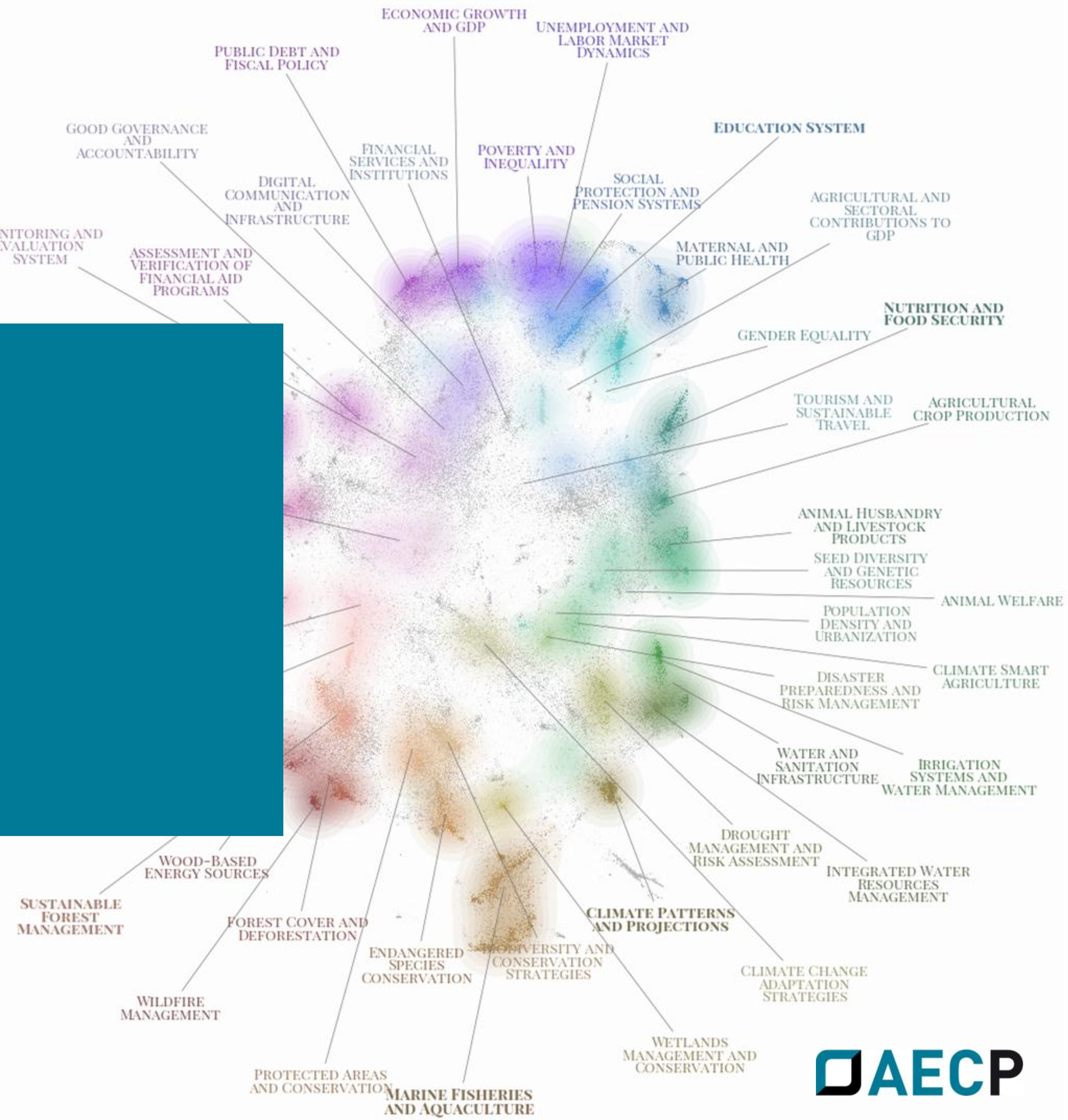
ETH zürich

# AI for Text Analysis in Agricultural Economics

**Christian Stetter**
Agricultural Economics and Policy Group, ETH Zurich

EAAE Masterclass | Webinar | 2025-10-30




AECP

PhD in AgEcon from TU Munich, Germany

Postdoctoral researcher at Agricultural Economics & Policy Group, ETH Zurich

Research interests:

- Agricultural production in the context of environmental change
- Use of AI and ML to enhance sustainable and resilient agriculture.

cstetter@ethz.ch

linkedin.com/in/christian-stetter-135958159/

https://scholar.google.de/citations?user=EOW7rSAAAAAJ&hl=en

5+ years experience in ML/AI

2+ years experience in AI for text analysis

# Learning Objectives

By the end of this session, you will hopefully:

- **Roughly capture the intuition** behind how AI and NLP models analyze text data.
- **Get familiar** with key tools and platforms for text analysis (e.g., Hugging Face, LLMs).
- **See demonstrations** of practical text analysis workflows.
- **Be inspired** to explore and apply these methods in your own AgEcon research.

# Outline

1. Motivation: Text as Data
2. What is Natural Language Processing?
3. Short Intro to Modern AI-based NLP
4. Models, Tools & Applications
   - Encoder-only models
   - Hands-on Sentiment Analysis Demo in Python

   - Decoder-only models (LLMs)
   - Hands-on Classification Demo in Python
5. Discussion & Conclusion

# Warm-up: Your Voice on AI in AgEcon

1. **Connect:**
   Scan the **QR Code** or go to the **link in the chat**

2. **Share:**
   Submit **one complete sentence in English** reflecting your **hope** or **concern** about the role of AI in agricultural economics.

3. **Submit:**
   Your responses will be used later for an in-class text analysis demo.

# Disclaimer before we begin

This session approaches text analysis from a **computational social science** perspective. While qualitative traditions have long engaged with text, our focus today is on **scalable, quantitative, and algorithmic methods** for extracting insights from large text corpora. We will not cover manual coding or interpretive frameworks typical of qualitative research. Instead, we'll explore how tools from machine learning and natural language processing can help agricultural economists analyze text as data.

**Things we do not cover:**

- How to build chatbots.
- How to use Ai to write your papers.
- How to use AI to help you code etc.
- Use LLMs to generate data.
- Finetuning.
- ...

*"For social scientists, the information encoded in text is a rich complement to the more structured kinds of data traditionally used in research."*

(Gentzkow et al., 2019, p.535)[6]

# Text as Data

# Our Traditional Data Comfort Zone…

**FA(/S)DN**

**Closed-form surveys**

**(Choice) Experiments**

**Statistical Offices (EuroSTAT, FAOStat, national statistical Bureaus)**

**Climate & Weather Data**

**Environmental data**

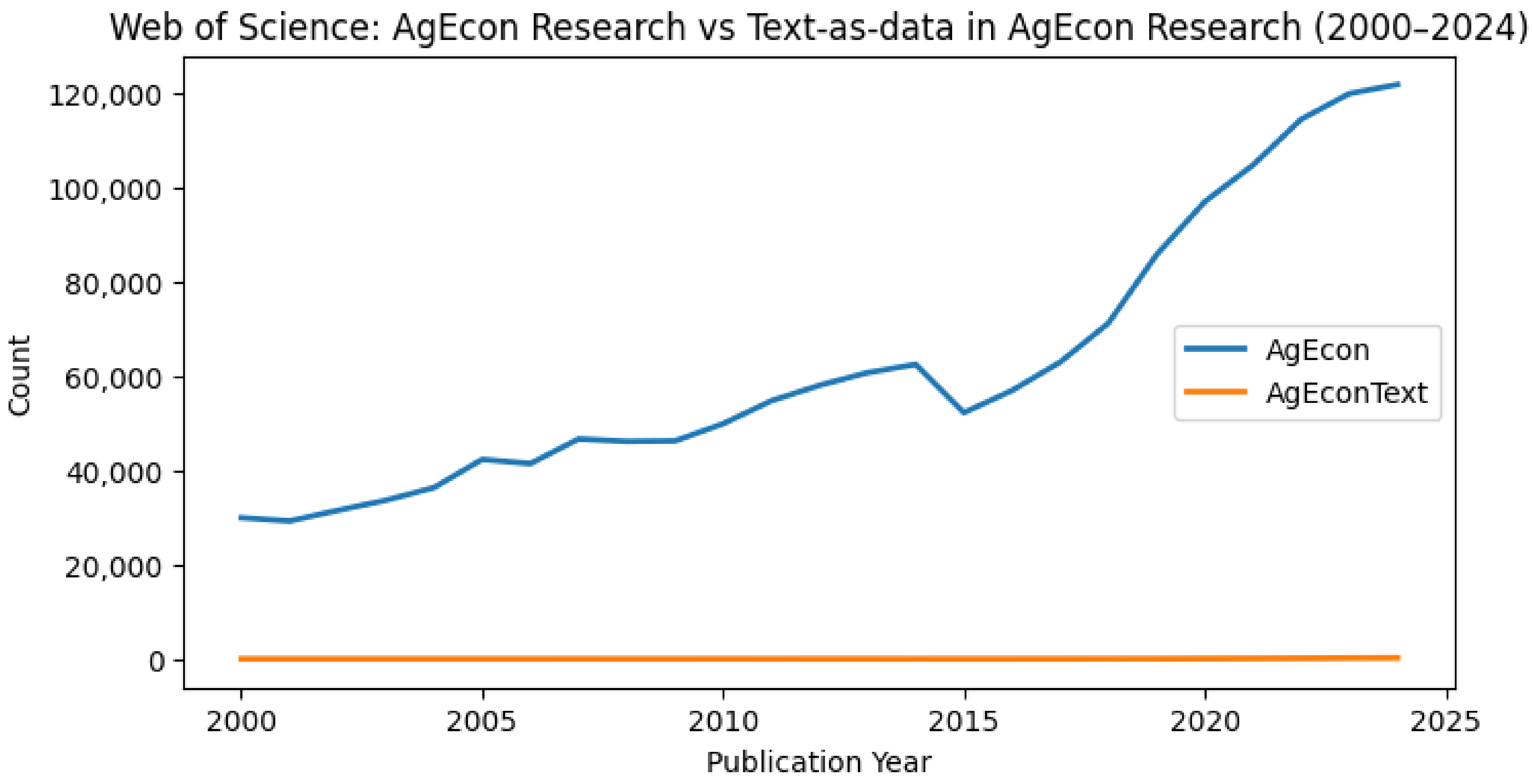
**IACS**

**Etc.**

Rigorous, numerical, clean, and directly usable for econometrics.

But by relying on these data, do we strip out the rich **qualitative context** and **nuance** found in more unstructured data such as text?

# Text data in AgEcon Research




**TS=("agricultural economics" OR "agriculture" OR "farm" OR "agribusiness" OR "agricultural policy")** and **Preprint Citation Index** (Exclude – Database) and **Research Commons** (Exclude – Database) and **Article** or **Review Article** (Document Types)

**TS=("text mining" OR "text analysis" OR "natural language processing" OR "sentiment analysis" OR "topic modeling") AND TS=("agricultural economics" OR "agriculture" OR "farm" OR "agribusiness" OR "agricultural policy")** and **Preprint Citation Index** (Exclude – Database) and **Research Commons** (Exclude – Database) and **Article** or **Review Article** (Document Types)

# Text data explosion

- Massive growth of unstructured data in recent decades
- Storage capacity
- Digitized text available for computational analysis


EUR-Lex **>1.15M** legal documents on EUR-LEX (Oct 2025)

amazon >**500M Amazon reviews (**May. 1996 to Sep. 2023)

**~7,5M** blog posts daily

>**140,000** documents FAO Knowledge Repository (Oct 2025)

WEB OF SCIENCE > **240M** records (journals, books, and proceedings) on Web of Science
>**200,000 AgEcon** related documents

# Text data as an opportunity!

There is an overwhelming amount of text available today.

Making manual analysis impractical.

Without modern tools and techniques to process and analyze large text corpora, we risk missing important patterns, trends, and knowledge.

# Text data is unstructured

- Different languages
- Different formats: txt, pdf, docx, etc.
- Different forms: Emails, social media posts, interview transcripts

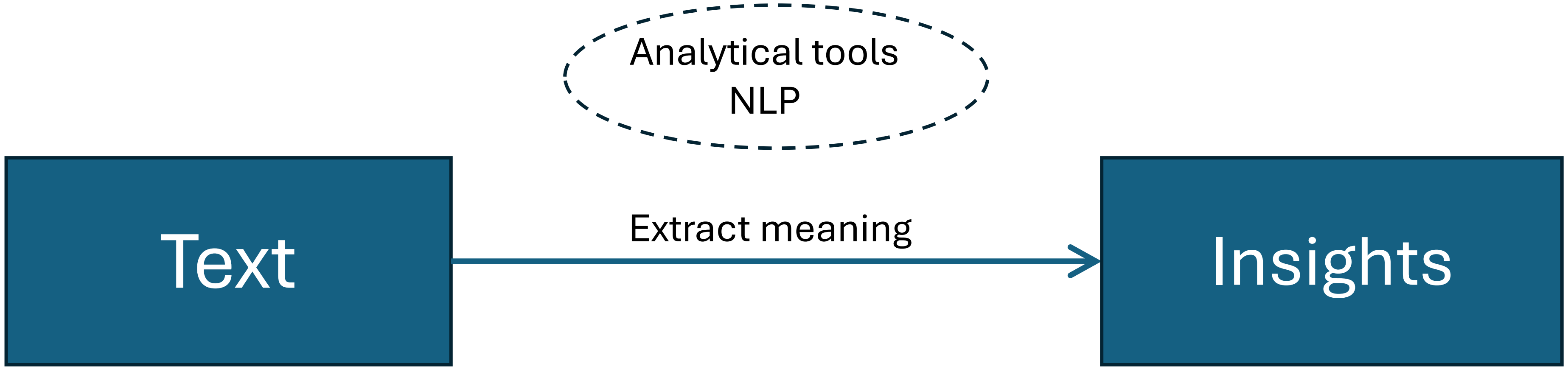
| Aspect | Structured Data | Unstructured Text Data |
|---|---|---|
| **Format** | Tabular (rows and columns) | Free-form text, variable length and format (sentences, paragraphs) |
| **Interpretation** | Machine-readable | Human-readable, context-dependent |
| **Consistency** | High consistency across records | Highly variable in length and content |
| **Information** | Quantitative, often limited to predefined fields | Qualitative (& quantitative), rich in nuance and context |

*What can we do with text data?*

# Natural Language Processing (NLP)

“Natural language processing(NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language.”

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227


Analytical tools
NLP
Text
Extract meaning
Insights

# NLP Tasks


Source:
https://www.nlplanet.org/course-practical-nlp/

# NLP Tasks

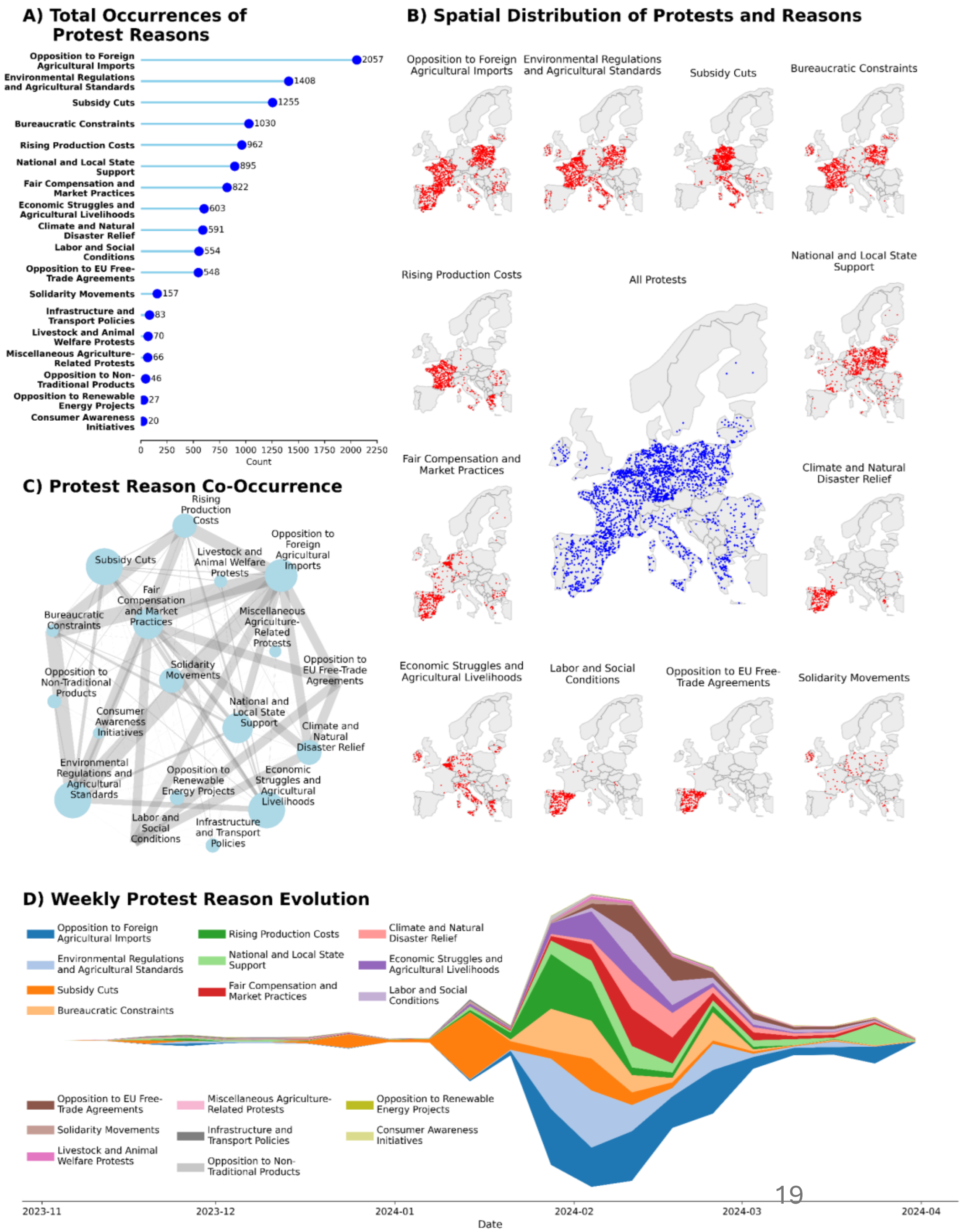

Source:
https://www.nlplanet.org/course-practical-nlp/

# Sentiment Analysis

- Quantify public sentiment about current agricultural practices.
- Quantify farmers’ sentiment about the current agricultural policy framework.
- Assess farmers’ sentiment about digitalization.

# Text Classification

Classify farmer protest descriptions according to protest reasons and quantify their spatiotemporal prevalence.



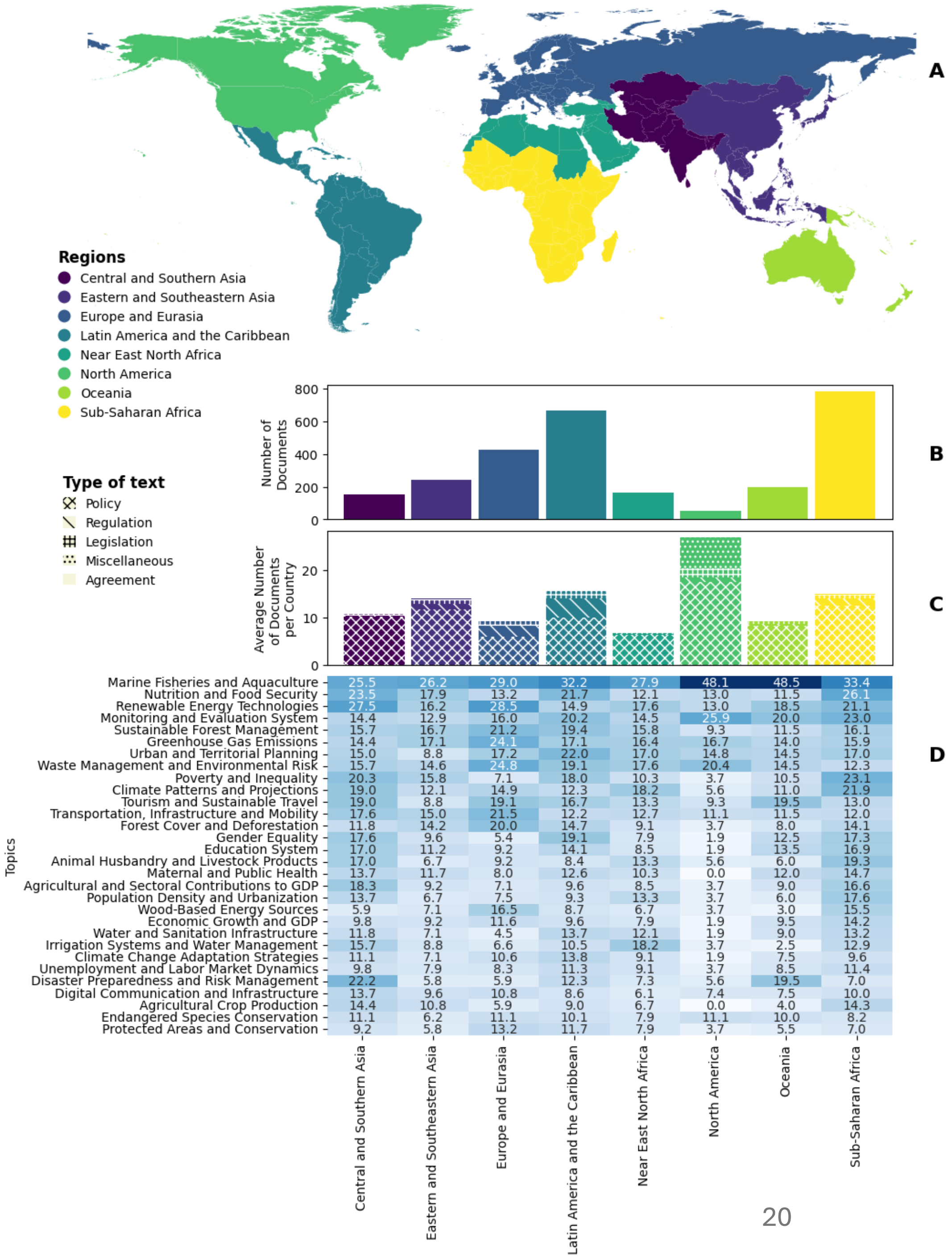

Stetter, C., Meemken, E.-M., Fürholz, A., & Finger, R. (2025). *Large language model analysis reveals key reasons behind massive farmer protests in Europe*. In Review. https://doi.org/10.21203/rs.3.rs-6652927/v1

# Topic Model

Assess spatiotemporal topic prevalence and trends globally based on legal documents.

*“This study employs AI-driven natural language processing to identify key patterns, trends, and themes in climate change legislation by analyzing 264,739 pages of climate-related legal texts from 189 countries in 51 languages between 2001 and 2022.”*

Stetter, C., Huber, R., & Finger, R. (in preparation). *Global Insights into Two Decades of Climate Legislation in Agriculture, Forestry and Fisheries*.



# Why do we need modern AI for NLP tasks?

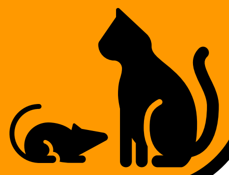
Because prior methods could not effectively address the **scale, complexity**, and **inherent lack of structure** in human language.

# Long-standing problems in NLP

## Word Order/Context Understanding

Bag-of-Words and early models ignored word order.

Example:
“The cat chased the mouse.” ≠
“The mouse chased the cat.”



## Multiple Meanings (Polysemy)

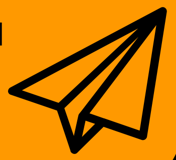
Models like Word2Vec gave one meaning per word.

Example:
“He sat by the river bank.”
“She works at the bank.”

## Long-Range Dependencies

RNNs forget earlier context — cannot link distant words or ideas.

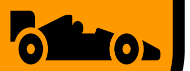
Example:
“John, who was born in France and speaks fluent French, moved to the US. …He still prefers reading books in ____.”



## Parallelization & Efficiency

RNNs process text token by token, slowing training.

Example:
Training on 10 million sentences:
RNN: sequential → days or weeks
Transformer: parallelized → hours



# Modern AI-based NLP

# Modern NLP Relies on Transformer Models

- Transformers are a type of deep learning model designed for processing text.
- Unlike earlier models that looked at words one at a time or in fixed-length chunks
  - transformers can **consider the *entire* context** of a sentence or paragraph at once.

## Key mechanisms:

- **Positional encoding:** Since transformers do not read text sequentially like humans do, they use positional information to keep track of word order.
- **Attention and self-attention:** Allow the model to focus on relevant words in a text chunk, no matter how far apart they are.
  - For example, *"The **cat** that was chased by the dog **ran away**,"* the model can connect *"cat"* with *"ran away"* even though separated by other words.

# Transformers: Architecture Intuition

**1** Text Input

"Hello World!"

Break into Words/Tokens

**2** Individual Words/Tokens

["Hello", "World", "!"]

Give Each **Word a Numerical Representation & Position Encoding**

**3** Words now have Numbers & Order

"Hello": [543, pos 1]
"World": [987, pos 2]
"!": [1629, pos 3]

Add Context with **Self-Attention**

**4** **Output**: Each word/token now has a contextualized numerical representation

$$\begin{array}{r l} \text{“Hello”} & \\ \text{“World”:} & \\ \text{“!”} & \end{array} \begin{bmatrix} 0.12, -0.05, 0.78, 0.33 \ldots \\ 0.85, 0.21, -0.62, 0.15 \ldots \\ 0.79, 0.18, -0.03, 0.00 \ldots \end{bmatrix}_{3x1024}$$

# Transformers – “Prediction Heads”

Add Context with
**(Self-)Attention**

4

**Output**: Each word/token now has a contextualized numerical representation

$$\begin{array}{r l} \text{“Hello”} & \multirow{3}{*}{} \\ \text{“World”:} & \\ \text{“!”} & \end{array} \begin{bmatrix} 0.12, -0.05, 0.78, 0.33 \ldots \\ 0.85, 0.21, -0.62, 0.15 \ldots \\ 0.79, 0.18, -0.03, 0.00 \ldots \end{bmatrix}$$

$3x1024$

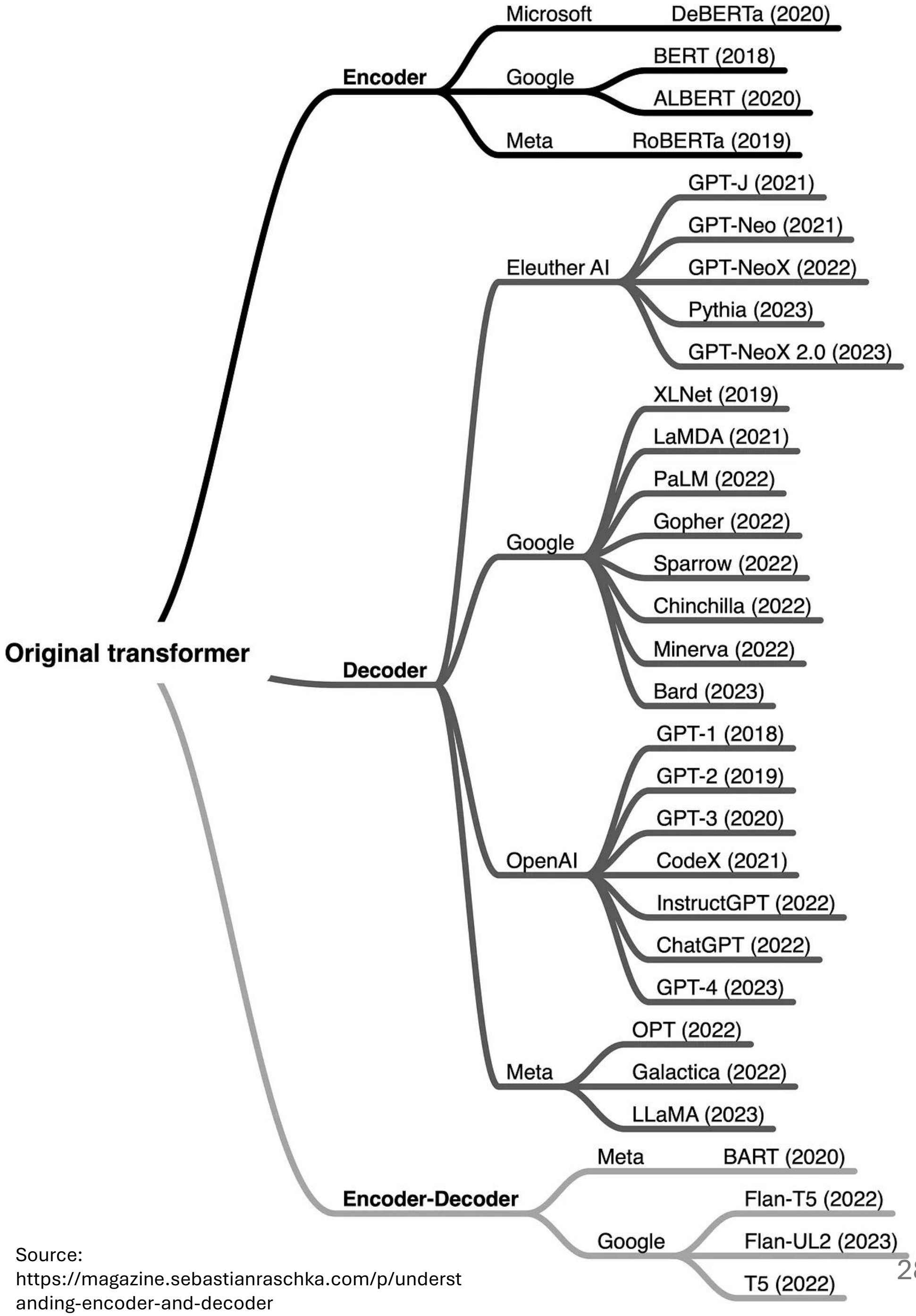
Additional **layers** or **mechanisms** added on top of the pre-trained base model to adapt it for specific tasks ***(often added during finetuning).***

5

**Output**: Classification, Regression, Next-token prediction, etc.

These models are typically **(pre-) trained on (massive) datasets** of text.

- Strong general understanding of language
- Usable for downstream tasks




Source: https://magazine.sebastianraschka.com/p/understanding-encoder-and-decoder

## Training

- The phase where the model **learns** from data.
- Involves **adjusting model parameters** (weights and biases) to minimize error.
- **Computationally intensive** — requires large datasets and high-performance hardware (e.g., GPUs/TPUs).
- OpenAI, Google, Meta, etc.

Output: a **trained model**, not necessarily ready for deployment.

## Fine-tuning

- Starts with a **pretrained model**
- The model is then **retrained** (partially or fully) on a **smaller domain-specific dataset** to adapt it to a particular task.
- Adjusts **some or all weights**, but typically with **fewer parameters updated** and **less data** than full training.
- Used to achieve **better accuracy or relevance** for specific use cases

Output: a **trained model** ready for deployment.

## Inference

- The phase where the pretrained model is **used to make predictions** on new, unseen data.
- Model parameters are **fixed** — no further learning occurs.
- **Faster and less resource-intensive** than training.
- End-users

Output: **predictions, classifications, or generated text.**

# Encoder-only models

# Encoder-only models (BERT, RoBERTa, MPNet)

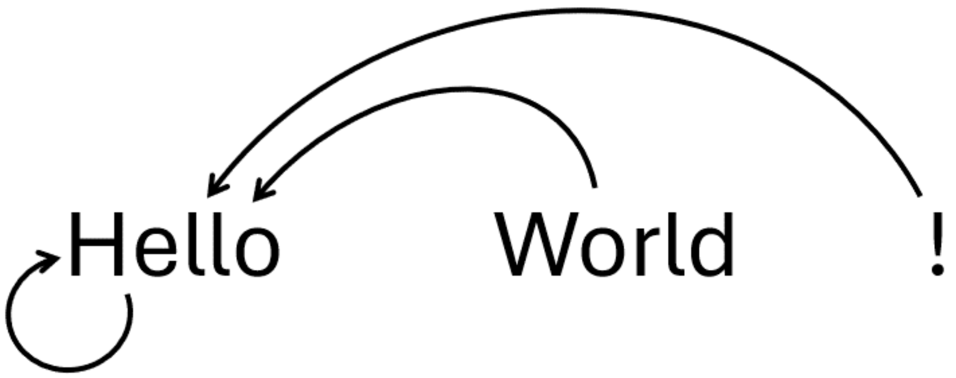
**Core Idea:**
Process input **bidirectionally** → every token sees every other token.

Learn **deep contextual embeddings** for understanding text.

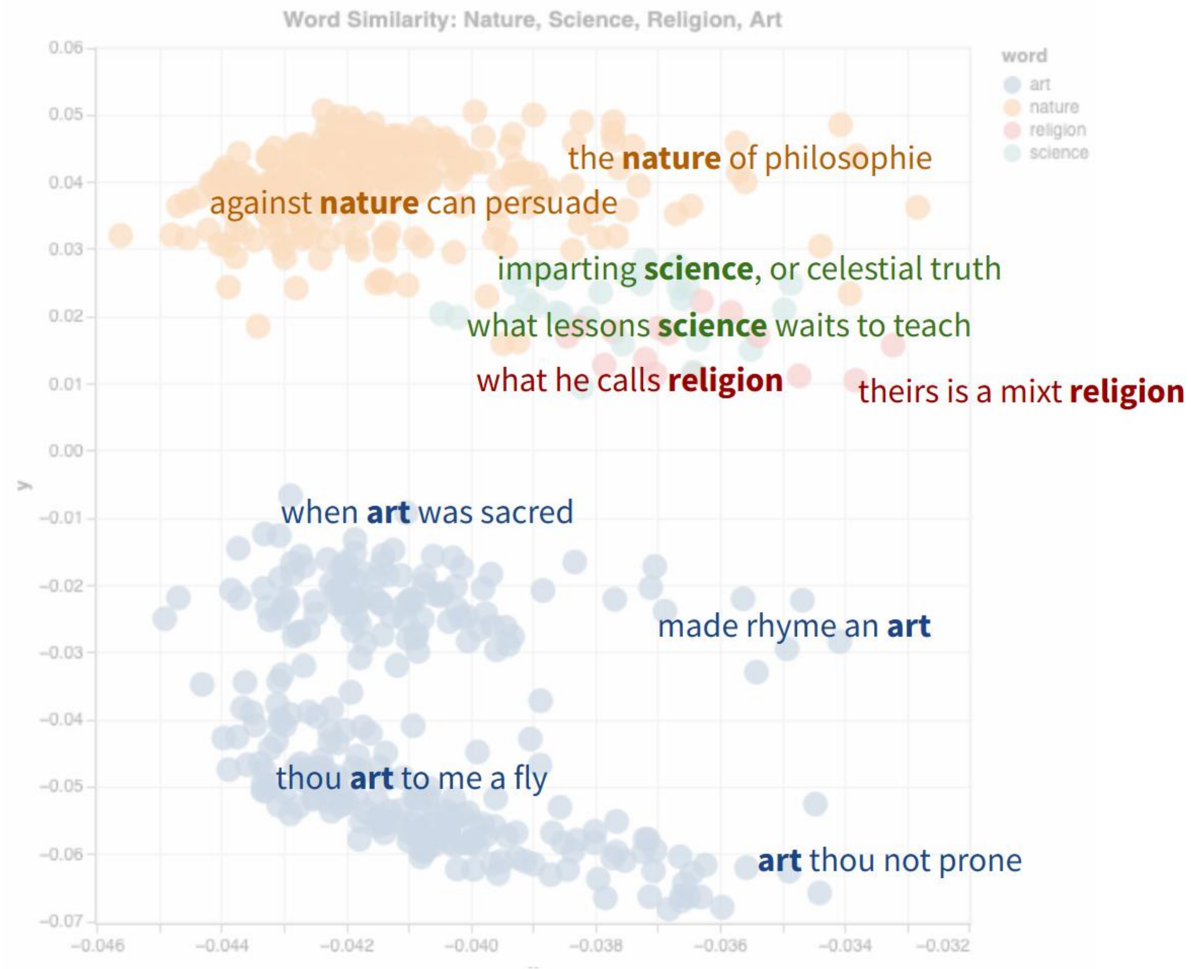
Not generative — they **understand** but don't **produce** new text

**Key Strengths:**

- Text Classification (sentiment, spam detection)
- Named Entity Recognition (NER)
- Semantic Search / Sentence Similarity
- Fast parallel processing for understanding tasks



# Contextualized Embeddings Illustrated




Antoniak, M., Walsh, M., Mimno, D., & Wilkens, M. (n.d.). *BERT for the social sciences and humanities.* Google Slides. https://docs.google.com/presentation/d/1PwibAk_Fvnk-sCfZ0C5-4UYvqGUEU969FAyFsdDuGyU/edit

# Hugging Face

**What is Hugging Face?**

- A leading open-source platform for Natural Language Processing (NLP) and (generative) AI.
- Hosts thousands of pre-trained (& finetuned) models, datasets, and other tools.
- Built around the idea of sharing and collaboration in AI research and application.

**Why it matters:**

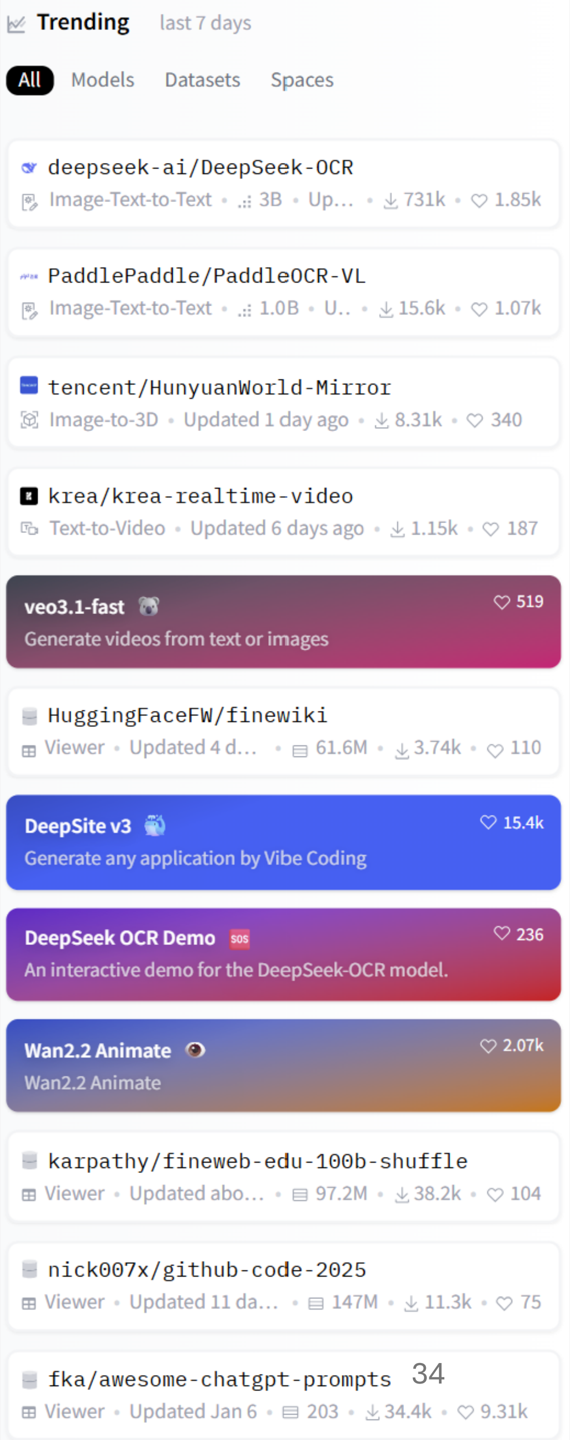
- Democratizes access to advanced language models (e.g., BERT, GPT, T5).
- Enables fast experimentation — “use powerful models in just a few lines of code.”

https://huggingface.co/

# Hugging Face Core Components

- **Transformers Library:**
  Use and finetune pre-trained models in Python.
- **Datasets Library:**
  Standardized access to public datasets.
- **Hugging Face Hub:**
  Central repository for sharing models & results.
- **Blog & Community**

  Regular **blog posts** on:
  - New model releases (e.g., BERT, BLOOM, Mistral)
  - Research explainers (attention, alignment, quantization)
  - Tutorials on using models in production
  - Responsible AI and ethics discussions
  - Users can **comment**, **share notebooks**, and **collaborate**.



# Demo 1

What is the sentiment of agricultural economists (interested in AI) associated with the role of artificial intelligence in their field?

https://colab.research.google.com/drive/1eNdqupHL8AfHfOJmtESZgSg6lTlovU1v?usp=sharing

# Limitations

- Small context window
- Finetuned models available but not necessarily fit own use case
- Finetuning often necessary for specific task
- Finetuning requires usually >1k samples

# Decoder-only models:
# Modern LLMs

# Decoder-Only Models (GPT, LLaMA, Claude)

- LLMs generate text by learning statistical patterns in language
- Generate text **autoregressively** — one token at a time, left to right. Learning to predict the next word (or token) in a sequence.
- This prediction is based on probabilities: given the words you have already seen, what is the most likely next word?
- Optimized for **generation**, not bidirectional understanding.


**Traditionally best for:**

- **Text Generation (stories, essays, code)**
- Creative Writing
- …

# Different Models

**Propriety**

- OpenAI GPT
- Anthropic Claude
- Google Gemini

**Key points:**

- Accessible through API
- Less transparent
- Often better performance
- Paid access
- Provide computational resources for inference

**Open-Source**

- Mistral
- OpenAI-GSS
- LLAMA

**Key points:**

- Freely available (e.g., via Hugging Face)
- Free to use & transparent
- Requires your own computational resources

# Decoder-only LLMs for both generation and understanding

**Trend:**

! Modern **decoder-only LLMs (GPT-4, Claude, Gemini, LLaMA 3, etc.)** now excel not only at text generation, but also at **understanding-focused tasks** once dominated by encoder-only models.

**Why This Shift Happened:**

**Instruction Tuning & RLHF:** Train LLMs to follow natural language instructions and reason.

**Massive Scale:** Billions of parameters and diverse training data improve semantic understanding

Conferences > 2023 IEEE International Confe...

**Prompting Large Language Models for Topic Modeling**

Publisher: IEEE Cite This PDF

Han Wang ; Nirmalendu Prakash ; Nguyen Khoi Hoang ; Ming Shan Hee ; Usman Naseem ; Roy Ka-Wei Lee

NeurIPS Proceedings

Large Language Models are Zero-Shot Reasoners

Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Main Conference Track

Bibtex Paper Supplemental

Authors

*Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa*

**ChatGPT outperforms crowd workers for text-annotation tasks**

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli Authors Info & Affiliations

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

**July 18, 2023** | 120 (30) e2305016120 | https://doi.org/10.1073/pnas.2305016120

arXiv > cs > arXiv:2306.17177

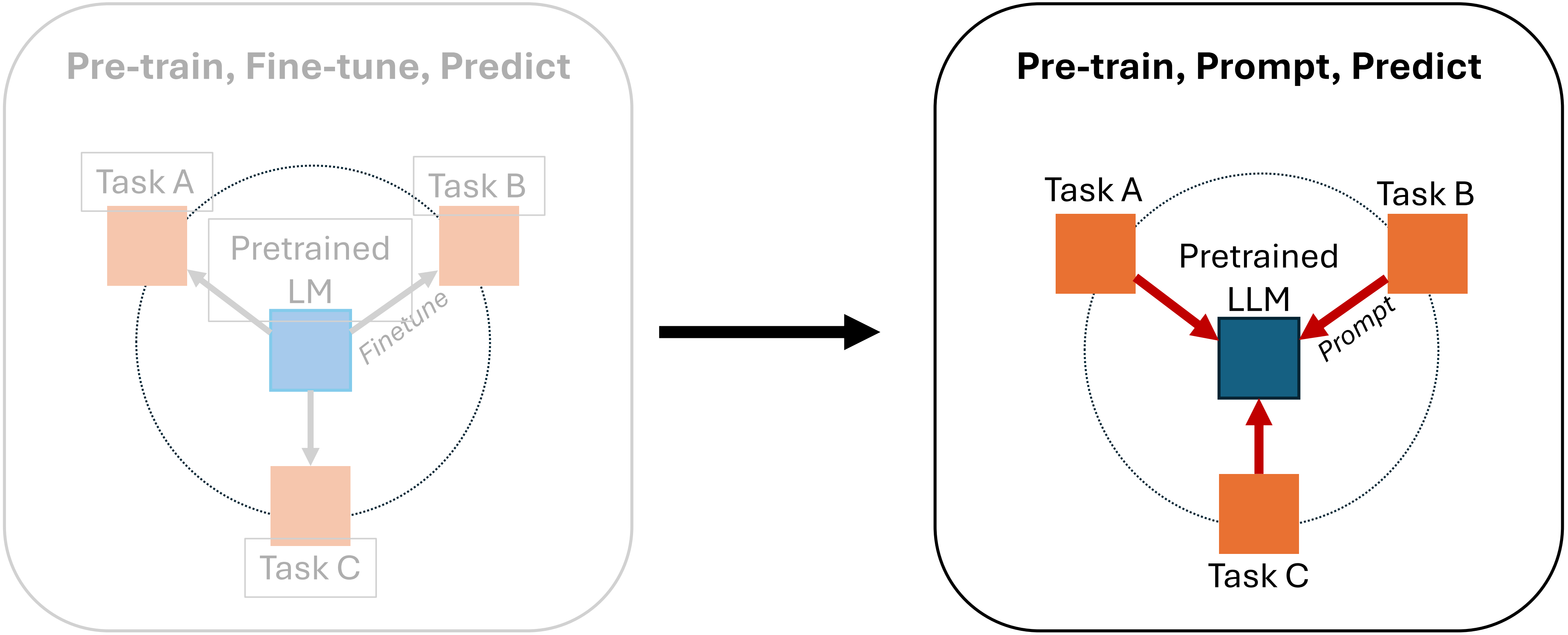
**Computer Science > Computation and Language**

*[Submitted on 18 Jun 2023]*

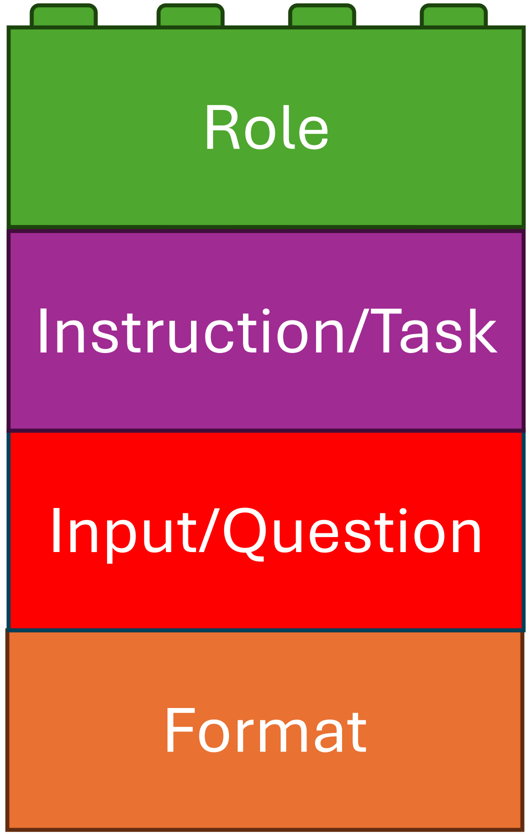
**Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis**

Mohammad Belal, James She, Simon Wong

# Paradigm shift: Pre-train, prompt, predict



# Prompt elements




**Role:**
You are an agricultural economist specializing in rural development and farm management.

**Instruction/Task:**
Classify the following farm enterprises based on their type of agricultural system (subsistence farming, commercial farming, mixed farming, plantation farming, or organic farming). Provide reasoning for each classification.

**Input/Question:**

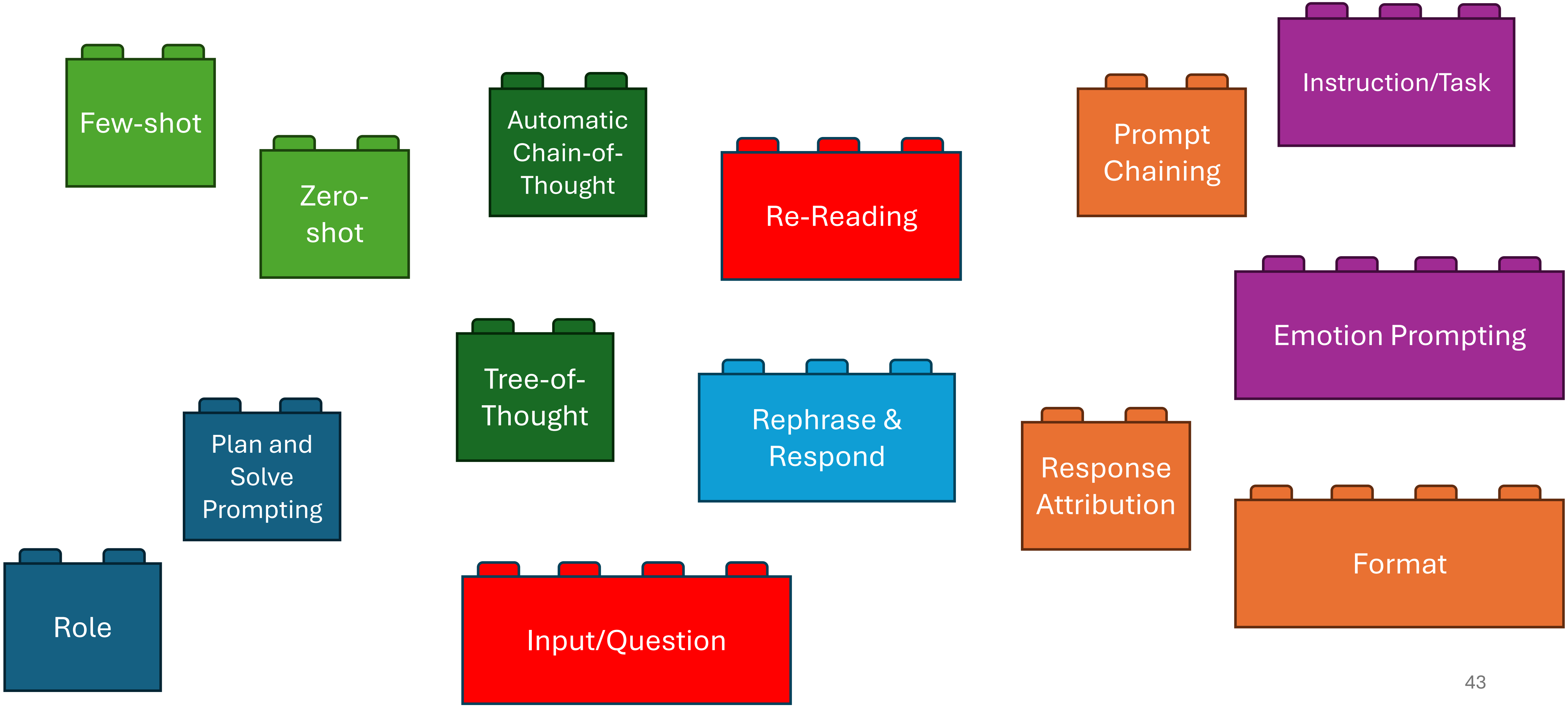
A) A 2-hectare maize farm in rural Kenya that produces mainly for family consumption.

B) A 5,000-hectare palm oil estate in Malaysia exporting to global markets.

C) A 10-hectare farm in France combining dairy and crop production for local markets.

**Format:**
Provide your answer in a **table** with the following columns:
| Case | Farm Description | Classification | Reasoning |

# Prompt Engineering: Building Block Thinking



# Prompt Engineering: Recommendations

## Start Simple

- Prompting is an **iterative process: experiment** to improve results.
- Begin with basic prompts and gradually add context and details.
- **Iterate** and refine prompts for better outcomes.
- Simplicity, specificity, and conciseness often yield the best results.
- **For complex tasks, break them into smaller subtasks** to manage complexity.

## Avoid Imprecision

- **Avoid being too vague** with prompts.
- Aim to be specific, clear, and direct in your wording.
- Effective communication: direct prompts produce better results.

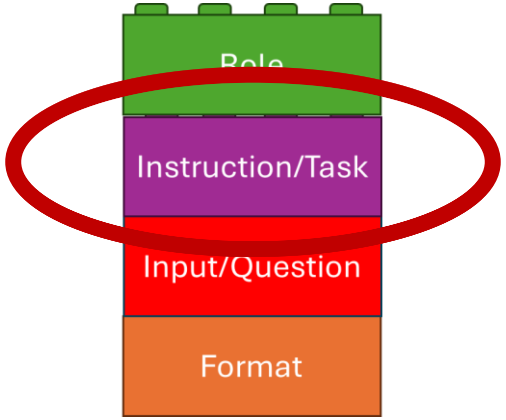
## Instructions

- **Experiment frequently** to find what works best for your task.
- Try different instructions, keywords, contexts, and data for optimization.
- **Specific and relevant context** usually leads to better result

## Specificity

- **Be specific and detailed in instructions to get better results**.
- Keep prompt length in mind. Avoid unnecessary details. Include only relevant information that supports the task.
- **Experiment and iterate** to find the right balance of detail and brevity.
- Describe clearly the desired outcome or style.

# Few-shot vs zero-shot learning




- Providing a few examples in the prompt can improve NLP performance of LLMs

## Zero-shot

**Task:** Classify each agricultural economics text into one of the categories: *Policy, Market Trend, or Technology.*

**Text:** "Government increased wheat subsidies to support rural farmers."

## Few-shot

**Task:** Classify each agricultural economics text into one of the categories: *Policy, Market Trend, or Technology.*

**Example 1:**
Text: "Corn prices fell due to global oversupply."
Label: Market Trend

**Example 2:**
Text: "Farmers adopted drones for crop monitoring."
Label: Technology

**Example 3:**
Text: "New tax credits for dairy producers were announced."
Label: Policy

**Now classify:**Text: "Government increased wheat subsidies to support rural farmers."

# Model Selection & Evaluation

- Potential criteria for choosing an LLM:
  - size
  - speed
  - cost
  - task-specific performance
  - open-source vs API.

# Demo 2

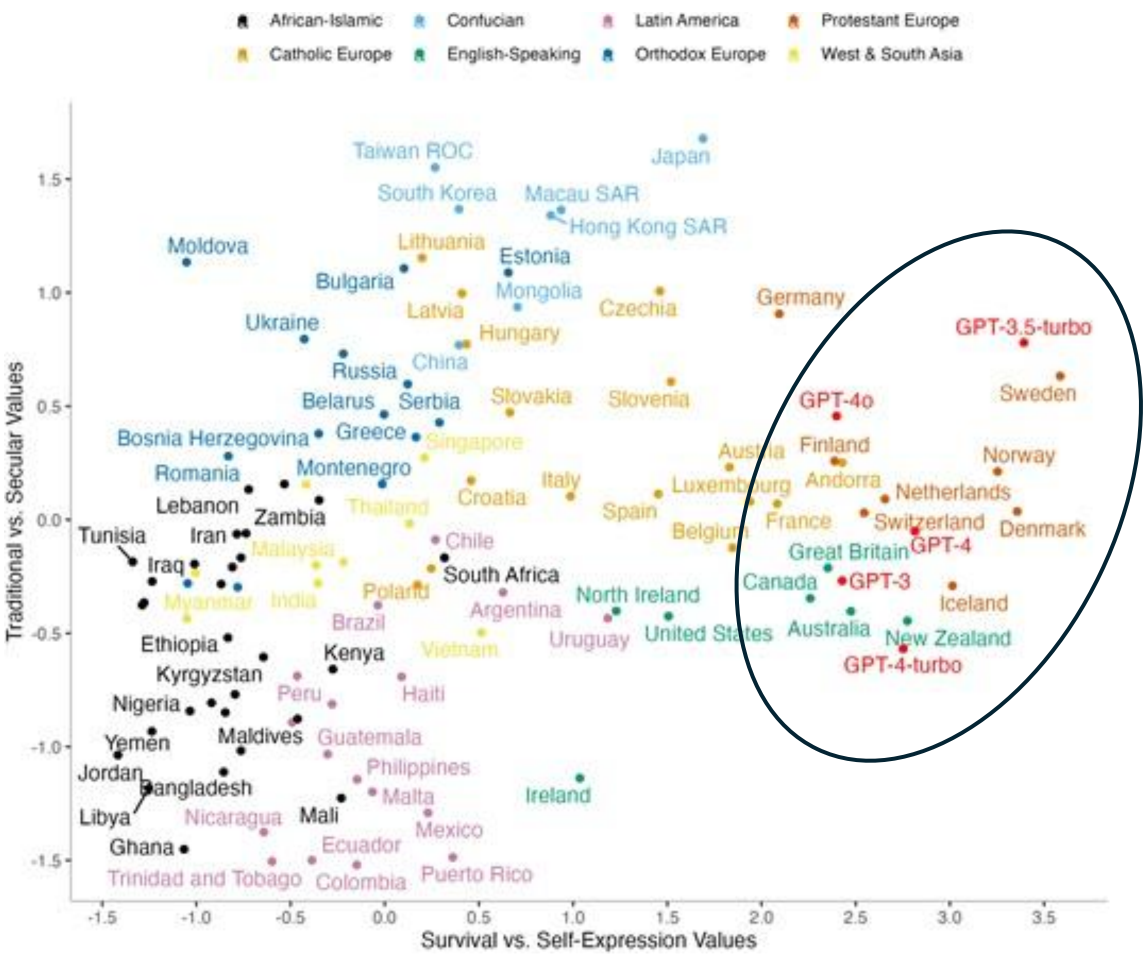
What scientific policy implications related to agriculture and climate change are there, and who are the intended recipients (farmers, policymakers, or both)?

https://colab.research.google.com/drive/1E1r3C_4TT2Aw26z7xPg0Y_sj3gqpKO4Z?usp=sharing

# A Key Concern

- Pre-trained models can inherit **biases** present in their training data.
- This can lead to skewed results, especially when analyzing texts related to gender, race, or other sensitive topics.
- It is important to be aware of these biases and take steps to mitigate them.

World Values Survey Map




Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

# Further Issues for Discussion

- The Role of Prompting
- Hallucinations
  - Factually incorrect, nonsensical, or irrelevant responses
  - Risk in decision-making and research interpretation
- Reproducibility
  - Model outputs can vary across runs or platforms
- Black Box Nature
  - Lack of transparency in how models reach conclusions
  - Difficult to explain or justify results in policy contexts
- Memorization instead of Generation
  - Training data of the model might contain text analyzed in one's task
- Quality Control & Uncertainty Quantification
  - Lack of standardized evaluation metrics. benchmarks
  - Need for confidence scores or error bounds in AI-assisted analysis
- Data Protection & Privacy
  - Sensitive farm, trade, or household data require secure handling
- Environmental Concerns
  - High energy consumption in model training and deployment
  - Trade-off between computational scale and sustainability
- Access and dependency issues
  - Big tech dominance
  - Open-source models need a lot of computational resources
  - Open-source models are catching up

# Key Takeaways


© marketoonist.com

### Data Opportunity

- Move beyond structured data to mine vast unstructured text (policy, reports, media).
- Scale vs. Depth: Modern AI achieves deep semantic understanding at massive scale.

### Method

- Leverage Pre-trained Models: Use powerful Pre-trained Transformers (BERT, GPT) that know language rules.
- Task Adaptation: Potentially fine-tune these models for specific AgEcon tasks.
- Need for benchmarks and training data in AgEcon domain.
- Prompt LLMs for simple and complex NLP tasks.

### Research Opportunities

- Produce potentially novel insights previously unattainable.
- Turn qualitative text into structured, quantitative variables (e.g., sentiment scores, policy targets).
- Combine extracted text data & results with traditional economic models.
- Unlock new questions: Explore market sentiment, public perception, and emerging policy trends.

# Further reading

- In-depth explanation of transformer architecture: https://poloclub.github.io/transformer-explainer/
- Open-access NLP intro book: https://www.nlplanet.org/course-practical-nlp/


- Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, *15*(1), 659–688. https://doi.org/10.1146/annurev-economics-082222-074352
- Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121
- Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), 1–3. https://doi.org/10.1073/pnas.2305016120
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3560815
- Saravia, E. (2022). Prompt engineering guide. *Https://Github.Com/Dair-Ai/Prompt-Engineering-Guide*.
- Stetter, C., Meemken, E.-M., Fürholz, A., & Finger, R. (2025). *Large language model analysis reveals key reasons behind massive farmer protests in Europe*. In Review. https://doi.org/10.21203/rs.3.rs-6652927/v1
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

# Thank you!

cstetter@ethz.ch

linkedin.com/in/christian-stetter-135958159/

https://scholar.google.de/citations?user=EOW7rSAAAAAJ&hl=en