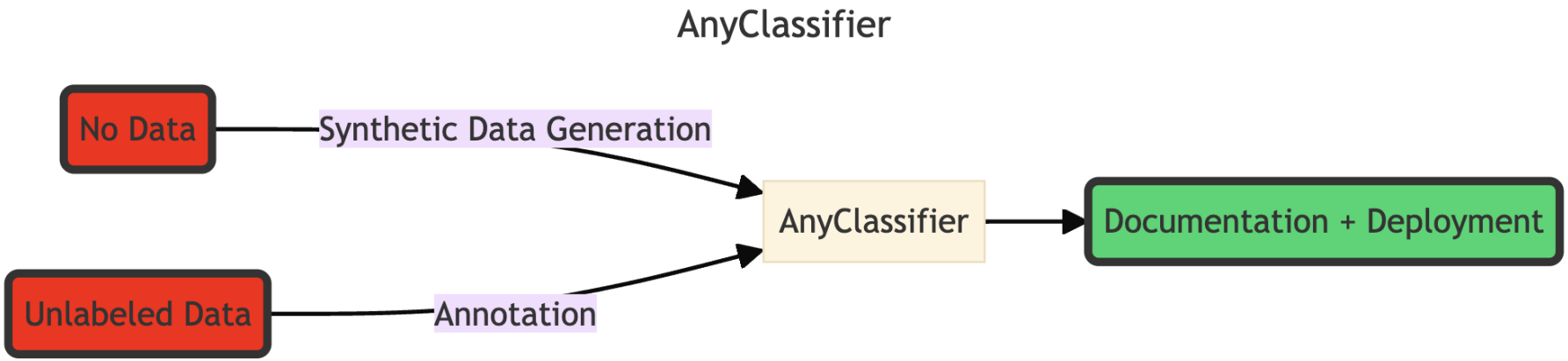
Challenges around social listening



→ and quickly changing topic space

How can we combine quality of LLMs with efficiency of bespoke models?

Enter solutions like `anyclassifier`



with many iterative steps going from model learning to evaluation in a circular fashion is overhead heavy and slow.

Model Options in `anyclassifier`

SetFit (default):

- Uses Sentence Transformers to embed text
- Trains a logistic regression or linear classifier on top
- Fast, efficient, scalable for large datasets
- Avoids full fine-tuning $\Rightarrow$ much faster than traditional transformers

FastText:

- Shallow classifier using n-gram features and embeddings
- Developed by Facebook; scales to millions of examples
- Closest to classical ML (e.g., SVM) in spirit and performance

Transformers (optional):

- Fully fine-tuned models (e.g., BERT)
- More accurate, but slower and computationally heavy

What's Automated?

- **LLM-driven labeling:** High-quality pseudo-labels from LLMs
- **Few-shot learning:** Improve label quality with example inputs
- **Synthetic data generation:** LLMs can generate labeled training samples

Plan

Working with LLMs

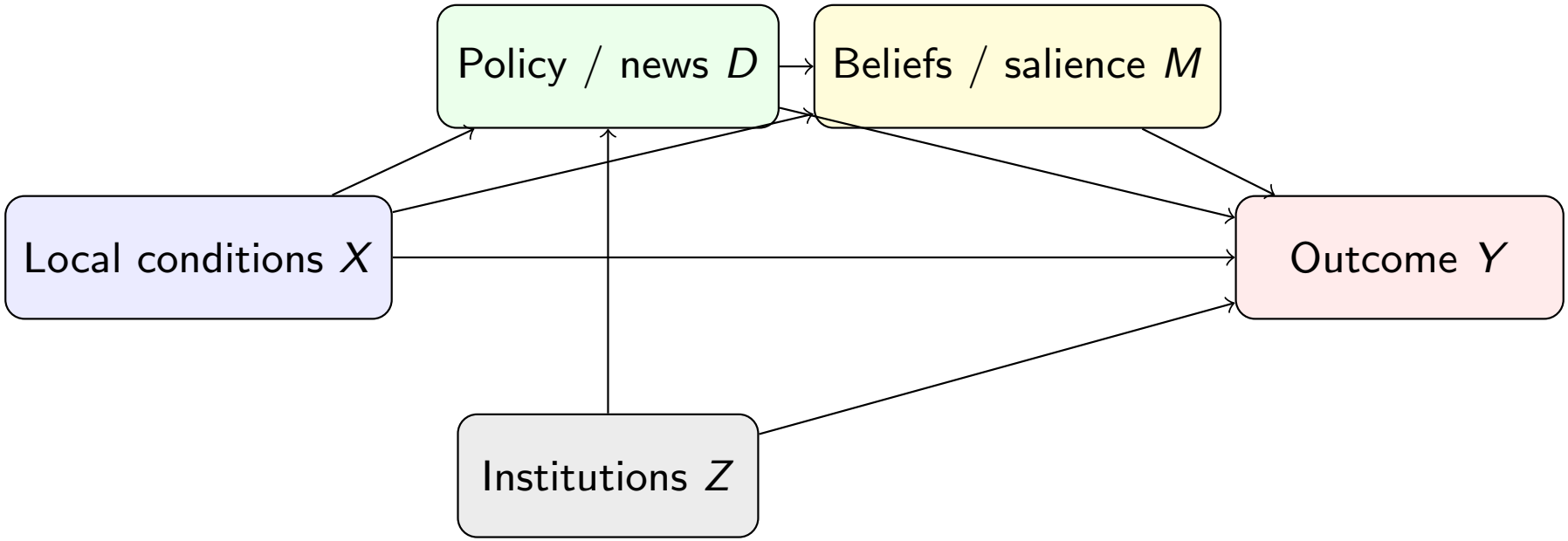
Causal Claims as an LLM Application

Working with embeddings

Developing scalable classifiers

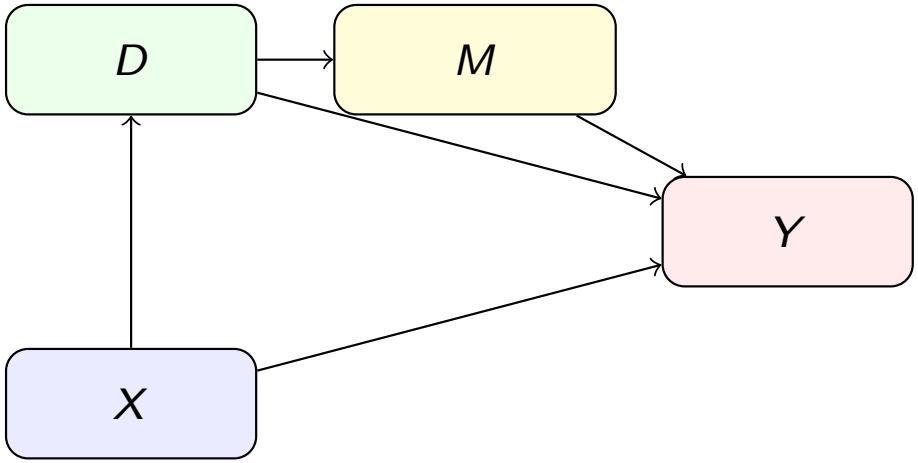
Graph Representations and Neural Networks

Economists Often Start from a DAG



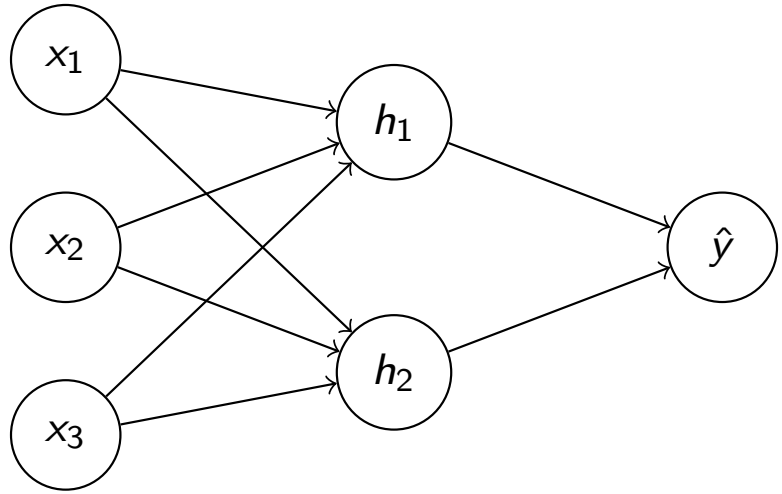
- A DAG encodes an economic mechanism: treatment, mediator, outcome and the backdoor paths that must be addressed.
- This is how economists write down identification assumptions before estimating anything.
- The node labels are substantive rather than generic: policy, beliefs, institutions and outcomes.

DAGs and Neural Nets Use the Same Visual Grammar



Economist's DAG

Edges carry a causal interpretation

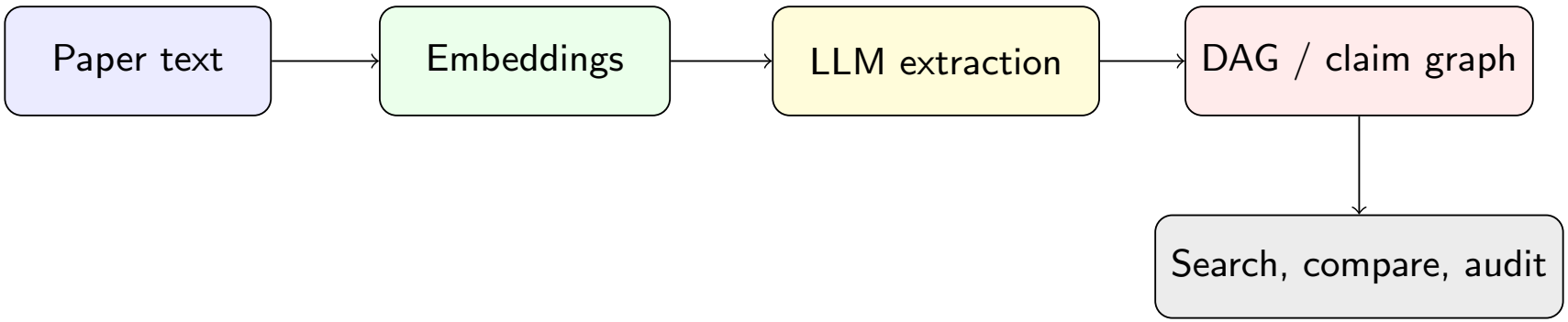


Neural network

Edges carry learned weights

- Both are directed graphs, but their semantics differ.
- In a DAG we specify assumptions; in a neural net we learn a representation that optimizes prediction or classification.
- For LLM applications, graph structure gives interpretability while neural nets give flexible function approximation.

Combining Economic DAGs with Neural Representations



- Neural nets supply the latent representations that make retrieval, matching and extraction possible.
- Economists still need a mechanism language: treatments, mediators, confounders, outcomes and identification checks.
- This combined workflow is exactly what the causal-claims application is trying to operationalize.