LLMs and Retrieval-Augmented Generation

Tools for Sustainable Finance Research and Practice

Vitali Alexeev

Finance Department, UTS Business School
University of Technology Sydney, Australia

Centre for Climate Risk and Resilience, UTS
Seminar Series
November 13, 2025

Table of Content

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Sentiment Analysis: Early Days

Loughran–McDonald Dictionary

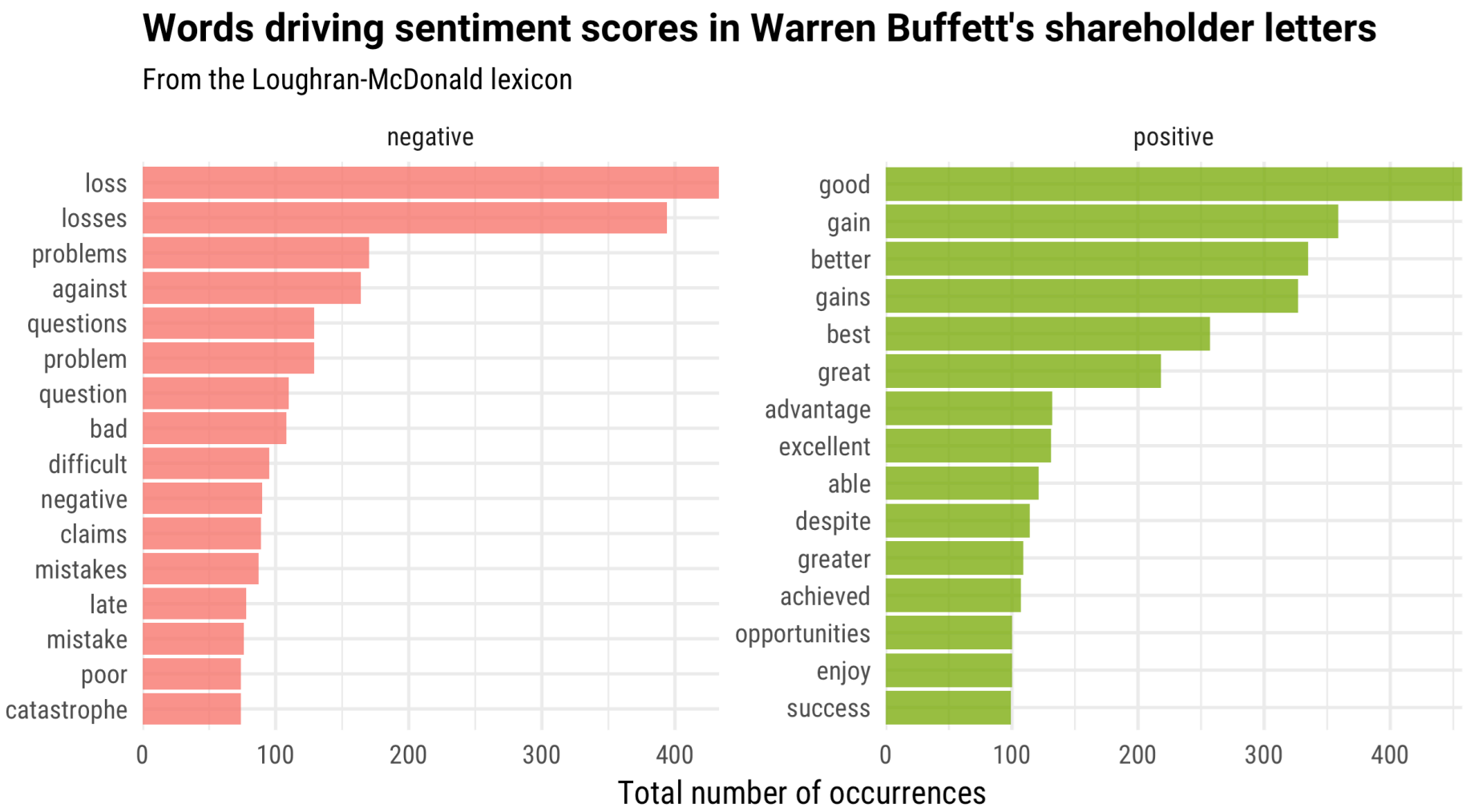
▸ XLSX download

- Count positive/negative words ⇒ sentiment proxy
  - e.g., *(pos − neg) / total*
- Transparent, fast, finance-specific
- **Strengths:** interpretable, reproducible
- **Weaknesses:**
  - Polysemy: high carbon intensity → positive word, negative meaning
  - Negation ignored (not significant)
  - Domain drift: evolving ESG vocabulary

Sentiment Analysis example

Warren Buffett's Letters to Shareholders

source



From Unigrams to Bigrams (N-grams)

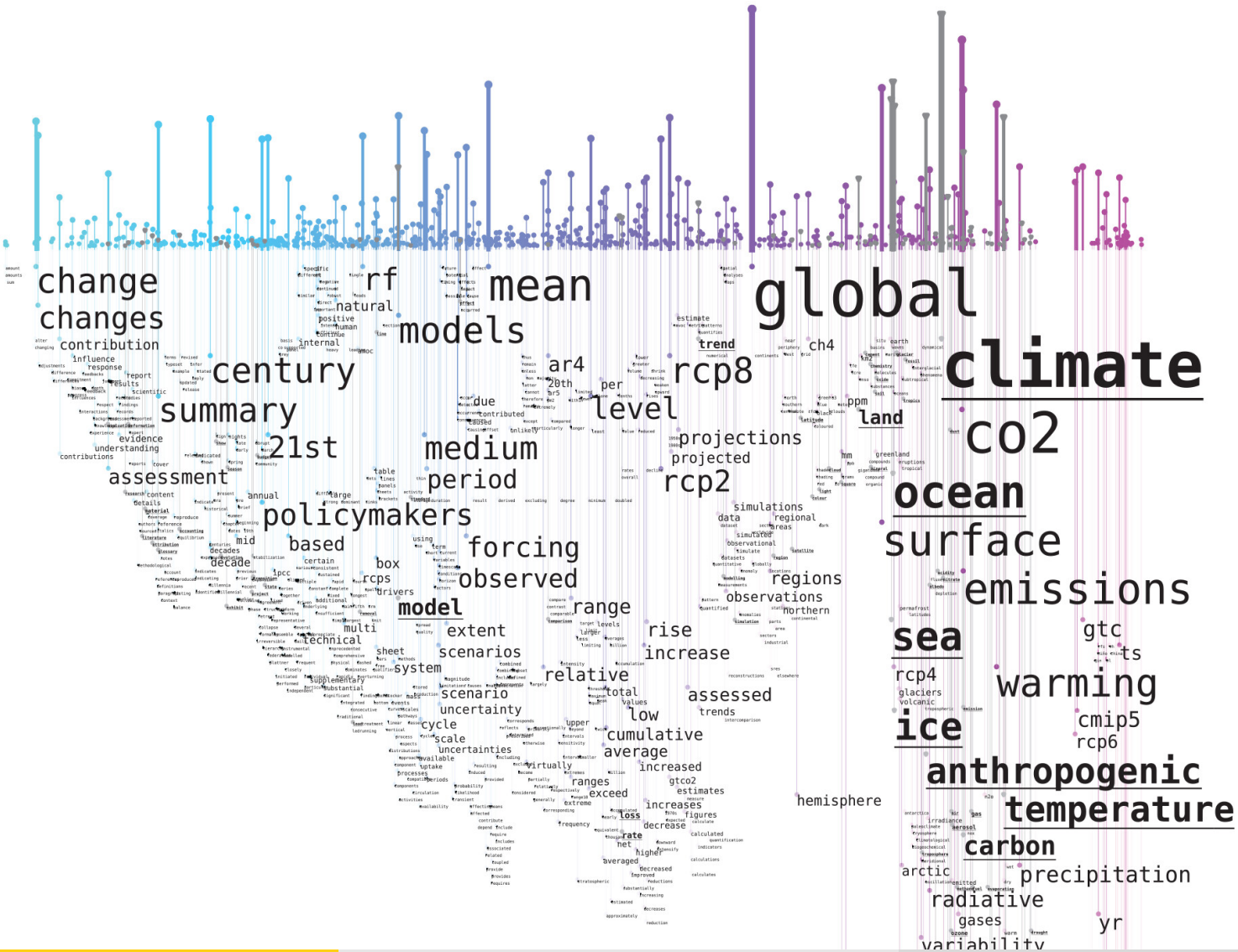
- Capture short phrases: carbon neutral, not compliant, scope 3, high risk, coal phase-out, methane leakage, board diversity

1-Gram	2-Gram	3-Gram
The	The Margherita	The Margherita pizza
Margherita	Margherita pizza	Margherita pizza is
pizza	pizza is	pizza is not
is	is not	is not bad
not	not bad	not bad taste
bad	bad taste	
taste		

Unigrams: "Climate Change 2021" report

Skeppstedt et al. 2024

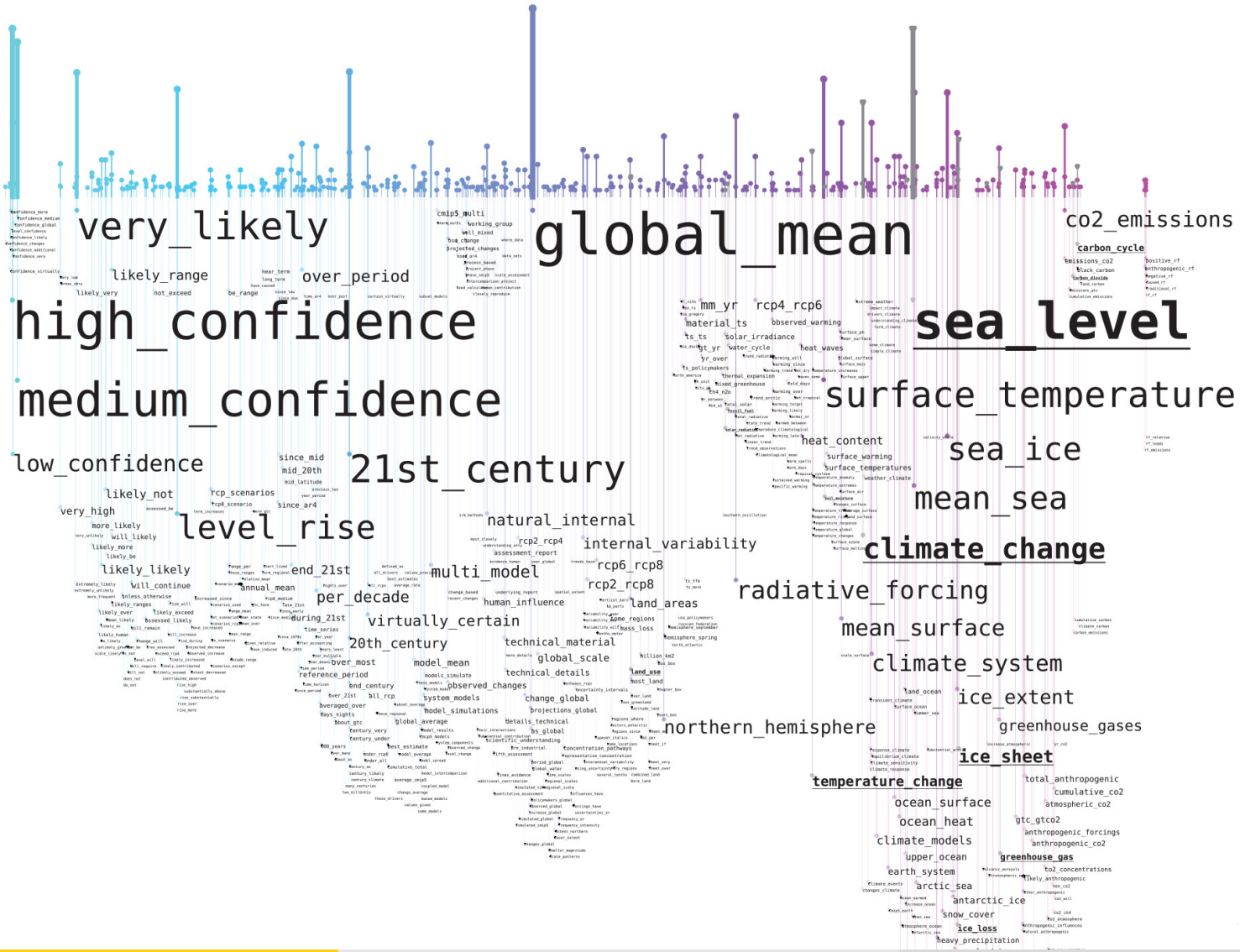
source



Bigrams: "Climate Change 2021" report

Skeppstedt et al. 2024

source



Word Embeddings

Introduction

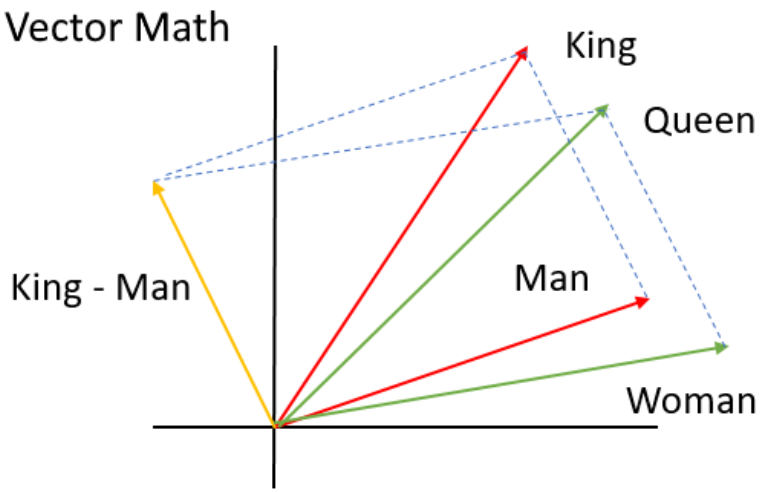
3D Embedding Projector | Word Analogies

What are Word Embeddings?

- Word embeddings are vector representations of words that capture **semantic relationships**.
- Traditional methods like **TF-IDF** represent words based on frequency.
- Modern embeddings like **Word2Vec**, **GloVe**, and **Transformers** capture **context and meaning**.

Word2Vec in Action

Vector Maths



Example of Word Relationships:

- **King - Man + Woman = Queen**
- **Paris - France + Germany = Berlin**

Why It Works:

- Words in similar contexts have similar vectors.
- Word2Vec learns relationships from raw text.

Transformers: Contextual Word Embeddings

BERT, GPT, and beyond

▸ Video: Concept ▸ Video: Transformers Visually Explained

Transformers use **self-attention mechanisms** to capture long-range dependencies in text, enabling advanced NLP applications.

- Unlike **Word2Vec**, words have different embeddings depending on context.
- **BERT (Bidirectional Encoder Representations from Transformers)**:
  - **Bidirectional Training**: Considers both left and right context.
  - **Masked Language Model (MLM)**: Predicts missing words in a sentence.
- **GPT (Generative Pre-trained Transformer)**:
  - **Autoregressive Training**: Predicts next word in a sequence.
  - **Fine-tuned for text generation**: Used in ChatGPT, summarization, and conversational AI.

Advantages:

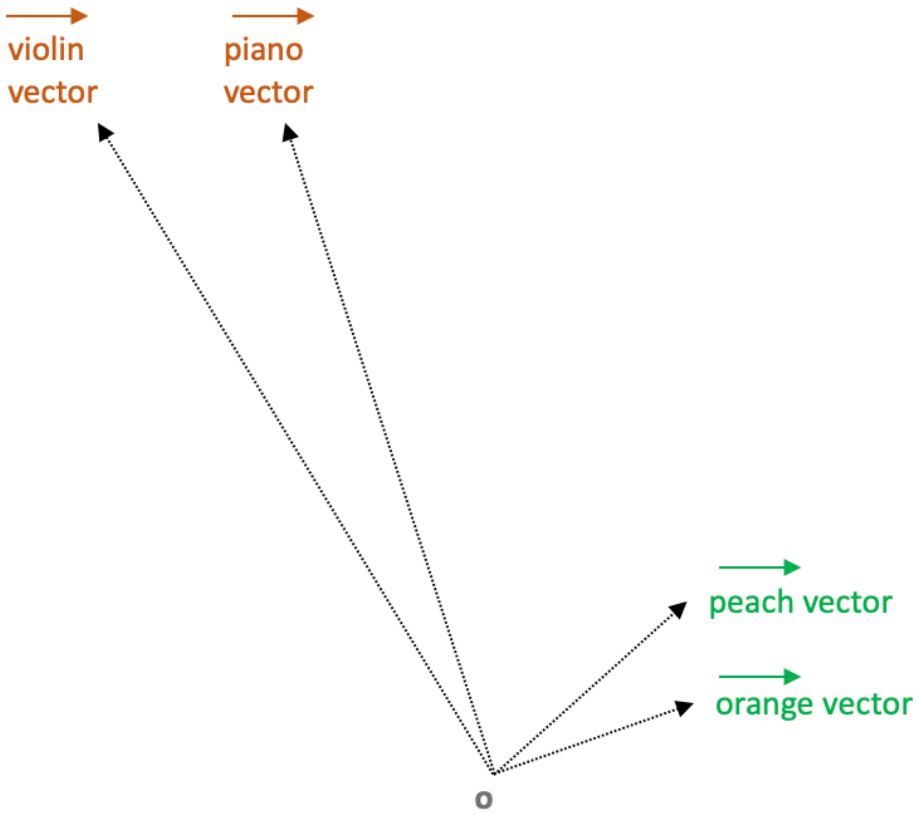
- ✓ Captures **contextual** meaning of words.
- ✓ Handles long-range dependencies in text.
- ✓ Powers state-of-the-art NLP applications (translation, Q&A, summarization).

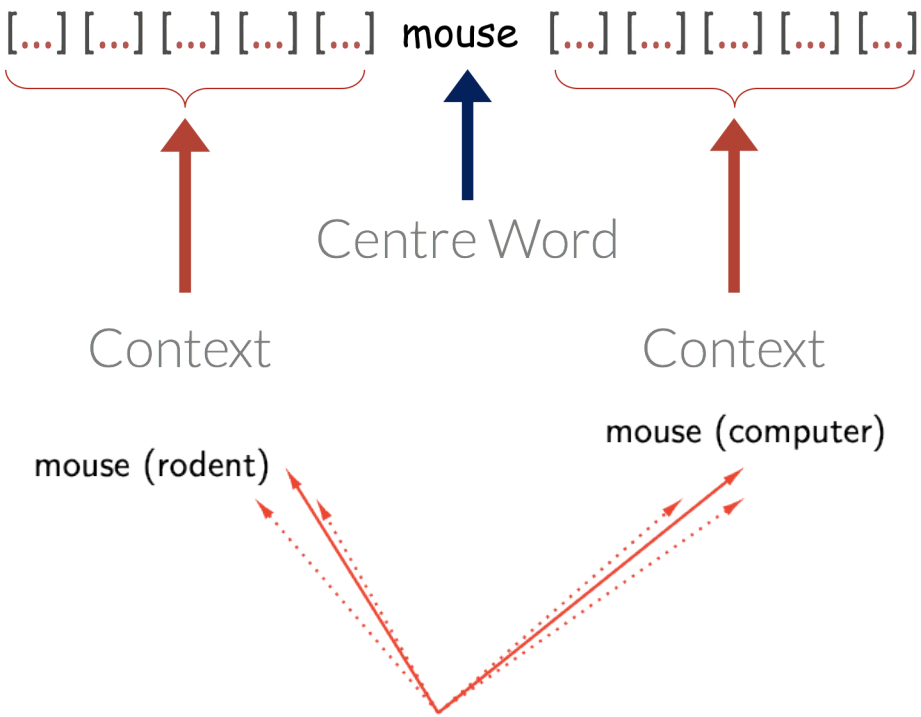
Limitations:

- ✗ Computationally expensive and requires large-scale hardware.
- ✗ May generate biased or misleading outputs.
- ✗ Requires extensive training data and fine-tuning.

Transformers: Contextual Word Embeddings

Words have **different embeddings** depending on context.





Example:

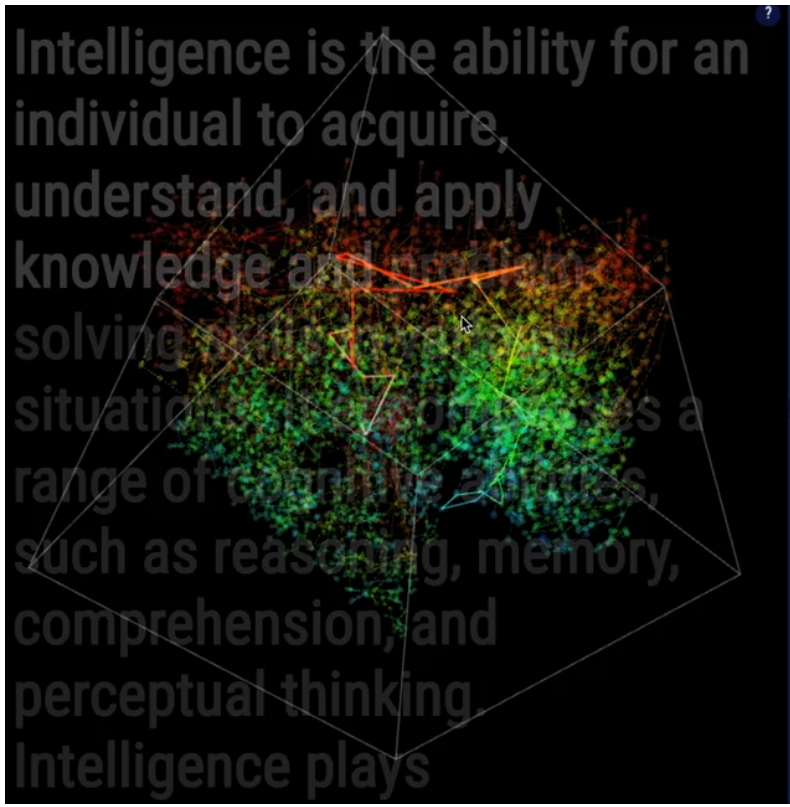
- "He went to the **bank** to withdraw money." → **bank** (finance).
- "The boat reached the **bank** of the river." → **bank** (geography).

Linguistic Highways

▸ Video ▸ Interactive

Mapped ChatGPT's thought process:

- Run the prompt *Intelligence is* hundreds of times with a high temperature setting (1.6).
- Capture the model's diverse responses.
- Use PCA, to compress the 1536-dimensional word embeddings into a 3D visualization, showcasing how AI builds sentences word by word.



3D cube displaying bifurcating response paths

tree diagram illustrating word probability choices

Transformer Models in Sustainable Finance NLP

Why Transformers? These models handle financial and ESG text with greater contextual understanding than traditional NLP.

- **FinBERT:** Fine-tuned BERT model for finance used to analyse sustainability-related sentiment in earnings calls, filings, and news.
  - Improves sentiment classification accuracy over rule-based models.
  - Used by asset managers for ESG sentiment tracking.
  - ▸ Paper
- **ClimateBERT:** Domain-specific model for climate-related financial disclosures (e.g., TCFD).
  - Classifies disclosures, extracts commitments and risk mentions.
  - Used in greenwashing detection and transition risk research.
  - ▸ Paper
- **Sustainability Summary (SusGen):** GPT-based model that summarises sustainability reports for stakeholders.
  - Aligns generated summary to SASB/GRI/TCFD frameworks.
  - Useful for investor briefings and automated ESG filings.
  - ▸ Paper

Transformers Models in Sustainable Finance NLP

BERT & GPT Models: Pretrained transformers improve classification, summarisation, and scoring of ESG content.

- **Use Case – ESG Scoring with E-BERT:** Fine-tuned BERT classifies reports with 93% accuracy into ESG performance levels. ▸ E-BERT Paper
- **Use Case – FinBERT-ESG:** Classifies ESG content by pillar (E/S/G) from annual reports. ▸ FinBERT-ESG
- **Use Case – ChatReport / SusGen:** LLMs summarise ESG reports aligned with TCFD. ▸ ChatReport ▸ SusGen

Hugging Face Ecosystem

Hugging Face

- **Model hub:** 500k+ models including finance/ESG
- **Dataset:** CDLA, climate reports, emissions corpora
- **Pipelines:** rapid experimentation; `transformers` for research-grade work

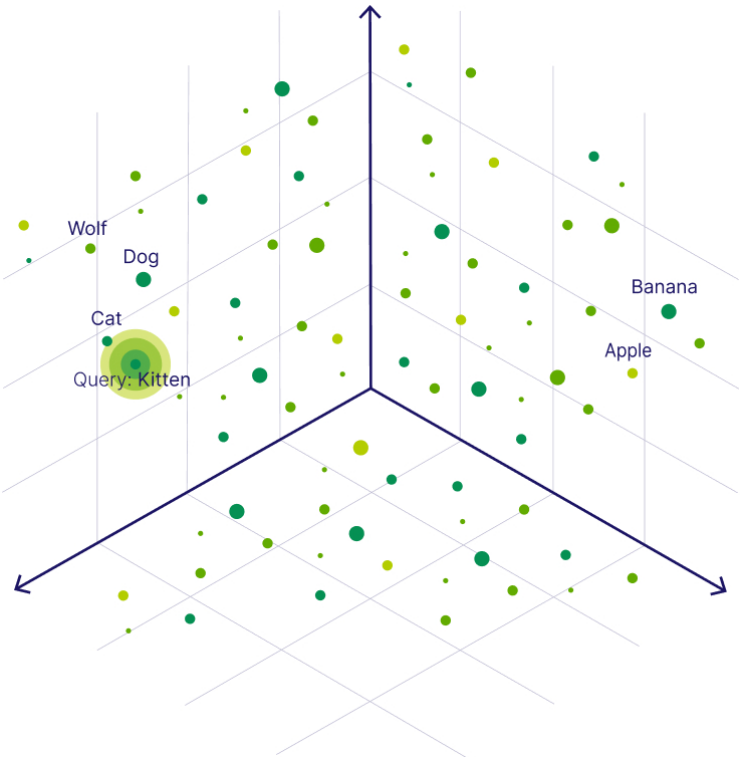
```python
from transformers import pipeline
model = pipeline("sentiment-analysis")

text1 = "We achieved science-based targets; emissions fell 22%."
text2 = "High carbon intensity persists despite investments."

print(model(text1))
print(model(text2))
```

Cosine Similarity in Word Embeddings

Concept



- Cosine similarity measures the similarity between two vectors by computing the cosine of the angle between them.
- In word embeddings, words are represented as high-dimensional vectors.
- Cosine similarity helps determine how similar two words are based on their vector representations.

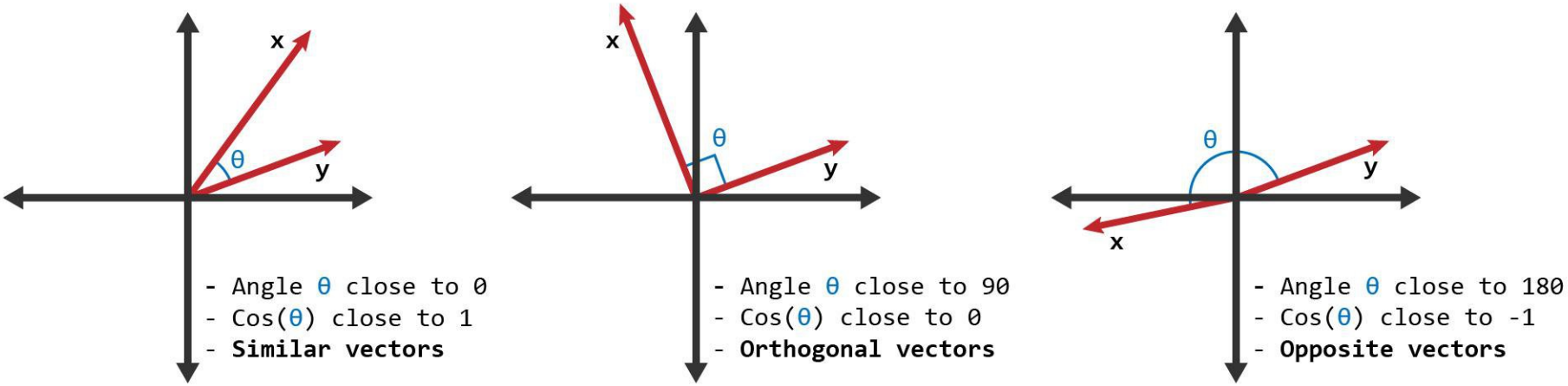
Mathematical Definition:

$$\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Cosine Similarity in Word Embeddings

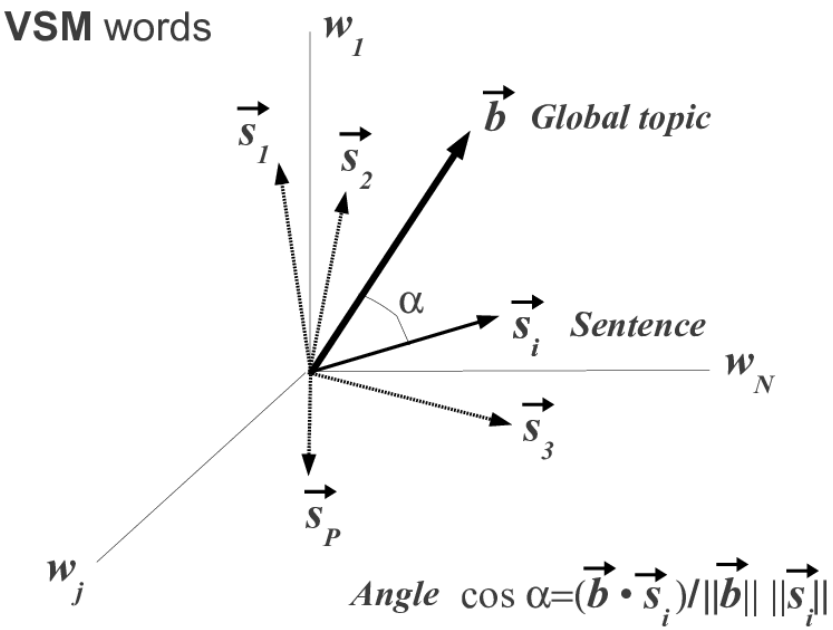
Intuition

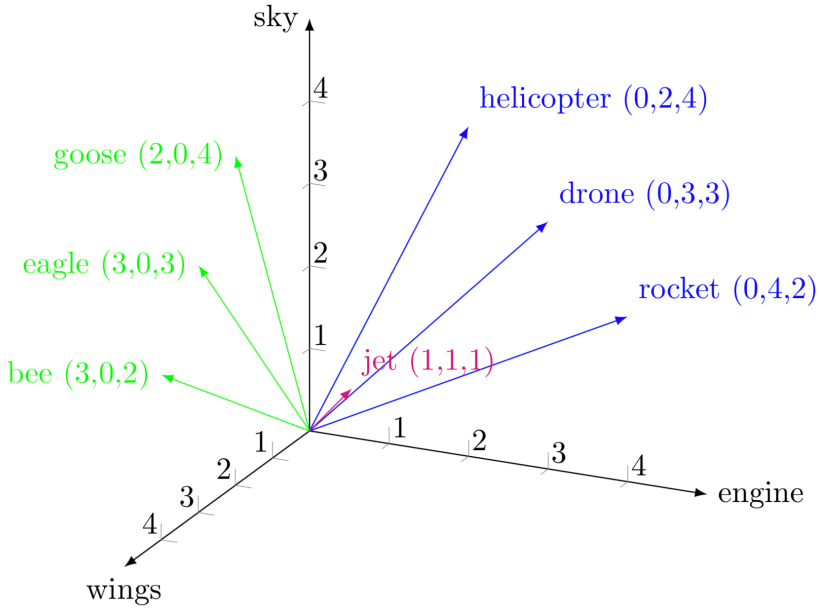
- If two words have similar meanings, their embeddings should be close in the vector space.
- Cosine similarity ranges from -1 to 1:
  - 1 indicates identical vectors (completely similar words).
  - 0 indicates orthogonal vectors (unrelated words).
  - -1 indicates completely opposite words (rare in word embeddings).
- Widely used in NLP for measuring similarity between words, sentences, and documents



Word embeddings

Cosine Similarity





Sentence Embeddings & Similarity Parsing

SBERT

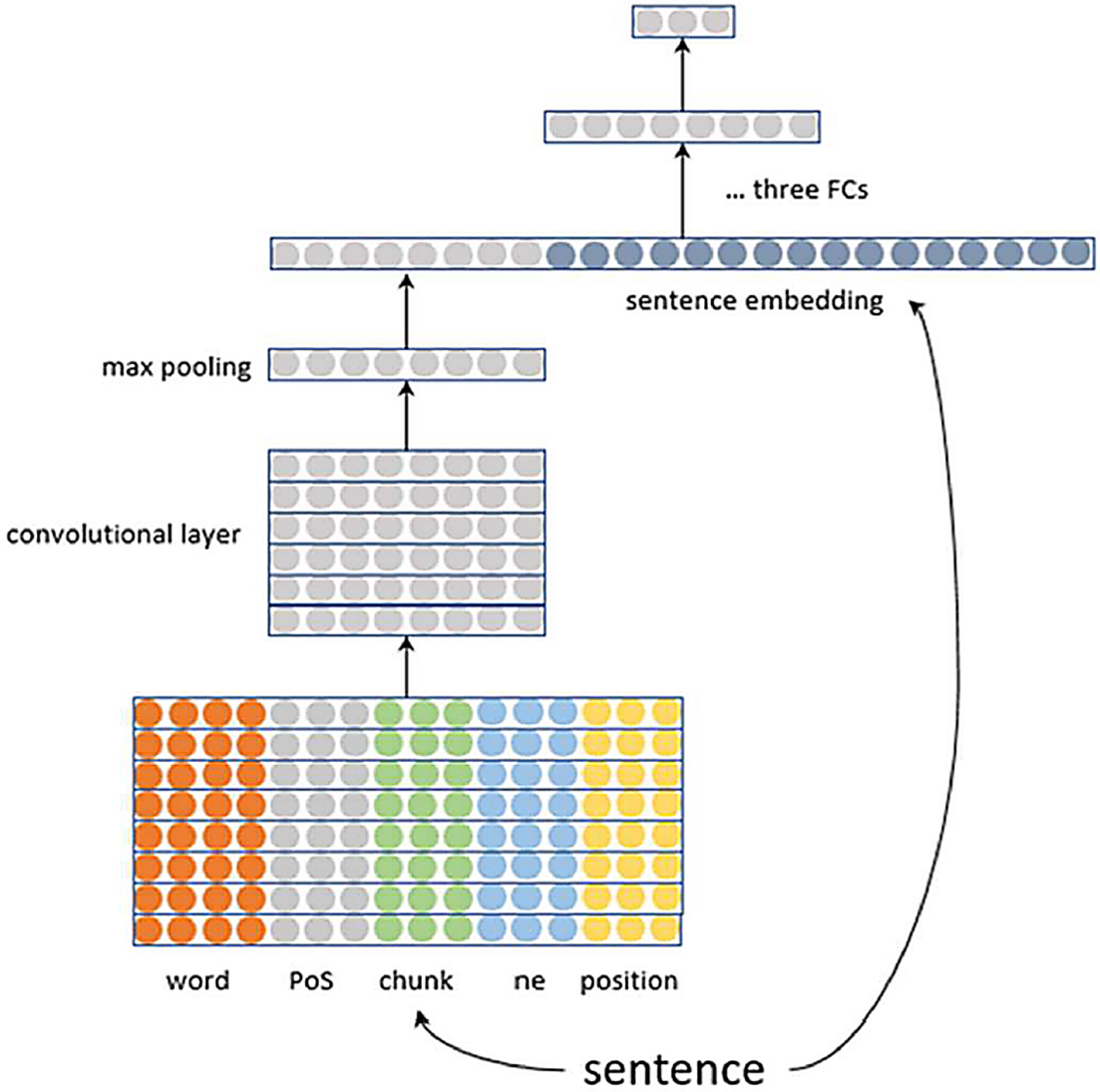
- **Why sentence-level?** ESG assertions live at sentence/paragraph granularity.
- **Contextual Meaning:** Capture semantic meaning of entire sentences, considering context and nuances beyond individual words.
- **Context-Dependent:** The same word may have different meanings depending on sentence usage → critical for nuanced ESG disclosures.

```
from sentence_transformers import SentenceTransformer, util
model = SentenceTransformer("all-MiniLM-L6-v2")

template    = "The company commits to net-zero Scope 1, 2, and
            3 emissions by 2050."
statement   = "Our goal is carbon neutral operations by 2030,
            excluding Scope 3."

emb = model.encode([template] + statement, convert_to_tensor=True)
sim = util.cos_sim(emb[0], emb[1:]).tolist()[0]
print(sim)
```

Sentence Embeddings



Comparison: Sentence Embedding **vs** Word Embedding

Feature	Sentence Embeddings	Word Embeddings
Scope	Represents entire sentences	Represents individual words
Representation	Captures overall meaning and context of sentences	Captures semantic relationships between individual words
Context Dependency	Context-dependent	Context-independent
Usage	Enhances understanding of sentence-level semantics	Focuses on semantic relationships between words
Applications	Effective for semantic search and retrieval	Suitable for tasks focusing on word-level semantics
Complexity	More computationally intensive	Less computationally intensive
Example Techniques	SBERT, Universal Sentence Encoder (USE), LSTM-based models	Word2Vec, GloVe, FastText
Advantages	Provides rich, contextually relevant information	Efficient for capturing word-level semantics
Disadvantages	Requires more resources and computation	May not capture full sentence context effectively

Sentence Correlations by Cosine Similarity

- **Data:** sustainability reports, annual reports/10-K, climate/ESG disclosures, CEO/CFO speeches, press releases, websites, CDP responses
- **Map:** **sentences** $\rightarrow$ **templates** (IFRS S2, CDP, SDGs)
- **Steps**:
  1. Curate templates: **unique** or **multiple**
     (e.g., reworded $\Rightarrow$ robustness to paraphrase $\Rightarrow$ ensemble/bootstrap)
  2. Encode (SBERT)
  3. Compute cosine similarity $\rightarrow$ threshold or Top-k
  4. Aggregate $\rightarrow$ coverage & quality dashboards
- **Output**: alignment scores, gaps, trend over time, etc...

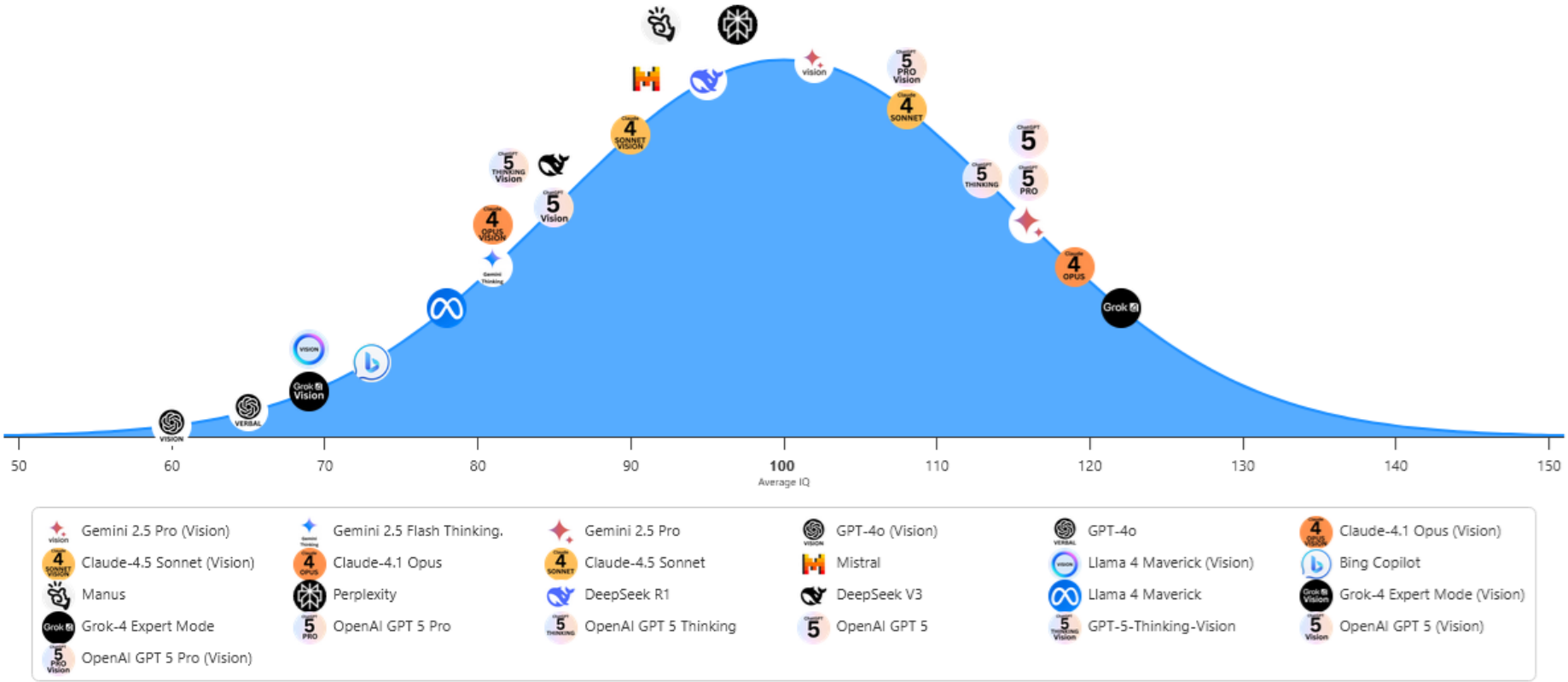
Hallucinations: Why They Happen

- Causes: Next-token prediction under uncertainty; pressure to be helpful.
  - training gaps, pressure to answer, prompt bias
- Mitigations:
  - Retrieval grounding,
  - citation checks,
  - "Don't-know" policy

How Smart Are LLMs?

A Comparison of IQ Test Results

interactive



Detecting Greenwashing with LLMs

Persona-Based Framework

Idea

- **Bhagwat et al. (2025)**: LLMs can *adopt demographic or ideological personas* (e.g., political, income, education) to simulate how diverse investors interpret identical news.
- The **persona lens** reveals systematic **disagreement** and latent **biases** in responses to firm information.

Adaptation to Greenwashing Detection

- Firms' sustainability disclosures often mix ambition with selective reporting.
- Endow an LLM with *stakeholder personas*:
  Assume: *Regulator* (compliance), *ESG analyst* (metrics), *Investor* (credibility), *NGO* (justice), *Consumer* (trust).
  Add: *age*, *gender*, *ethnicity*, *polotical views*, etc.

Detecting Greenwashing with LLMs

Persona-Based Framework

Methodological Pipeline

1. **Input:** Corporate sustainability or climate disclosure text.
2. **LLM Prompting:** "Assume the role of a [persona]. From this perspective, how credible and transparent are the firm's claims?"
3. **Elicit Scores:** Credibility score (1–10) and justification.
4. **Compute Disagreement:**

$$D_i = \sqrt{\sum_j w_j (s_{ij} - \bar{s}_i)^2}$$

5. **Interpretation:** High D_i strong divergence $\rightarrow$ potential greenwashing risk.

Detecting Greenwashing with LLMs

Persona-Based Framework

Illustrative Example

- Disclosure: "We achieved carbon neutrality through verified offsets."
  Regulator: 3/10 – weak verification.
  Investor: 5/10 – limited data.
  ESG Analyst: 4/10 – missing Scope 3.
  NGO: 1/10 – offset integrity doubtful.
  $\Rightarrow$ **High persona disagreement signals possible greenwashing.**

Persona-based LLMs quantify stakeholder disagreement: turns narrative perception into measurable greenwashing risk.

LLMs & RAG: Why this matters

Large Language Models (LLMs)

- LLMs write and reason like a well-read assistant.
- They are brilliant at producing clear text, summaries, and explanations.
- **But:** they sometimes *hallucinate*—sounding confident when facts are missing.

Retrieval–Augmented Generation (RAG)

- RAG fixes this by letting the model *look things up* first.

Net effect:

- fluent answers *grounded in real documents* with sources you can check.

LLM in one slide

- Think of an LLM as a **very advanced autocomplete**.
- It has read large amounts of text and learnt patterns of language and ideas.
- When you ask a question, it **predicts the next words** that best fit your prompt.
- This makes it great at drafting, summarising, translating, and explaining.

When LLMs shine (and when they don't)

Strengths

- **Clarity:** drafts, summaries, explanations, rephrasings.
- **Structure:** outlines, step–by–step plans, checklists.
- **Reasoning** over given text: compare, contrast, synthesise.

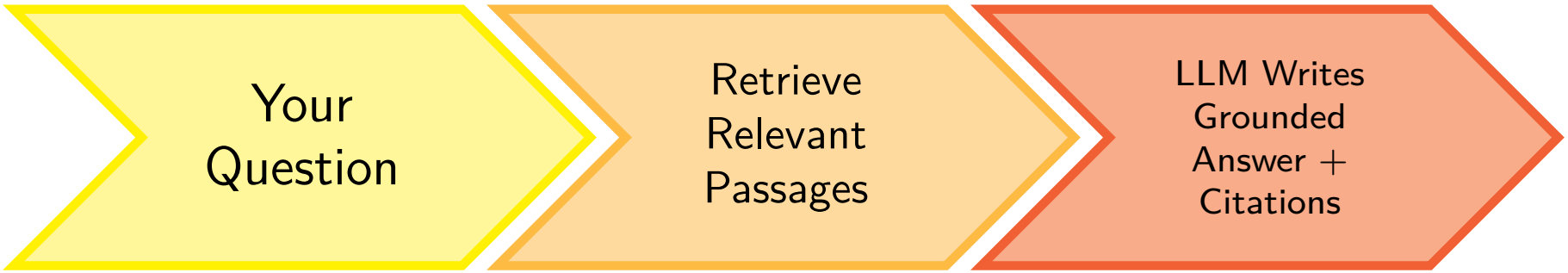
Weaknesses

- **Hallucinations:** may make up facts if not grounded.
- **Staleness:** internal "knowledge" may be out–of–date.
- **Opacity:** without sources, you cannot easily verify claims.

Retrieval–Augmented Generation (RAG)

- **Goal:** reduce hallucinations and improve trust.
- **Method:** *retrieve* relevant documents first, then let the LLM *generate* an answer using those passages.
- **Analogy:** a student checks the textbook before answering, instead of relying on memory.

Three steps you can remember



- The model is *instructed* to use only what was retrieved.
- Output includes **short quotes** and **page/section references**.

Benefits in practice

- **Accuracy:** answers are anchored in topically relevant text.
- **Transparency:** you see the *evidence* (snippets, page numbers).
- **Freshness:** you can plug in the latest PDFs, reports, or websites.
- **Compliance:** easier to audit reasoning paths and sources.

Example 1: Pulling KPIs from a PDF report

Task: "What are the company's Scope 1 emissions for the latest year?"

RAG flow:

1. Search the sustainability report for phrases like "Scope 1", "tCO_2e", "GHG inventory".
2. Retrieve **the table or paragraph** that lists numbers for each year.
3. Ask the LLM to extract **value, unit, year** and return a **verbatim quote** and **page**.

Outcome: A short, checkable answer (e.g., "*325,000 tCO_2e in FY2024*"), plus where it came from.

Example 2: Verifying a claim

Claim: "We will be net zero by 2035 across all scopes."

RAG flow:

1. Retrieve statements about "net zero", "targets", "scope coverage".
2. LLM decides: **SUPPORTS**, **REFUTES**, or **INSUFFICIENT**.
3. Include a short quote and the page/section as evidence.

Outcome: A **verdict** with **traceable reasoning**—useful for due diligence and teaching critical reading.

Grounded extraction prompts (copyable patterns)

Metric extraction (single KPI)

Instruction: "Use only the *Context*. If the metric is not explicitly reported, return `NOT FOUND`. Provide: value, unit, latest reporting year, short evidence quote (40 words), page number."

Claim verification

Instruction: "Given the claim and *Context*, decide `SUPPORTS` | `REFUTES` | `INSUFFICIENT`. Return a one–sentence rationale plus a short quote and the page reference."

Talk vs Walk

Instruction: "Label evidence as `talk` (policy/targets) or `walk` (measured outcomes/spend)."

Simple guardrails that raise quality

- "**Use only the Context**—do not rely on prior knowledge."
- "**Quote** a short passage (40 words) and include **page/section**."
- "If the exact figure isn't present, return `NOT FOUND` and say why."
- "Prefer **consolidated** totals over segments; explain any reconciliation."
- "State your **confidence** (high/medium/low) and any caveats."

From passages to a trustworthy answer

- Ask for a **clear schema**: value, unit, year, quote, page, file, confidence.
- Encourage **short evidence quotes** rather than long copies.
- Require **"NOT FOUND"** when information is missing or ambiguous.
- Save outputs in a **table** so you can scan and compare quickly.

Common issues & how to mitigate

Pitfalls

- **Ambiguous labelling:** units (t vs kt vs Mt), market– vs location–based Scope 2.
- **Boundary mismatches:** group vs subsidiary; calendar vs fiscal year.
- **Fragmented evidence:** figures split across notes, tables, and figures.

Mitigations

- Make the model **report the unit** and **the year** explicitly.
- Prefer **consolidated** totals; explain reconciliation in a *notes* field.
- Ask for a **confidence** rating and require a **verbatim quote**.

Remember this

- **LLM** = fluent assistant that writes and reasons from patterns.
- **Problem** = may hallucinate or miss recent facts.
- **RAG** = look up sources first, then answer using those passages.
- **Outcome** = answers that are both *clear* and *checkable*.

Metric extraction (single KPI)

Instruction: *"Use only the Context. If the metric is not explicitly reported, return `NOT FOUND`. Provide: value, unit, latest reporting year, a short quote (40 words), and the page number."*

Good search terms (examples): *Scope 1*, *GHG*, *tCO_2e*, *latest year*, *market–based*, *assurance*.

Claim verification (support / refute / insufficient)

Instruction: *"Given the claim and Context, decide `SUPPORTS` | `REFUTES` | `INSUFFICIENT`. Return a one–sentence rationale, a short quote (40 words), and the page reference."*

Tips: Prefer policy statements for *targets* (talk), audited tables for *measurements* (walk).

Talk vs Walk labelling

Instruction: *"Classify the evidence as `talk` (targets, policies, commitments) or `walk` (measured outcomes, actual spend, realised performance). If both appear, return both and explain briefly."*

Persuading Investors: A Video-Based Study

Hu Ma (2025). "Persuading Investors: A Video-Based Study". In: *Journal of Finance*.
DOI: 10.1111/jofi.13471

- **Main Goal:** Examine how the delivery of startup pitches – including visual cues (facial expressions), vocal tone, and verbal content – influences investors' funding decisions and subsequent startup performance.
- **Methods:** Analyze over 1,000 pitch videos, quantifying multi-modal persuasion features (visual, vocal, verbal positivity). Also conducted a controlled experiment with MBA students as investors to identify whether positive delivery sways investor beliefs or preferences.

ESGReveal: LLM + RAG for ESG Data Extraction

Zou et al. (2023). "ESGReveal: An LLM-based approach for extracting structured data from ESG reports". In: *arXiv preprint arXiv:2312.17264*

- **Main Goal:** Introduce **ESGReveal**, a framework to auto extract structured E, S, and G data from lengthy corporate sustainability reports, improving consistency and accuracy in ESG disclosures.
- **Methods:** Utilizes LLMs with RAG. **ESGReveal**'s pipeline comprises an ESG metadata module (for defining query criteria), a report preprocessing module (building a knowledge base of report content), and an LLM agent module for targeted data extraction. Evaluated on 166 ESG reports using GPT-4, achieving 77% accuracy in data extraction and 84% in ESG disclosure analysis, outperforming baseline models.

Enhancing LLMs with Climate Resources

Kraus et al. (2023). *Enhancing Large Language Models with Climate Resources*. SSRN Working Paper #4407205

- **Main Goal:** Improve the reliability and up-to-dateness of LLMs in the climate change domain by augmenting them with external **climate data resources**, addressing LLMs' lack of recent information and precision.
- **Methods:** Treats the LLM as an agent that can perform multi-source retrieval. The prototype system integrates an emissions database (ClimateWatch) and live web search into the LLM's workflow, allowing it to retrieve current, precise climate data and context. This agentic retrieval approach grounds the LLM's responses in factual data, reducing hallucinations and imprecise language in climate-related answers. Demonstrated via experiments that linking LLMs to climate databases and search leads to more accurate and trustworthy outputs for climate queries.

AI for Climate Finance: Agentic RAG for EWS Investments

Vaghefi et al. (2025). *AI for Climate Finance: Agentic Retrieval and Multi-Step Reasoning for Early Warning System Investments*. Preprint (2025)

- **Main Goal:** Enable automated tracking and classification of Early Warning System (EWS) investment finance (a climate adaptation area) by parsing heterogeneous financial reports from development banks, which traditionally lack standardized reporting for EWS projects.
- **Methods:** Developed an agent-based Retrieval-Augmented Generation pipeline that orchestrates context retrieval with internal chain-of-thought reasoning. The agent ingests unstructured project PDFs (with tables and text) to identify and categorize climate adaptation funding across defined pillars, and extracts budget allocations with evidence. Evaluated on 25 climate-project documents against multiple baselines (zero-shot classifier, fine-tuned transformer, few-shot CoT models): the agentic RAG achieved 87% accuracy (89% precision, 83% recall), significantly outperforming other methods. The study highlights the transparency and explainability gains of a tool-using glass box LLM agent over end-to-end black-box models for climate finance tracking.

ESGenius: Benchmarking LLMs on ESG Knowledge

He et al. (2025). "ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge". In: *arXiv preprint arXiv:2506.01646*

- **Main Goal:** Present **ESGenius**, the first comprehensive benchmark to evaluate and improve how well large language models understand ESG and sustainability topics, especially in question-answering tasks.
- **Methods:** (i) ESGenius-QA, a dataset of 1,136 multiple-choice questions covering a broad range of ESG and sustainability issues (questions generated by LLMs and validated by domain experts, each linked to its source text for answer grounding); and (ii) ESGenius-Corpus, a collection of 231 authoritative ESG documents (standards, frameworks, reports) used as reference material. A two-stage evaluation (zero-shot and with retrieval) was applied to 50 LLMs (0.5B to 670B parameters). Results showed that zero-shot accuracy on ESG questions is moderate (55–70%), while incorporating RAG (grounding answers in the ESG corpus) substantially improves performance (e.g., boosting a mid-sized model from 64% to 80% accuracy). This demonstrates the crucial role of grounding LLM responses in trusted ESG sources.

ESG-CID: Disclosure Content Index for GRI and ESRS

Ahmed et al. (2025). *Enhancing Retrieval for ESGLLM via ESG-CID: A Disclosure Content Index Finetuning Dataset for Mapping GRI and ESRS.* Preprint (2025)

- **Main Goal:** Facilitate retrieval-augmented ESG reporting by creating a labeled dataset that links standard disclosure requirements to actual report content, thereby helping train models to automatically map and generate ESG disclosures under frameworks like GRI and ESRS.
- **Methods:** Built the ESG-CID dataset by leveraging the "disclosure content index" sections in existing sustainability reports. Past GRI-compliant reports contain indices mapping specific GRI disclosures to report sections; these mappings (and analogous ones for the new ESRS) were extracted as weak supervision data. An LLM was used as a judge to refine and validate these query–section pairs for quality. The resulting dataset provides thousands of (query, relevant passage) examples. Using ESG-CID, the team fine-tuned and evaluated various text embedding models for ESG document retrieval. Results indicate that fine-tuned BERT-based retrievers outperform state-of-the-art commercial embeddings (including OpenAI's) on this task. Notably, these models generalize well under temporal and cross-standard evaluations (e.g. training on GRI-based data, testing on the newer ESRS format), showing the value of ESG-CID for improving ESG RAG systems.

Climate Finance Bench: QA over Climate Disclosures

Mankour et al. (2025). "Climate Finance Bench". In: *arXiv preprint arXiv:2505.22752*

- **Main Goal:** Provide an open benchmark to drive progress in question-answering on corporate climate disclosures using LLMs, ensuring that models can reliably extract and reason about key climate-related financial information from sustainability reports.
- **Methods:** Compiled Climate Finance Bench, a dataset of 33 recent corporate sustainability reports (covering all 11 GICS industry sectors) and 330 expert-curated Q&A pairs. The questions span factual extraction (e.g. specific emission figures), numerical reasoning (calculations or comparisons), and logical inference about climate strategies. Using this benchmark, the study evaluates multiple Retrieval-Augmented Generation pipelines: comparing retrieval techniques (dense neural vs. hybrid neural+BM25, with rerankers), LLM answer generators (e.g., GPT-4 and others, including quantized versions), and different prompting strategies. The evaluation finds that the retriever's ability to find the correct evidence is the primary bottleneck for QA accuracy – even advanced LLMs struggle if relevant text is not retrieved. Furthermore, the benchmark reports the carbon footprint of each configuration, highlighting trade-offs between accuracy and efficiency. Techniques like model quantization are discussed as ways to reduce compute cost. This benchmark serves as a test-bed for developing trustworthy and efficient climate QA systems grounded in corporate disclosures.

Responsible RAG for Climate Decision-Making

Juhasz et al. (2024). "Responsible Retrieval Augmented Generation for Climate Decision Making from Documents". In: *arXiv preprint arXiv:2410.23902*

- **Main Goal:** Enhance the trustworthiness and safety of LLM-powered assistants in climate policy and law by establishing a framework for responsible Retrieval-Augmented Generation (RAG) deployment, focusing on domain-specific evaluation and user assurance.
- **Methods:** Proposed a novel evaluation framework with climate-specific dimensions (e.g. factual accuracy, completeness, and hallucination rates in climate law/policy documents) to rigorously assess RAG systems. This framework was applied to a question-answering prototype that uses RAG to help policymakers query dense climate legislation and policy texts. The study introduces domain-tailored metrics and a human-annotated evaluation dataset, and uses them to compare different RAG configurations' retrieval quality and generation faithfulness. Key principles for responsible AI in this high-stakes domain are demonstrated: involve climate law experts in the development loop; use retrieval to ground answers in authoritative documents (minimizing unsupported content); implement transparency and user controls (so users can trace sources and verify answers easily). A defense-in-depth approach is advocated, combining multiple safeguards since no single technique (retrieval, filtering, etc.) catches all errors. The work provides a toolkit (including a live demo and open-source evaluation code) to catalyze safer deployment of RAG-based assistants for climate decision-makers.

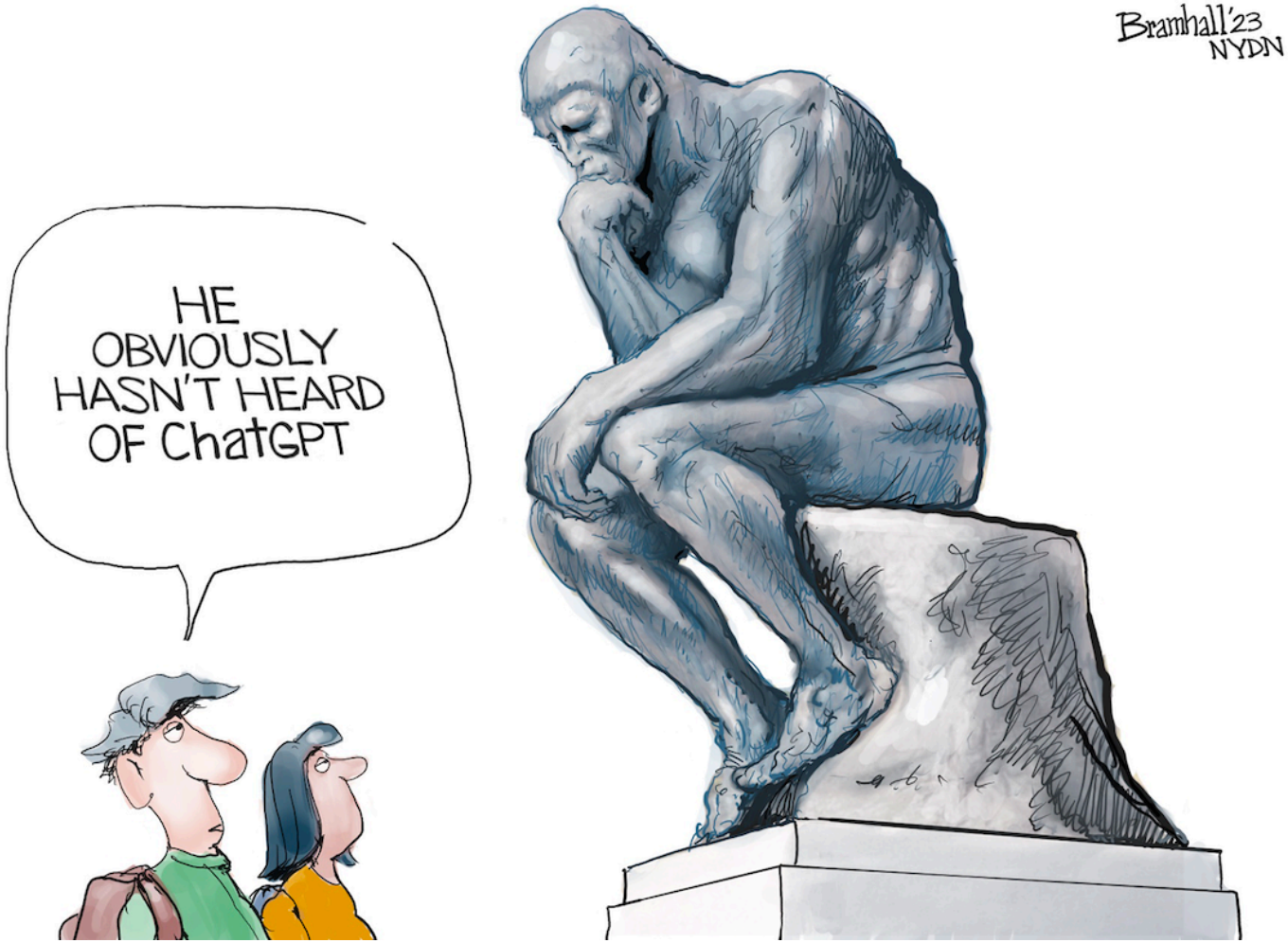
Closing Reflections: When Machines Are Watching

"You can fool some of the people all of the time, and all of the people some of the time, but you can not fool all of the people all of the time." — *Abraham Lincoln*

- The rise of AI-driven tools is reshaping how companies approach disclosure.
- Firms are now aware that machines (not just humans) are parsing every word, number, and image.
- Recent evidence shows that disclosures are becoming:
  - **Optimised for machine readability:** simplified structure, clearer headings.
  - **Strategically softened:** sentiment is subtly adjusted to avoid negative triggers.
  - Cao et al. (2023). "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI". In: *The Review of Financial Studies* 36.9, pp. 3603–3642
- Visuals and language are more curated, possibly leading to **"machine-conscious greenwashing"**.

Implications:

- AI is raising the bar on accountability, but also changing the game.
- The next challenge is **adaptive transparency**: ensuring disclosures remain truthful under algorithmic scrutiny.


Bramhall'23
NYDN
HE OBVIOUSLY HASN'T HEARD OF ChatGPT

Legal Framing - Greenwshing

Principles for trustworthy environmental and sustainability claims

1. Make accurate and truthful claims.
2. Have evidence to back up claims.
3. Do not leave out or hide important information.
4. Explain any conditions and qualifications.
5. Avoid broad and unqualified claims.
6. Use clear and easy-to-understand language.
7. Ensure that visual elements do not give the wrong impression.
8. Be direct and open about the sustainability transition.

1. Make Accurate and Truthful Claims

- **Goal:** Confirm that claims reflect actual facts and measurements.
- **Techniques:**
  - LLM-based fact-checkers with Retrieval–Augmented Generation (RAG) (e.g., Climinator).
  - Satellite and sensor validation tools (e.g., ClimateTRACE).
  - Consistency checking across time-series or multi-source data.
- **Example:** A company claims emissions have declined. AI compares reported Scope 1 values with remote-sensed emissions and historical disclosures.

2. Have Evidence to Back Up Claims

- **Goal:** Ensure claims are traceable to quantitative or referenced sources.
- **Techniques:**
  - RAG systems retrieving regulatory or peer disclosures.
  - ClimateBERT and cheap-talk classifiers for identifying unsupported language.
  - AI pipelines extracting disclosures and checking for evidence fields (e.g., emission method, assurance).
- **Example:** AI flags net-zero claims with no timeline, baseline, or emission coverage.

3. Do Not Leave Out or Hide Important Information

- **Goal:** Detect selective disclosure or material omissions.
- **Techniques:**
  - NER coverage ratio: proportion of concrete references (dates, metrics, locations).
  - Comparison with peer disclosures (topic modeling, section length).
  - TCFD/ESRS compliance mapping.
- **Example:** A fossil-fuel firm avoids discussing methane; AI detects absent keywords and skipped disclosure sections.

4. Explain Any Conditions and Qualifications

- **Goal:** Ensure disclaimers, scope, and limitations are transparent.
- **Techniques:**
  - Modal verb and hedging detection (e.g., “may”, “under certain conditions”).
  - LLM summarization prompts: “What conditions apply to this claim?”
  - Regex and syntactic parsing of qualifiers.
- **Example:** A renewable share claim is qualified by a single region; AI identifies limited scope and missing context.

5. Avoid Broad and Unqualified Claims

- **Goal:** Detect exaggerated or vague marketing language.
- **Techniques:**
  - Lexical specificity analysis (e.g., propositional idea density).
  - Greenwash text classifiers trained on vague/unsubstantiated statements.
  - Sentiment and adjective detection for promotional bias.
- **Example:** "100% sustainable operations" is flagged due to lack of numeric detail, scope, or evidence.

6. Use Clear and Easy-to-Understand Language

- **Goal:** Encourage accessible, transparent disclosure for stakeholders.
- **Techniques:**
  - Readability scoring (Flesch-Kincaid, Gunning Fog).
  - Plain language rewriting using LLMs.
  - Jargon detection and simplification prompts.
- **Example:** Passive voice and industry acronyms flagged for simplification in sustainability disclosures.

7. Ensure Visual Elements Do Not Mislead

- **Goal:** Prevent deceptive use of imagery or infographics.
- **Techniques:**
  - Vision-language models (e.g., CLIP, Flamingo) for image-scene tagging.
  - Video frame classification into green/non-green themes.
  - Layout and chart analysis for data omission (e.g., misleading y-axes).
- **Example:** An ad featuring lush forests is flagged when paired with oil expansion news.

8. Be Direct and Open About the Sustainability Transition

- **Goal:** Promote transparent, verifiable, and forward-looking transition strategies.
- **Techniques:**
  - ESGReveal and ClimateBERT for checking target and strategy sections.
  - RAG prompts: "Summarise the transition plan in 3 concrete actions."
  - Coverage testing against known disclosure frameworks (e.g., TCFD pillars).
- **Example:** AI finds that "transition" is only mentioned in vague terms, with no capex or scope 3 plan.

Comparing AI Methods Across the Eight Principles

Principle	Main AI Tools	Key Advantage
Accurate Claims	RAG, Fact-checking LLMs, Emission cross-checks	Verifies statements against data
Evidence Present	RAG, ClimateBERT, Disclosure QA agents	Flags unsupported assertions
Omission Detection	NER, topic modeling, coverage comparison	Spots silent gaps or omissions
Qualifiers Clear	Hedging detection, LLM summarization	Parses caveats, scope, or limitations
Avoid Broad Claims	Specificity scoring, Greenwashing classifiers	Detects exaggeration and vagueness
Clarity	Readability scoring, plain language rewriting	Improves transparency and accessibility
Visual Honesty	VLMs (CLIP, Flamingo), chart parsers	Flags nature-washing and misleading imagery
Transition Openness	ClimateBERT, ESGReveal, RAG Q&A	Assesses credibility of transition narratives

References I

Ahmed et al. (2025). *Enhancing Retrieval for ESGLLM via ESG-CID: A Disclosure Content Index Finetuning Dataset for Mapping GRI and ESRS*. Preprint (2025).

Cao et al. (2023). "How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI". In: *The Review of Financial Studies* 36.9, pp. 3603–3642.

He et al. (2025). "ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge". In: *arXiv preprint arXiv:2506.01646*.

Hu Ma (2025). "Persuading Investors: A Video-Based Study". In: *Journal of Finance*. DOI: 10.1111/jofi.13471.

Juhasz et al. (2024). "Responsible Retrieval Augmented Generation for Climate Decision Making from Documents". In: *arXiv preprint arXiv:2410.23902*.

Kraus et al. (2023). *Enhancing Large Language Models with Climate Resources*. SSRN Working Paper #4407205.

Mankour et al. (2025). "Climate Finance Bench". In: *arXiv preprint arXiv:2505.22752*.

Skeppstedt et al. (2024). "From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts". In: *Information Visualization* 23.3, pp. 217–238. URL: https://doi.org/10.1177/14738716241236188.

Vaghefi et al. (2025). *AI for Climate Finance: Agentic Retrieval and Multi-Step Reasoning for Early Warning System Investments*. Preprint (2025).

References II

Zou et al. (2023). "ESGReveal: An LLM-based approach for extracting structured data from ESG reports". In: *arXiv preprint arXiv:2312.17264*.