Aligning Language Model Evaluators with Human Judgement

Sharan Maiya

Department of Earth Sciences
University of Cambridge

This dissertation is submitted for the degree of
Master of Research in Environmental Data Science

Fitzwilliam College

July 2024

Declaration

This report is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text and/or bibliography. This report contains **5376** words.

Sharan Maiya
July 2024

Acknowledgements

I would like to thank my supervisors Ramit Debnath, Laura Cimoli, and Anna Korhonen for their advice and guidance through this project and their encouragement throughout these past few months. I would also like to express my gratitude to Yinhong Liu for his thoughts and suggestions.

Abstract

Large Language Models (LLM's) are capable automatic evaluators, highly suited for problems in which large datasets of text samples are evaluated on a numerical or Likert scale e.g., scoring factual accuracy or the quality of generated natural language. However, LLM's are still sensitive to prompt design and exhibit biases in such a way that their judgements may be misaligned with human assessments. *Pairwise Preference Search* (PAIRS) [43] is a search method designed to exploit LLMs' capabilities at conducting pairwise comparisons instead, in order to partially circumvent the issues with direct-scoring methods; however, this approach still relies heavily on prompt design. Here, we make use of interpretability techniques and introduce *Pairwise Preference Search with Linear Probing* (PPAIRS), which uses contrast pairs and linear probing to directly align an evaluator with human assessments. PPAIRS achieves state-of-the-art performance on text evaluation benchmarks and domain-specific problems of fact-checking and uncertainty estimation: we present an analysis of the communication of expert confidence in IPCC assessment reports from AR3 to AR6, demonstrating the effectiveness of PPAIRS as a research tool for textual uncertainty assessment in this process. We show that PPAIRS can be deployed for out-of-context reasoning tasks where ground truth labels are limited. Code and data are available at `https://github.com/maiush/PPairS` and `https://zenodo.org/records/12627714`.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

In a recent Nature Climate Action paper, Debnath et al. [18] consider how best to harness machine learning for effective action in the climate science domain, suggesting that human-in-the-loop AI systems can lead to scientific progress and better decision-making, while avoiding inequalities concerning fairness, ethics, and bias. Large Language Models (LLM's) [59, 11] are AI systems trained on vast amounts of text data to understand and generate human-like language across a wide range of tasks and domains when prompted by human users. Modern fine-tuned LLM's show impressive instruction-following capabilities [54, 17], making them ideal for human-in-the-loop AI systems, not just in the climate domain. Indeed, LLM's have recently emerged as useful tools for automatic evaluation tasks [15, 72, 73, 3] in which data are labelled on either a numerical or Likert scale.

Current models are brittle however, showing sensitivity to prompt design [76, 73] or various response-biases such as order effects [14], making them misaligned with human judges. Recently, Liu et al. [43] introduce a new method, ***Pair****wise Preference* ***S****earch* (PAIRS), to mitigate these issues by employing LLM's to provide relative comparisons between samples of a dataset, rather than absolute scores to single samples. However, their approach still relies heavily on prompting and is therefore susceptible to the same issues, albeit to a lesser degree.

An alternative approach to eliciting knowledge from an LLM via prompting is to employ interpretability techniques, in which weights or hidden-layer activations are analysed to infer properties of the LLM itself. **In this work we introduce *Pairwise Preference Search with Linear Probing* (PPAIRS), which utilises LLM interpretability methods to extract latent knowledge well-aligned with human judges for automatic evaluation tasks.** PPAIRS is a human-in-the-loop AI system, which we apply to fact-checking tasks and to investigate the communication of expert confidence in reports published by the Intergovernmental Panel on Climate Change (IPCC). In doing so, we answer the following research questions:

- Are LLM interpretability techniques superior to prompting for the task of extracting latent knowledge?
- Can such knowledge be tailored to a specific domain, given a problem setting?
- How is expert confidence communicated across different sub-domains within reports published by the IPCC?
- Can the assessment of this confidence be automated through the use of our method?

We find PPAIRS significantly improves on the current state of the art, achieving competitive performance with frontier proprietary models using smaller, cheaper, open-source alternatives. It also improves on previous best solutions to the problem of predicting expert confidence in scientific claims within the domain of climate science, following the guidelines set out by the IPCC for uncertainty communication.

Report Structure

The structure of this thesis is as follows:

- Chapter 2 introduces LLM's and PAIRS [43]. It goes on to review the growing literature on the use of language models in climate science.
- Chapter 3 presents the definitions and theory underpinning our method before outlining the algorithm itself.
- Chapter 4 is a compilation of results from three experiments. The first is a comparison between our method and the current state of the art. The second demonstrates its impressive performance on domain-specific tasks. The third is a novel application to a complex problem in the climate domain, where we investigate the communication of expert confidence in reports published during the last four assessment cycles of the IPCC.
- Chapter 5 concludes this work by summarising our main results. We discuss the key limitations of our method, and suggest directions for future research to address them.

Chapter 2

Background and Related Work

2.1 Large Language Models

Language modelling is the machine learning task of predicting the likelihood of a sequence of words (or sub-words, known as **tokens**) conditional on other sequences of words. **Autoregressive** or **causal** language modelling factors the conditional probability of a single sequence into a product for each token given its preceding tokens. The sequence $Y = \{y_1, y_2, \ldots, y_n\}$ is modelled as,

$$\begin{aligned}\mathbb{P}(y_1, y_2, \ldots, y_n) &= \mathbb{P}(y_1) \cdot \mathbb{P}(y_2|y_1) \cdot \mathbb{P}(y_3|y_1, y_2) \cdot \ldots \cdot \mathbb{P}(y_n|y_1, y_2, \ldots, y_{n-1}) \\ &= \prod_{t=1}^{n} \mathbb{P}(y_t|y_1, \ldots, y_{t-1}).\end{aligned}$$

This distribution can be learned in a semi-supervised fashion given a corpus of text for training, by considering every possible sequence of tokens and minimising the cross-entropy between a model's predicted distribution and the true distribution via gradient descent [7, 69, 59, 20].

The dominant paradigm for training such a model involves a neural network architecture known as a **transformer**. For a full derivation we refer the reader to Vaswani et al. [69] in which it was introduced, or *The Illustrated Transformer* [2] for an accessible resource. In this work we use a variant of this architecture known as a **decoder-only transformer**, which can be examined in Figure 2.1, from Anthropic's *Mathematical Framework for Transformers* [4, 22].

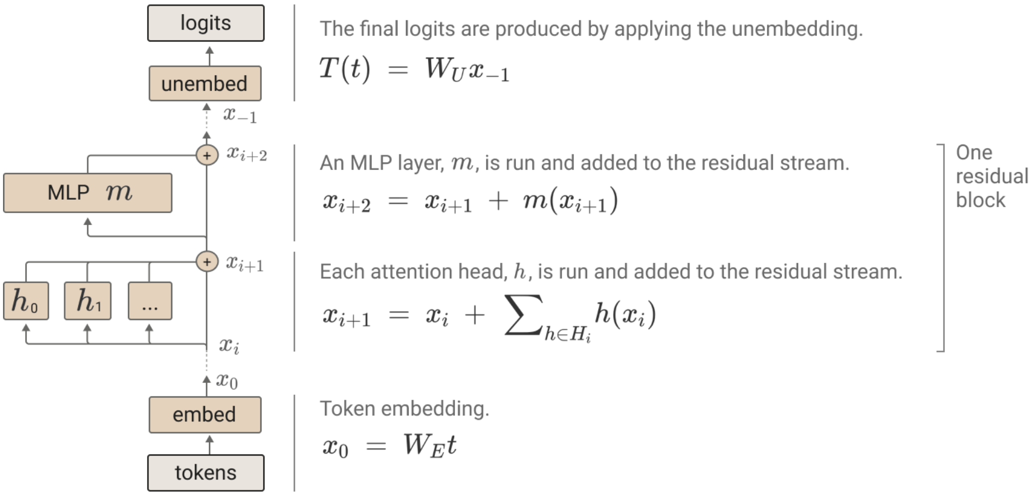

logits
unembed
x_{-1}
x_{i+2}
MLP m
x_{i+1}
h_0
h_1
...
x_i
x_0
embed
tokens
The final logits are produced by applying the unembedding.
$T(t) = W_U x_{-1}$
An MLP layer, m, is run and added to the residual stream.
$x_{i+2} = x_{i+1} + m(x_{i+1})$
Each attention head, h, is run and added to the residual stream.
$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$
One residual block
Token embedding.
$x_0 = W_E t$

Fig. 2.1 The decoder-only transformer architecture from Elhage et al. [22].

In reference to Figure 2.1, the transformer predicts the next token in a given sequence in the following manner:

1. Each individual token t is mapped to a high-dimensional **embedding** vector x_0. The space of all possible embedding vectors is the **latent space** of the model and the collection of input embedding vectors is now known as the **residual stream** (x_i).
2. The residual stream passes through a number of **residual blocks** in which:
   (a) **Attention heads** h_i "read" and process *shared* information between different tokens, and "write" this information back.
   (b) **MLP heads** m "read" and process *new* useful information, and "write" this information back.[1]
3. The final state of the residual stream x_{-1} for the last token is **unembedded** into a single, very high-dimensional vector known as the **logits**. The size of the logits is the total number of possible tokens, with each entry corresponding to a single token.
4. The *softmax* operation $\left(\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } \mathbf{z} \in \mathbb{R}^K\right)$ converts the logits into a probability distribution over all tokens. This distribution can be sampled from (in e.g., a greedy manner by choosing the index with the highest probability) in order to produce the next token. Further tokens are generated autoregressively.

[1] The function of attention and MLP heads here is described informally.

Large Language Models (LLM's) are so-called due to their number of parameters (e.g., the number of residual blocks, size of the latent space, or size of the MLP heads) and all LLM's studied in this work follow the above process, albeit with minor differences and modifications irrelevant to the methodology of PPairS. Additionally, all models we consider have been **fine-tuned** in various ways [54, 17] such that they predict tokens in a manner that endows them with the persona of a helpful assistant. The most widely known example of such a model is ChatGPT [52]. Interactions with such models are usually formatted as turn-based conversation, and the act of instructing or asking a question of an assistant-model is known as **prompting**.

2.2 Pairwise Preference Search

Recent works have achieved promising performance using LLM's for tasks of **automatic evaluation**, where data are rated on a numerical or Likert scale by prompting a given model [15, 72, 73, 3]. An example of such a task is factual evaluation, and a typical prompt for automatic evaluation takes the following form:

```
Consider the following statement:
<Statement>
Rate the factual accuracy of this statement on a scale of 1
(completely inaccurate) to 5 (absolutely correct).
```

However, for current LLM's, minor differences in prompt design can result in major differences in response [76, 73], or responses themselves can be biased in various ways [70, 61, 75].

Several recent methods aim at mitigating such issues [75, 41, 42] and this work focuses on one known as ***Pair**wise Preference **S**earch* (PairS) [43] - the approach is illustrated in Figure 2.2.

The key element of this method is the use of **pairwise comparisons** instead of directly scoring text samples. In a pairwise comparison, the evaluator is asked to consider which of two samples is "*more* F", where F is the quality on which each sample is to be evaluated. This differs from the **direct-scoring** approach of asking the evaluator to assign a score of "how F" each sample is. In computer science, several algorithms exist which utilise pairwise comparisons to sort a list of objects; PairS employs the MERGE SORT algorithm [36] to then sort samples into ascending order. In benchmark experiments, PairS is better correlated with human judgements than several other modern methods for automatic evaluation.

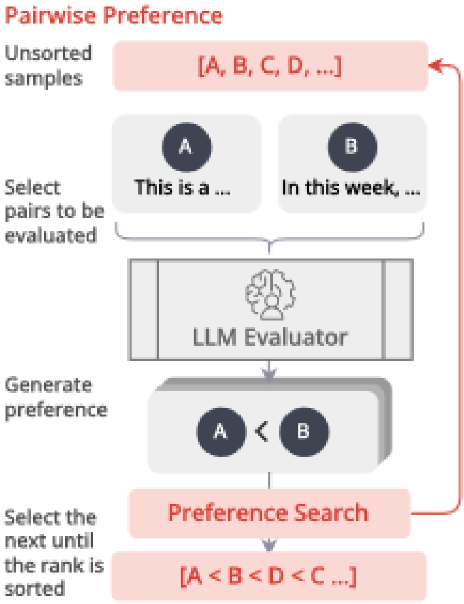


Fig. 2.2 An illustration of PAIRS from Liu et al. [43].

The output of a pairwise comparison is a **preference probability**: for a pair of text samples (A, B) to be evaluated on their factual accuracy, $\mathbb{P}(A \succ B)$ denotes the probability A is more factually accurate than B. There are several ways this could be calculated, but PAIRS uses **logit comparisons** (also known as **zero-shot** prompting) with LLM's. A typical prompt will look like,

```
Consider the following two statements:
A: <Statement A>
B: <Statement B>
Which is more factually accurate? Answers must be a single choice.
```

The logits of the LLM evaluator will have entries l_A and l_B, for the probabilities the next token is "A" or "B" respectively. Applying the softmax function to these logits results in the preference probabilities of each statement.

In direct-scoring, pairwise comparisons are performed by simply comparing scores, and are always one of $\{0, \frac{1}{2}, 1\}$, for $A \not\succ B$, $A \sim B$, and $A \succ B$, respectively.

2.3 Language Models for Climate Science

LLM’s are already used for applications in healthcare, education, and social interaction [39, 29, 63, 24, 37]. In this work we develop a novel method for extracting latent knowledge in an LLM and apply it to a setting in climate science in Chapter 4, Experiment 3. This adds to the growing body of work employing LLM’s for various problems in climate science.

CLIMATEBERT [71] is an LLM fine-tuned on a large corpus of climate-related text and released to the research community for work on downstream tasks such as Q&A systems for reliable information retrieval [67] and text classification [8] of climate-related statements in corporate documents. Others have tackled similar tasks such as automated fact-checking [40] through the use of proprietary models like GPT-4 [53]. Such methods are often dialogue-based, encouraging discussion between different LLM’s in order to reach a consensus. We examine automated fact checking in Chapter 4, Experiment 2, instead making use of smaller open-source models yet achieving higher performance due to the reliability of PPAIRS.

Related methods from the domain of natural language processing have been employed to analyse climate-related texts. In works like Jiang et al. [33], HKIMR [27], and Hu et al. [28], Latent Dirichlet Allocation is used to cluster climate-related text into topics, before performing further analysis such as identifying sentiment or links between different corpuses. Other similar approaches include joint sentiment-topic modelling [19] and word embeddings [50]. These works are largely exploratory, and we see our contribution in PPAIRS as a general method for augmenting research in any of these areas due to the ability to analyse larger datasets faster with reliable automatic evaluation.

To facilitate further work in this area, several datasets have been collected and curated. These include CLIMATEXT [68], CLIMATE-FEVER [21], and DEEP CLIMATE CHANGE [66]. We significantly increase the scope of CLIMATEXT in Chapter 4, Experiment 3 - see Chapter 6 for further details on accessing this new dataset.

More generally, it is clear that advances in all subdomains of AI can lead to opportunities for positive climate action, and recent works such as Debnath et al. [18] aim to build effective frameworks for the application of human-in-the-loop AI systems in this space which address issues of fairness, ethics, and biases.

Inspired by these frameworks, we build a human-in-the-loop AI system in PPAIRS and apply it to the automatic assessment of expert confidence in climate-related texts. This problem is only considered in the literature in a narrower setting in Lacombe et al. [38], using relatively expensive, proprietary models. We discuss this previous work and expand on it significantly in Chapter 4, Experiment 3. Additionally, to our knowledge, PPAIRS represents the first application of interpretability techniques with LLM's to problems in the climate domain.

Chapter 3

Methodology

Pairwise Preference Search with Linear Probing

In this work we introduce a method for automatic evaluation with LLM's. Our method is called PPAIRS: ***Pair**wise Preference **S**earch with Linear **P**robing*, a direct successor to PAIRS [43]. To intuit our method, we begin by defining a fundamental unit of transformer interpretability: a **feature**.

It is empirically evident that neural networks have interpretable linear directions in latent space. This has been demonstrated for word embeddings [49] and transformers [10], but also other architectures like CNNs [13, 74, 5], RNNs [35, 58], and GANs [6]. The idea that models compose linear directions in latent space to represent different concepts is informally known as the "linear representation hypothesis" [55].

There is no widely accepted formal definition of a feature, but for the purposes of this work we consider one to be any linear direction in latent space which correlates with some human-understandable concept. For PPAIRS we are particularly interested in binary features of truth or knowledge: a given text may or may not be consistent with the knowledge the LLM has learned during training, and we wish to identify a feature correlated with this property. Such a feature should identify that "the capital of France is Rome" is false while "the capital of France is Paris" is true (assuming the LLM has learned this fact).

To identify such a feature we make use of **contrast pairs** [12]. We begin with a diverse set of binary statements or questions $S = \{s_i\}_{i=1}^N$. The contrast pairs are a dataset of prompts $X = \{(x_i^+, x_i^-)\}_{i=1}^N$ constructed by appending contrasting tokens to each s_i. Suppose for example that $s_i =$ "The capital of France is Paris." A contrast pair for factual accuracy on

s_i would have x_i^+ = "The capital of France is Paris. True" and x_i^- = "The capital of France is Paris. False". Both prompts are then used as inputs to an LLM, and the final state of the residual stream on each contrasting token is harvested, leading to two embedding vectors $\phi(x_i^+)$ and $\phi(x_i^-)$.

Under the linear representation hypothesis, both $\phi(x_i^+)$ and $\phi(x_i^-)$ can be decomposed into several features, many of which are shared (since both are derived from the statement s_i). Let us write both as a linear combination of features :

$$\phi(x_i^+) = \sum_{i=1}^{n} \mathscr{F}_i^{shared} + \sum_{j=1}^{m} \mathscr{F}_i^{+} + \varepsilon^+,$$
$$\phi(x_i^-) = \sum_{i=1}^{n} \mathscr{F}_i^{shared} + \sum_{j=1}^{k} \mathscr{F}_i^{-} + \varepsilon^-,$$

with all $\mathscr{F}^{shared}$ features shared by both embeddings, $\mathscr{F}^{+/-}$ features unique to each element of the contrast pair and remaining information $\varepsilon^{+/-}$.

Contrast pairs become incredibly powerful when we consider their difference $\phi(x_i^+) - \phi(x_i^-)$, removing the effect of all $\mathscr{F}^{shared}$ and leaving only **contrastive features**. Two immediately obvious contrastive features are:

- $\Delta_{syntax} := \mathscr{F}^{True} - \mathscr{F}^{False}$, the syntactical difference in the prompts x^+ and x^-.
- $\Delta_{knowledge} := \mathscr{F}^{\top} - \mathscr{F}^{\perp}$, the logical difference between both prompts: one is true while the other is false.

Given a dataset of contrast pair differences $D = \{\phi(x_i^+) - \phi(x_i^-)\}_{i=1}^{N}$, a centering step can be performed to remove Δ_{syntax} before taking this difference:

$$\tilde{\phi}(x_i^+) := \phi(x_i^+) - \mu^+,$$
$$\tilde{\phi}(x_i^-) := \phi(x_i^-) - \mu^-,$$

where μ^+ and μ^- are the mean embedding vectors of $\{\phi(x_i^+)\}$ and $\{\phi(x_i^-)\}$ respectively.

Given ground truth labels for each pairwise comparison, we can model the probability with a classifier:

$$\mathbb{P}(x^+ \text{ true}) = \sigma(\mathbf{w}^T(\tilde{\phi}(x_i^+) - \tilde{\phi}(x_i^-))).$$

Fitting a linear classifier as above in the latent space of a neural network is often referred to as **linear probing** [1], and the probe direction vector **w** in our case corresponds to our desired binary knowledge feature. **We propose the use of linear probing through contrast pairs as a method for estimating the preference probability in PAIRS**, naming our approach *Pairwise Preference Search with Linear Probing* or PPAIRS. The full approach of this method is therefore:

1. Begin with a dataset of text samples $T = \{t_i\}_{i=1}^{N}$ to be automatically evaluated along a given dimension F e.g., factual accuracy.
2. Convert T to a new dataset of pairwise comparisons $P = \{(t_i, t_j)\}$, with the goal of evaluating which element in a given pair is “more F”.
3. Produce a dataset of contrast pairs $X = \{(x_{i,j}^{+}, x_{i,j}^{-})\}$: prompts in which either t_i or t_j is stated to be “more F”.
4. Construct a dataset $D = \{\tilde{\phi}(x_{i,j}^{+}) - \tilde{\phi}(x_{i,j}^{-})\}$ of centered embedding vector differences for each contrast pair.
5. Train a linear probe on D using ground truth labels to classify which element in a contrast pair is true, thereby determining a preference probability.
6. Proceed to sort T using these preference probabilities as in PAIRS.

We demonstrate in Chapter 4 that PPAIRS leads to significantly higher accuracy on pairwise comparisons against ground truth human judgements than the original PAIRS method.

Chapter 4

Experiments

4.1 Experiment 1: Text Evaluation Benchmarks

Before demonstrating the effectiveness of PPAIRS for domain-specific tasks, we first quantify its improvement in performance over previous state-of-the-art methods using benchmark datasets. We replicate the same experiments reported in Liu et al. [43] to allow for a direct comparison.

4.1.1 Experimental Set-Up

We use the NEWSROOM [26] and SUMMEVAL [23] datasets, which both consist of news articles collected over several years. Each article is tagged with several machine-generated text summaries, which are each then evaluated by multiple human judges. Four dimensions along which these summaries are evaluated are coherence (CH), consistency (CON), fluency (FLU), and relevance (RE), which are rated on a scale from 1 (very low) to 5 (excellent). We seek to perform the automatic evaluation task of assigning ratings to summaries which are aligned with those of the human judges. For details on metrics, models, baselines, and the full prompts used, see Appendix A.

4.1.2 Results

Classification accuracy for pairwise comparisons using PPAIRS is significantly higher than using either PAIRS or direct-scoring, as shown in Table 4.1. This is true for both datasets, all four aspects, and all three models. It is especially interesting to note that for both direct-scoring and PairS, we see higher accuracy when evaluating Llama-3 as opposed to Llama-2 or Mistral, while the performance of PPairS is roughly the same regardless of the LLM used. This hints that PPairS may be at or approaching an upper-bound on the knowledge representations of LLM's at this size, likely due in part to the supervised setting of linear probing.

Model	NEWSROOM				SummEval			
	CH	CON	FLU	RE	CH	CON	FLU	RE
Mistral-7B-Instruct-v0.1								
Direct-Scoring	0.57	0.55	0.58	0.56	0.56	0.84	0.77	0.58
PairS	0.52	0.44	0.51	0.53	0.58	0.55	0.58	0.59
PPairS	**0.81**	**0.81**	**0.79**	**0.79**	**0.83**	**0.92**	**0.86**	**0.81**
Llama-2-7b-chat								
Direct-Scoring	0.57	0.57	0.55	0.57	0.55	0.84	0.78	0.58
PairS	0.64	0.63	0.65	0.68	0.58	0.67	0.60	0.59
PPairS	**0.80**	**0.82**	**0.79**	**0.80**	**0.86**	**0.92**	**0.87**	**0.82**
Llama-3-8b-Instruct								
Direct-Scoring	0.66	0.70	0.64	0.69	0.68	0.77	0.69	0.64
PairS	0.71	0.72	0.72	0.71	0.63	0.59	0.56	0.63
PPairS	**0.79**	**0.81**	**0.79**	**0.79**	**0.85**	**0.91**	**0.86**	**0.82**

Table 4.1 Accuracy [0, 1] on pairwise comparisons for our benchmark experiments. In all cases PPairS substantially improves on direct-scoring or PairS.

After sorting the original text summaries using all three methods, we again see significantly better performance in all settings when using PPairS in Table 4.2. In several cases, the correlation with human judgements when using PPairS approaches or even exceeds that of direct-scoring or PairS when using GPT-4-Turbo, the current frontier proprietary LLM [44, 16]. This is not only a massive inference cost and hardware cost reduction for performance on automatic evaluation tasks, this represents a significant cost reduction to the user when factoring in current API pricing for models like GPT-4-Turbo (as of writing, US$30 / 1M generated output tokens [51]). With PPairS, the resource constraints of many works cited in Section 2.3 and similar research could be loosened significantly.

Model	NEWSROOM				SummEval			
	CH	CON	FLU	RE	CH	CON	FLU	RE
Mistral-7B-Instruct-v0.1								
Direct-Scoring	0.32	0.20	0.26	0.39	0.23	0.37	0.19	0.19
PairS	0.55	0.48	0.48	0.53	0.28	0.30	0.24	0.27
PPairS	**0.59**	**0.59**	**0.61**	**0.55**	**0.74**	**0.39**	**0.45**	**0.68**
Llama-2-7b-chat								
Direct-Scoring	0.02	-0.02	0.01	0.14	0.12	0.18	0.06	0.11
PairS	0.43	0.37	0.28	0.43	0.17	0.31	0.18	0.24
PPairS	**0.57**	**0.60**	**0.57**	**0.52**	**0.72**	**0.36**	**0.44**	**0.64**
Llama-3-8b-Instruct								
Direct-Scoring	0.21	0.25	0.07	0.25	0.35	0.19	0.14	0.19
PairS	0.43	0.44	0.38	0.30	0.25	0.26	0.14	0.29
PPairS	**0.59**	**0.68**	**0.57**	**0.56**	**0.67**	**0.38**	**0.45**	**0.64**
GPT-4-Turbo-preview								
Direct-Scoring	0.55	0.57	0.60	0.54	0.44	0.46	0.42	0.53
PairS	0.64	0.67	0.60	0.61	0.53	0.47	0.48	0.59

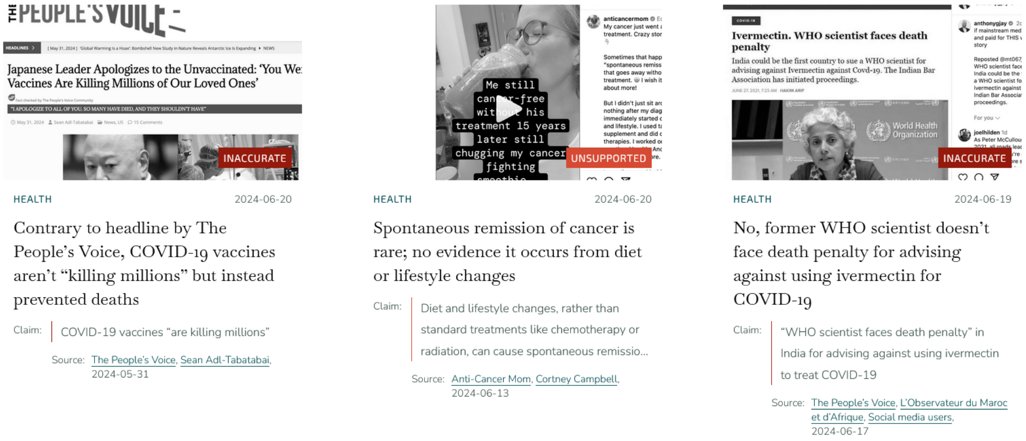
Table 4.2 Spearman's rank correlation [-1, 1] against ground truth sorted scores, after sorting using pairwise comparisons via different methods. Correlations for Mistral, Llama-2, and GPT-4-Turbo using direct-scoring and PairS are lifted from Liu et al. [43]. In all cases, PPairS outperforms both methods. In many cases PPairS is comparable and even exceeds performance with GPT-4-Turbo.

4.2 Experiment 2: Fact Checking

We now consider more domain-specialised tasks, where we feel PPairS holds the most promise, beginning at a natural starting point: fact checking.

4.2.1 Experimental Set-Up

Science Feedback [62] is an online publication which aims to verify the credibility of claims often cited as scientific fact in the domains of *climate*, *health*, and *energy*. These are evaluated by panels of domain-experts, and tagged with their credibility. A typical page is shown in Figure 4.1.



HEALTH 2024-06-20

Contrary to headline by The People's Voice, COVID-19 vaccines aren't "killing millions" but instead prevented deaths

Claim: COVID-19 vaccines "are killing millions"

Source: The People's Voice, Sean Adl-Tabatabai, 2024-05-31

HEALTH 2024-06-20

Spontaneous remission of cancer is rare; no evidence it occurs from diet or lifestyle changes

Claim: Diet and lifestyle changes, rather than standard treatments like chemotherapy or radiation, can cause spontaneous remissio...

Source: Anti-Cancer Mom, Cortney Campbell, 2024-06-13

HEALTH 2024-06-19

No, former WHO scientist doesn't face death penalty for advising against using ivermectin for COVID-19

Claim: "WHO scientist faces death penalty" in India for advising against using ivermectin to treat COVID-19

Source: The People's Voice, L'Observateur du Maroc et d'Afrique, Social media users, 2024-06-17

Fig. 4.1 A typical page on SCIENCE FEEDBACK [62]. Claims are tagged with their scientific credibility (coloured boxes). We manually merge these into three categories: inaccurate, misleading, and accurate. For example, the above "unsupported" claim is re-labelled as inaccurate for our analysis.

We scrape all claims in all three domains from SCIENCE FEEDBACK and their credibility tags. The total number of unique claims in the domains of climate, health, and energy are 161, 722, and 11, respectively. These are manually binned into three categories: inaccurate, misleading, and accurate. The classification task of identifying the correct tag for a given claim is converted into a pairwise comparison task by prompting an LLM for the "more factually accurate / less ambiguous" claim. Details on metrics, models, baselines, and prompts are included in Appendix B.

4.2.2 Results

Model	Domain		
	Climate	Health	Energy
Direct-Scoring	0.82	0.86	**0.91**
PAIRS	0.58	0.56	0.56
PPAIRS	**1.00**	**1.00**	0.88 / 0.87

Table 4.3 Accuracy [0, 1] for ground truth pairwise comparisons. For the Climate and Health domains, PPAIRS elicits perfect classification accuracy. For the Energy domain, we train a probe on either the Climate (red) or Health (blue) domain statements, and find they transfer with competitive accuracy, hinting that our probes may be locating a more general knowledge feature than the domain of their training data.

Classification accuracy for pairwise comparisons is shown in Table 4.3. Across all three domains, we notice higher accuracy for direct-scoring than PAIRS. Such a gap is not evident in analogous results for Experiment 1 (Table 4.1). We offer an explanation for this observation in Appendix B.1.

We observe 100% classification accuracy of PPAIRS in both the climate and heath domains. Perfect generalisation like this suggests that the language model used may have been trained on data from the SCIENCE FEEDBACK publication, but it should be noted that neither direct-scoring nor PAIRS led to perfect accuracy. It appears that PPAIRS has allowed us to elicit the model's true capability in terms of knowledge.

Consider now the interesting results for the energy domain. Due to the small number of unique claims, the training of a linear probe is met with difficulty. We instead take this opportunity to examine the cross-domain generalisation of probes trained on pairwise comparisons from either the climate or the health domains, and find they generalise well to the energy domain, achieving classification accuracies of 88% and 87% respectively. While direct-scoring leads to higher accuracy, we find this result noteworthy as it suggests that the linear probes trained may align with a general knowledge feature of the model representations, as opposed to knowledge only within the particular domain on which the probe was trained. While this is speculation based on a single early result, we strongly feel it deserves further investigation in future work.

4.3 Experiment 3: Expert Confidence

4.3.1 Expert Confidence in the IPCC Reports

The Intergovernmental Panel on Climate Change (IPCC) is the United Nations body for assessing and synthesising the current understanding of various problems in climate science [65]. It is comprised of three *Working Groups* and a *Task Force*, which work to produce reports[1] summarising relevant scientific literature over several year-long *assessment cycles*, of which there have been six. Since the third of these cycles, authors have followed guidelines to, where appropriate, assign confidence levels to claims, with the aim of representing *"collective judgement in the validity of a conclusion based on observational evidence, modeling results, and theory"* [30]. Confidence is measured on a 5-point Likert scale with tags *very low*, *low*, *medium*, *high*, and *very high*.

[1] https://www.ipcc.ch/reports/

The IPCC reports cover a wide range of sub-domains and it is interesting to consider how confidence has been communicated across them since these guidelines were introduced. Such analysis could lead to insights on over- or under-confident areas of scientific research, or patterns between different topics.

Topic Modelling

We set out to conduct this analysis on all longform working group and synthesis reports since 2001: a total of 17 publications. Each is preprocessed and all tagged claims are extracted by searching for each possible tag e.g., *medium confidence*. We cluster these tags into 29 distinct topics e.g., “forests” and “aerosols” - our methodology for doing so can be found in Appendix C.

Among these 29 topics we find the numbers represented in the third, fourth, fifth, and sixth assessment cycles are 14, 11, 26, and 29 respectively. Rather than an indication of an increase in diversity of topics covered in more recent assessment cycles, we instead believe this is due to the fact the confidence tagging framework was not adopted uniformly during the third and fourth assessment cycles. Indeed in both cases only the synthesis and Working Group II reports adopt this framework, leading to an unbalanced distribution of topics mostly covering the impacts of climate change on human and natural systems (the focus of Working Group II) as opposed to the physics of climate systems (Working Group I) or climate change mitigation (Working Group III) [65].

We assign a single score to capture how uncertainty is communicated for a given topic by mapping the scores -2, -1, 0, 1, and 2 to the confidence tags *very low*, *low*, *medium*, *high*, and *very high* respectively, and calculating a weighted average using the proportion of claims within a given topic for each tag. The full distribution of these scores across all topics during each assessment cycle is shown in Figure 4.2.

What is most apparent in Figure 4.2 is that across all topics in all reports the confidence of claims skews high. This is especially true during the most recent assessment cycle, with no scores lower than ∼0.4. While this may simply reflect the scientific literature - it is intuitively more likely to find published science of higher certainty [45, 31] - it indicates a potential misalignment with the goals of the IPCC. If reports are intended to be a synthesis of the current state of climate science it may be helpful and even necessary to consider and discuss low confidence yet high risk topics. While such discussions must obviously be handled with care, we feel one group best prepared to do so is the IPCC.

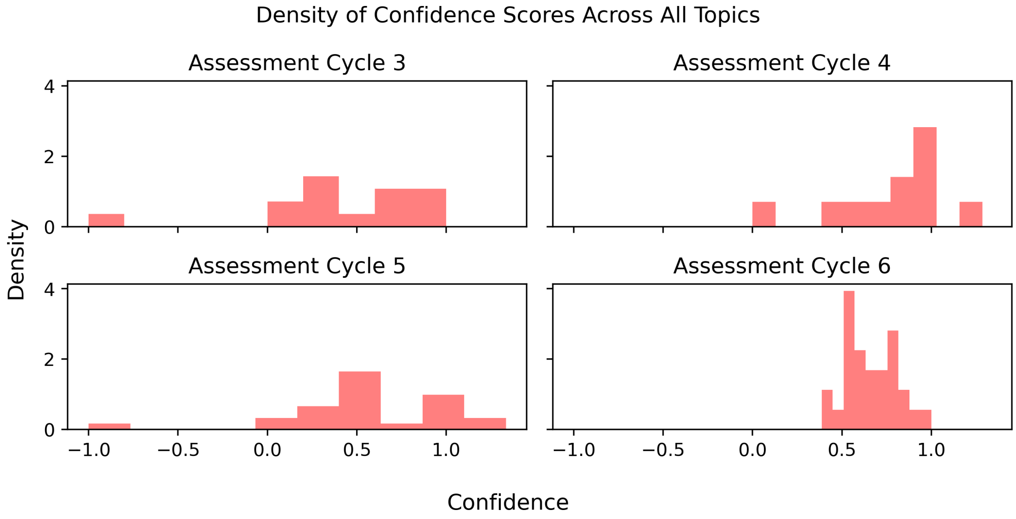


Fig. 4.2 Distribution of confidence scores across all topics during each assessment cycle. Positive indicates higher confidence while negative indicates higher uncertainty.

In order to investigate the distribution of confidence tags across different topics themselves, we perform a UMAP [46] projection of all generated context embeddings for each assessment cycle to two dimensions, visualising the projections in Figure 4.3.

Each context (with its central claim) in Figure 4.3 is colour-coded with its confidence tag. Due to the UMAP projection, contexts under the same or similar topics are likely to be clustered nearer to each other in each sub-figure, yet we find no clear patterns in the assignment of confidence tags. Instead, we again note two findings:

- The higher relative proportion of *very high* confidence claims in the sixth assessment cycle suggests authors opted to focus on well-understood ideas.
- A strong sample-imbalance exists over time. This complicated analysis due to confounding by the presence (or lack thereof) of tagged claims for a given topic in earlier reports.

On the latter above point, note it is not necessarily the case that earlier reports of the IPCC discussed fewer topics than more recent work, rather that certain topics covered only in publications from Working Groups I and III are not written using the same confidence tagging framework. If it were possible to not only extract claims from these earlier reports but tag them automatically, a more complete analysis of this kind would be possible. We believe this is exactly the type of problem PPairS is well-suited for, and shift our focus now to tackling it.

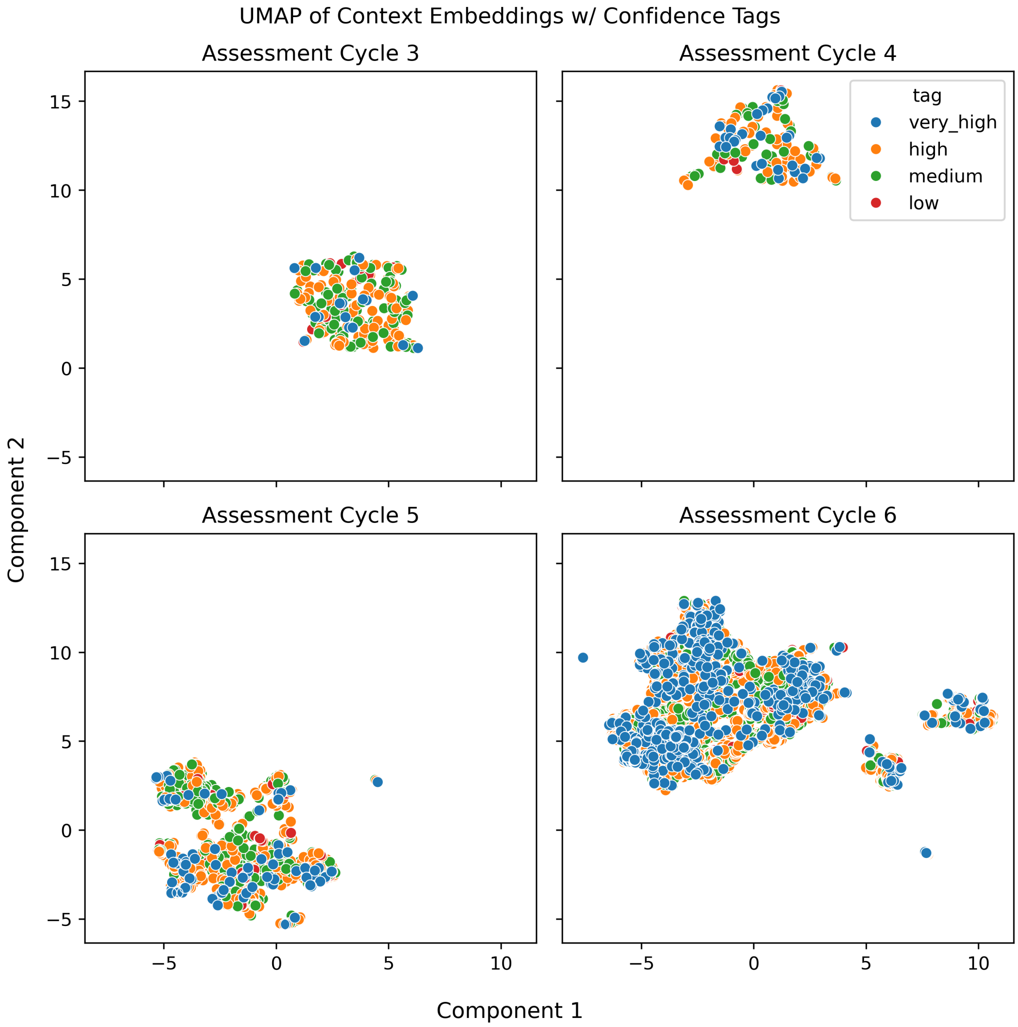


Fig. 4.3 UMAP [46] projection of all context embeddings to two dimensions. Each context (a single point) is colour-coded with the confidence tag of its main claim.

4.3.2 Experimental Set-Up

This problem set-up is considered in Lacombe et al. [38] where the CLIMATEX dataset is introduced. 8094 claims from reports during the sixth assessment cycle of the IPCC are extracted and tagged in a direct-scoring method using three proprietary LLM's. Of the three, GPT-4 is shown to perform best, achieving an accuracy of 44.3% and 47.0% in zero-shot and few-shot settings respectively. This poor accuracy obtained with GPT-4 is unsatisfying: we propose instead to make use of relatively smaller, open-source models and PPAIRS. In doing so we extend the CLIMATEX dataset by including all working group and synthesis reports since the third assessment cycle of the IPCC (resulting in a total of 10065 claims), and set up a pairwise comparison task of prompting an LLM to compare two claims and assess which is “expressed with higher confidence”. Further details on data, metrics, models, baselines, and prompts are provided in Appendix C.

4.3.3 Results

We obtain an accuracy of 72.7% on pairwise comparisons between claims using a trained linear probe. This classifier closely aligns with the first principal component of the contrast pair differences - we visualise this in Appendix C.5.

We then use a modified binary search algorithm we call PPAIRS_INSERT, to automatically assign confidence tags to a held-out test set of unseen claims. The algorithm is described in detail in Appendix D. The accuracy of this method is shown in Table 4.4.

Assessment Cycle				
Third	Fourth	Fifth	Sixth	**Overall**
0.69	0.67	0.64	0.60	0.64

Table 4.4 Classification accuracy [0, 1] of PPAIRS_INSERT assigning confidence tags to unseen claims in reports from each assessment cycle, and an overall accuracy.

We find the use of PPAIRS_INSERT to result in an overall accuracy of 64% when assigning a confidence tag to unseen claims from IPCC reports. For the sixth assessment cycle in particular, we achieve an accuracy of 60%: a significant improvement over 47% achieved by GPT-4 in a few-shot setting from Lacombe et al. [38]. As with Experiment 1, we find that smaller, cheaper, open-source models can be used to obtain better performance than proprietary models like GPT-4 through the use of latent knowledge in model internals.

While PPAIRS does lead to this higher accuracy, we feel an overall accuracy of 64% is not high enough to justify its use to extend the topic analysis presented earlier, due to the potential for claims to be tagged incorrectly. With the impressive performance of PPAIRS in Experiments 1 and 2 in mind, it is likely that the nature of the pairwise comparison problem set up in this experiment is inherently difficult (a sample of three non-expert humans achieved an accuracy of 36.3% in Lacombe et al. [38]). In order to probe this question further, it would be prudent to examine probe performance and the performance of PPAIRS_INSERT in more detail through the topics identified previously, with the aim of identifying particular sub-domains in which PPAIRS performs well or poorly.

Chapter 5

Discussion and Conclusion

LLM's appear in many ways to be perfect tools for automatic evaluation. The size and diversity of their training data leads to impressive knowledge in many domains; their instruction-following ability makes them ideal tools for human-in-the-loop AI systems; and their capability to quickly process batches of text results in a significant speed-up compared to human evaluators. Current LLM evaluators are however sensitive to prompt design or other biases, leading to misalignment with human judges. PAIRS, introduced in Liu et al. [43], addresses many of these issues by exploiting LLMs' capability at pairwise comparisons over direct-scoring, improving on the performance of zero-shot prompting.

We build on this work by extracting latent knowledge in LLM internals using linear probes, establishing state-of-the-art performance on a range of tasks in Chapter 4 with our method, PPAIRS. This is demonstrated first on standard benchmark tasks in Experiment 1, but we believe the true promise of PPAIRS lies in domain-specific tasks. Indeed, we use it in Experiment 2 to achieve perfect classification accuracy on tasks in the domains of climate and health science when prompting methods do not. We finally go on to show PPAIRS improving on the performance of prompting with frontier proprietary models in Experiment 3, on a complex out-of-context reasoning task in the climate domain of assessing expert confidence.

It should be noted that one reason for PPAIRS' impressive performance is its supervised nature: we directly align with human judgements using a relatively small number of labels. In some domains such as frontier science or complex decision-making it may be less feasible or even impossible to obtain accurate ground-truth data; research into unsupervised variations may be more future-proof for such domains.

The supervised nature of our method also implies that the directions identified by trained linear probes may not hold causal influence over the predictions of the LLM itself. We feel this is an important consideration and would be excited to see more work on ablating probe directions from the latent space and noting performance changes. If the LLM's "true" knowledge feature is identifiable it may even be possible to use such a feature to steer model behaviour [64].

A related question concerns the generality of such a knowledge feature. In Experiment 2 we find some evidence that the direction identified by a given linear probe may align with the LLM's knowledge *in general*, as opposed to within the particular domain on which the probe was trained. This should be more thoroughly investigated in order to build a theory of exactly how knowledge and belief is represented within the internals of LLM's.

The work in Experiment 3 highlights that the performance of PPairS is of course fundamentally limited by the representations of the LLM it is used with. This is especially the case for complex, broad domains such as climate science. We believe it is particularly important to investigate probe performance in these domains through an exploration of sub-topics. In the example of Experiment 3, such work could identify particular topics where the LLM's knowledge appears to be lacking, informing future priorities with regards to pretraining or fine-tuning to address these gaps in knowledge. We feel this method is in some ways a better approach to establishing the limitations of a given LLM's knowledge than prompting alone. It should be noted however that PPairS still requires the prompting of an LLM, and while addressing many of the above stated issues with direct-scoring methods, it may still be sensitive to prompt design in less obvious ways.

The most significant limitation of PPairS is by design: access to model internals is required to fit a linear probe. Current frontier models such as GPT-4 are proprietary and do not provide access to such internals, resulting in a gap between the practical and theoretical maximum performance of PPairS. We hope that as the research community continues to develop more powerful open-source models such as Meta's Llama series [48], this gap will decrease.

Fundamentally, through this work and the avenues for future research described above, we hope to make progress on a theory of LLM knowledge representations. Such a theory could better equip practitioners with tools for eliciting the true capabilities of LLM's for research and decision-making tasks, building powerful human-in-the-loop AI systems which accelerate the progress of science on the whole, including complex domains such as climate. We look forward to new work in this direction.

Chapter 6

Code and Data Availability

PPairS is implemented as a Python package and released under an MIT License. It can be installed via `pip` through the GitHub repository at `https://github.com/maiush/PPairS`. Full implementations of all experiments in Chapter 4 are provided. All data downloaded, processed, and produced - including harvested LLM activations - are archived at `https://zenodo.org/records/12627714`, where links to original datasets are also included should users wish to download these from source.

References

[1] Alain, G. and Bengio, Y. (2017). Understanding Intermediate Layers Using Linear Classifier Probes. In *5th International Conference on Learning Representations (ICLR) Workshop Track*.

[2] Alammar, J. (2018). The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer/. Accessed: June 2024.

[3] Alvarez, R. M. and Morrier, J. (2024). Evaluating the Quality of Answers in Political Q&A Sessions with Large Language Models. *arXiv preprint arXiv:2404.08816*.

[4] Anthropic (2021). https://www.anthropic.com/. Accessed: March 2024.

[5] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6541–6549.

[6] Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.

[7] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

[8] Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N. (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

[9] Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA. NLTK Version 3.7.

[10] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

[11] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford,

A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[12] Burns, C., Ye, H., Klein, D., and Steinhardt, J. (2022). Discovering Latent Knowledge in Language Models Without Supervision. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22245–22258. Curran Associates, Inc.

[13] Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M., and Olah, C. (2020). Curve Detectors. *Distill*. https://distill.pub/2020/circuits/curve-detectors.

[14] Chen, X., Chi, R. A., Wang, X., and Zhou, D. (2024). Premise Order Matters in Reasoning with Large Language Models. *arXiv preprint arXiv:2402.08939*.

[15] Chen, Y., Wang, R., Jiang, H., Shi, S., and Xu, R. (2023). Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Atlanta, Georgia. Association for Computational Linguistics.

[16] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv preprint arXiv:2403.04132*.

[17] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[18] Debnath, R., Creutzig, F., Sovacool, B. K., and Shuckburgh, E. (2023a). Harnessing human and machine intelligence for planetary-level climate action. *npj Climate Action*, 2(1):20.

[19] Debnath, R., Ebanks, D., Mohaddes, K., Roulet, T., and Alvarez, R. M. (2023b). Do fossil fuel firms reframe online climate and sustainability communication? A data-driven analysis. *npj Climate Action*, 2(1):47.

[20] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[21] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. (2021). CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. *arXiv preprint arXiv:2012.00614*.

[22] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

[23] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation.

[24] Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., and Wicks, P. (2023). Large language model AI chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398.

[25] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[26] Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

[27] HKIMR (2022). A Machine Learning Based Anatomy of Firm-Level Climate Risk Exposure. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4265246. Available at SSRN.

[28] Hu, W., Li, K., and Yu, T. (2022). Towards Objectivity and Interpretability: An Anatomy of Climate Change Exposure. Available at SSRN.

[29] Hwang, G.-J. and Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 31(7):4099–4112.

[30] IPCC (2001). Third Assessment Report. https://www.ipcc.ch/reports/?rp=ar3. Accessed: April 2024.

[31] Janzwood, S. (2020). Confident, likely, or both? The implementation of the uncertainty language framework in IPCC special reports. *Climatic Change*, 162(3):1655–1675.

[32] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

[33] Jiang, Y., Song, X., Harrison, J., Quegan, S., and Maynard, D. (2017). Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 25–30. Association for Computational Linguistics.

[34] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, 2 edition.

[35] Karpathy, A., Johnson, J., and Fei-Fei, L. (2015). Visualizing and Understanding Recurrent Networks. In *International Conference on Learning Representations (ICLR) Workshop*.

[36] Knuth, D. E. (1998). *The Art of Computer Programming*, volume 3. Addison-Wesley, Reading, Massachusetts.

[37] Kocielnik, R., Prabhumoye, S., Zhang, V., Jiang, R., Alvarez, R. M., and Anandkumar, A. (2023). BiasTestGPT: Using ChatGPT for Social Bias Testing of Language Models. *arXiv preprint arXiv:2302.07371.*

[38] Lacombe, R., Wu, K., and Dilworth, E. (2023). ClimateX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements? *arXiv preprint arXiv:2311.17107.*

[39] Lee, P., Bubeck, S., and Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13):1233–1239.

[40] Leippold, M., Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Luterbacher, J., and Huggel, C. (2024). Automated Fact-Checking of Climate Change Claims with Large Language Models. *arXiv preprint arXiv:2401.12566.*

[41] Li, C., Zhou, H., Glavaš, G., Korhonen, A., and Vulić, I. (2023). On Task Performance and Model Calibration with Supervised and Self-Ensembled In-Context Learning. *arXiv preprint arXiv:2312.13772.*

[42] Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., and Zhang, Q. (2023). Calibrating LLM-Based Evaluator. *arXiv preprint arXiv:2309.13308.*

[43] Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., and Collier, N. (2024). Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. *arXiv preprint arXiv:2403.16950.*

[44] LMSYS (2024). Chatbot Arena. https://chat.lmsys.org/?leaderboard. Accessed: March 2024.

[45] Mach, K. J., Mastrandrea, M. D., Freeman, P. T., and Field, C. B. (2017). Unleashing expert judgment in assessment. *Global Environmental Change*, 44:1–14.

[46] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426.*

[47] Meta AI (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288.*

[48] Meta AI (2024). Llama 3. https://llama.meta.com/llama3/. Accessed: May 2024.

[49] Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

[50] Müller-Hansen, F., Repke, T., Baum, C. M., Brutschin, E., Callaghan, M. W., Debnath, R., Lamb, W. F., Low, S., Lück, S., Roberts, C., Sovacool, B. K., and Minx, J. C. (2023). Attention, sentiments and emotions towards emerging climate technologies on Twitter. *Global Environmental Change*, 83:102765.

[51] OpenAI (2020). OpenAI API Pricing. https://openai.com/api/pricing/. Accessed: June 2024.

[52] OpenAI (2022). ChatGPT. https://chatgpt.com/. Accessed: March 2024.

[53] OpenAI (2024). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[54] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training Language Models to Follow Instructions with Human Feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

[55] Park, K., Choe, Y. J., and Veitch, V. (2023). The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint arXiv:2311.03658*.

[56] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[57] PyPDF Contributors (2024). PyPDF: A Pure-Python PDF Library. https://github.com/py-pdf/pypdf.

[58] Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to Generate Reviews and Discovering Sentiment. *arXiv preprint arXiv:1704.01444*.

[59] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[60] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[61] Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. (2023). Verbosity Bias in Preference Labeling by Large Language Models.

[62] Science Feedback (2017). https://science.feedback.org/. Accessed: May 2024.

[63] Skjuve, M., Følstad, A., Fostervold, K. I., and Brandtzaeg, P. B. (2021). My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, 149:102601.

[64] Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. (2024). Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248*.

[65] United Nations (1988). Intergovernmental Panel On Climate Change. https://www.ipcc.ch/about/. Accessed: April 2024.

[66] Vaghefi, S., Muccione, V., Huggel, C., Khashehchi, H., and Leippold, M. (2022). Deep Climate Change: A Dataset and Adaptive domain pre-trained Language Models for Climate Change Related Tasks.

[67] Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Allen, S., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Wang, Q., Webersinke, N., Huggel, C., and Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4(1):480.

[68] Varini, F. S., Boyd-Graber, J., Ciaramita, M., and Leippold, M. (2021). ClimaText: A Dataset for Climate Change Topic Detection. *arXiv preprint arXiv:2012.00483*.

[69] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. Curran Associates, Inc.

[70] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. (2023). Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926*.

[71] Webersinke, N., Kraus, M., Bingler, J., and Leippold, M. (2022). ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

[72] Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., and Chen, D. (2024). Evaluating Large Language Models at Evaluating Instruction Following. *The Twelfth International Conference on Learning Representations*.

[73] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

[74] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object Detectors Emerge in Deep Scene CNNs. *arXiv preprint arXiv:1412.6856*.

[75] Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., and Roy, S. (2024). Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. *arXiv preprint arXiv:2309.17249*.

[76] Zhou, H., Wan, X., Vulić, I., and Korhonen, A. (2023). Survival of the Most Influential Prompts: Efficient Black-Box Prompt Search via Clustering and Pruning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Appendix A

Further Details on Experiment 1

A full implementation of Experiment 1 is provided at `https://github.com/maiush/PPairS/tree/main/src/dev/benchmarks`.

Metrics. For a given text summary, a ground truth score for each of the four aspects is calculated by averaging the scores assigned by all human judges. A pairwise comparison can then be made between two text summaries along a given aspect by comparing these scores. We report the classification accuracy for agreement between these ground truth pairwise comparisons and those obtained via direct-scoring, logit comparison (as in PAIRS) or linear probes (as in PPAIRS). Each of these three sets of pairwise comparisons leads to (potentially) different sortings of all text summaries: we also report Spearman's rank correlations between them all.

Models. For a fair comparison with the experiments carried out in Liu et al. [43], we select the same two open-source models used: MISTRAL-7B-INSTRUCT-V0.1 [32] and LLAMA-2-7B-CHAT [47]. We include results from a third open-source model, LLAMA-3-8B-INSTRUCT [48], which can be considered of comparable size yet higher capability [44]. Results using GPT-4-TURBO (-preview) are also lifted from Liu et al. [43] for an additional comparison with a frontier proprietary model, when considering Spearman's rank correlations.

Baselines. For our main baseline we report results applying PAIRS with a beam size of 1000 and uncertainty threshold at 0.6 (following Liu et al. [43]), with the same hyperparameter choices used for PPAIRS when sorting. We additionally report results using direct-scoring.

A.1 Prompt Templates

For both direct-scoring and PAIRS we follow the prompt-templates used in Liu et al. [43].

For direct-scoring we use the following prompt template:

```
<INSTRUCTION>
Source text: <ARTICLE>
Summary: <SUMMARY>
Please rate on a scale from 1 to 5, where 1 represents very low
<ASPECT>, and 5 indicates excellent <ASPECT>. You must only return
an int score.
```

For PAIRS, the template for a pairwise comparison is:

```
<INSTRUCTION>
Source text: <ARTICLE>
Summary A: <SUMMARY A>
Summary B: <SUMMARY B>
Question: Which summary is more <ASPECT>?
If the summary A is more <ASPECT>, please return 'A'.
If the summary B is more <ASPECT>, please return 'B'.
You must only return the choice.
```

Here, *<ASPECT>* is one of coherence, consistency, fluency, and relevance, while *<INSTRUCTION>* adds detail on each aspect to aid the LLM in scoring / comparing. For pairwise comparisons these details are:

- **Coherence**: *"Compare the coherence of the two following summaries."*
- **Consistency**: *"Evaluate and compare how two summaries consistently follow the source text. A summary is consistent with the article if it faithfully reflects the main points, facts, and tone of the article. A summary is inconsistent if it introduces any errors, contradictions, or distortions of the original article."*
- **Fluency**: *"Evaluate and compare the fluency of the two following summaries."*
- **Relevance**: *"Evaluate the relevance of the following summaries. A summary is relevant if it captures the main points from the article, without leaving out any crucial details*

> *or adding any unnecessary or inaccurate ones. A summary is more relevant if it uses the same or similar terms and expressions as the article. A summary is less relevant if it omits some of the key facts from the article, or if it introduces irrelevant information that is not supported by the article.”*

In PPairS, these details and implied by ground truth labels and learned by a linear probe, and are therefore not required[1]. The prompt for PPairS is consequently much simpler:

```
Consider the following article: <ARTICLE>
Below are two summaries of this article:
Choice 1: <SUMMARY1>
Choice 2: <SUMMARY2>
Which summary is more <ASPECT>?
```

To construct a contrast pair we append the following text to the end of the above prompt: *“Between Choice 1 and Choice 2, the more <ASPECT> choice is Choice <CHOICE>”*, with <*CHOICE*> varied between ‘1’ and ‘2’.

A.2 Linear Probe Training

In order to simulate a realistic setting we use ground truth scores for 100 summaries when training a linear probe (we felt this was a reasonable number of human judgements to collect for most evaluation tasks). This results in a total of 9900 pairwise comparisons. Probes are trained on these data using `Python`’s `scikit-learn` library [56] for a maximum of 10000 epochs, employing an early-stopping criterion on a validation set. We find weight-decay regularisation to limit the performance of trained probes, and an L_1 regularisation term for sparsity results in no clear improvement.

[1] We see this as a significant advantage of PPairS.

Appendix B

Further Details on Experiment 2

A full implementation of Experiment 2 is provided at `https://github.com/maiush/PPairS/tree/main/src/dev/sciencefeedback`.

Metrics. We report classification accuracy against ground truth pairwise comparisons for direct-scoring, logit comparison (PAIRS) and linear probes (PPAIRS).

Models. In Experiment 1 we find LLAMA-3-8B-INSTRUCT to be the most competitive open-source model used. In this experiment all pairwise comparisons for all three methods are carried out using this model, in the same manner as the methodology of Experiment 1.

Baselines. As in Experiment 1, we report PAIRS as our main baseline (using the same hyperparameter settings) and additionally report results using direct-scoring.

After scraping all tagged claims from the SCIENCE FEEDBACK publication [62] we manually bin them into three categories: inaccurate, misleading, and accurate. This is done to narrow down the number of discrete tags as we find many to be synonymous. Specifically, we group the following tags together:

- **Inaccurate**: *inaccurate*, *incorrect*, *unsupported*, *reasoning*
- **Misleading**: *misleading*, *context*, *imprecise*, *but...*
- **Accurate**: *accurate*, *correct*

B.1 Direct-Scoring vs PAIRS

In the results presented in Table 4.3, we see significantly higher performance using direct-scoring than the original PAIRS method. We believe the most likely explanation is the set-up of the experiment itself, specifically the classification of claims into only three categories: inaccurate, misleading, and accurate. Consider two inaccurate claims of different magnitude: one makes a more egregious error than the other. A pairwise comparison would find one claim "more factually accurate" than the other, while a ground truth label would consider both claims equal in the sense of them both being inaccurate. This issue could perhaps be ameliorated by allowing for more than three (and therefore more nuanced) categories, but due to the discrete nature of classification the core of the issue will always persist. Indeed, we feel this issue can *only* be eliminated by a supervised method, such as PPAIRS, able to infer the threshold for non-equality in pairwise comparisons from labels.

B.2 Prompt Templates

For direct-scoring we use the following prompt template:

```
Consider the following claim:
Claim: <CLAIM>
Is this claim accurate, inaccurate, or misleading?
Responses must be a single choice.
```

For both PAIRS and PPAIRS, we use the same prompt template:

```
Compare the factual accuracy of the following two claims:
Claim 1: <CLAIM 1>
Claim 2: <CLAIM 2>
Question: Which claim is more factually accurate / less ambiguous?
Responses must be a single choice.
```

To construct a contrast pair we append the following text to the end of the above prompt: *"Between Claim 1 and Claim 2, the more factually accurate choice is Claim <CHOICE>"*, with <*CHOICE*> varied between '1' and '2'.

B.3 Linear Probe Training

We follow the same training regime as described in Appendix A.2, using 100 ground truth tagged claims for a total of 9900 pairwise comparisons. Probes are trained for the same number of epochs (10000), employing an early-stopping criterion and no parameter regularisation.

Appendix C

Further Details on Experiment 3

A full implementation of Experiment 3 is provided at `https://github.com/maiush/PPairS/tree/main/src/dev/climatex`.

Metrics. We report classification accuracy of pairwise comparisons against ground truth labels obtained using the true confidence tags for each claim. Additionally, we use a modified binary search algorithm we call PPAIRS_INSERT (Appendix D) to tag a held-out test set of claims with confidence tags using our trained linear probe, reporting classification accuracy here too.

Models. We again make use of LLAMA-3-8B-INSTRUCT.

Baselines. As we have demonstrated in both Experiments 1 and 2 the improvement in performance of PPAIRS over other state-of-the-art techniques, we opt not to include baseline results for this experiment. We focus instead on our accuracy metrics.

C.1 Text Preprocessing

We use the *Natural Language Toolkit* (`NLTK`) [9] for all text processing. Each report is processed using the `py-pdf` library [57] and subseqeuently tokenised into sentences. We remove all parenthetical references, author references, and section indicators, and replace all common abbreviations and acronyms e.g., *ERF* → *effective radiative forcing*. Each processed sentence is then searched for confidence tags and paired with its context: the preceding and succeeding five sentences within the same section.

C.2 Further Details on Topic Modelling

We extract a total of 10065 claims, with 255, 145, 1170, and 8495 claims from the periods of the third, fourth, fifth, and sixth assessment cycles respectively. Vector embeddings of each claim and context are generated using SBERT [60]. The collection of embeddings is then clustered using BERTopic [25] into 206 different topics, after removing common English stop words and considering both unigrams and bigrams. After fitting, we employ maximal marginal relevance with a diversity parameter of 0.2 to increase the diversity of topic words.

The 206 original topics are grouped via hierarchical clustering into 49. These can each be identified by commonly occurring words e.g., *ocean*, *coral*, *ecosystems*, *fish*, *reefs*, *organisms*, *phytoplankton*. This particular group of words implies a relatively clear label for the topic on the whole: *marine ecosystems*. Others are less obvious, and a final set of 29 topics is created by discarding them. The full list of topics is as follows:

- Sustainability
- Decision-Making
- Climate Models
- Emissions
- Aerosols
- Energy
- Technology
- Transport
- Precipitation
- Temperature
- Sea Level
- Ocean Acidification
- Marine Ecosystems
- Biodiversity
- Forests
- Polar Regions
- Tropical Islands
- Coasts
- Mountains
- Australasia
- Economics
- Migration
- Disease
- Farming
- Fishing
- Water (supply)
- Peat
- Geoengineering
- Bioenergy

C.3 Prompt Templates

We use the following prompt template for PPairS:

```
Compare the confidence of the following two statements,
considering their respective contexts:
```

```
Context 1: <CONTEXT 1>
Statement 1: <STATEMENT 1>

Context 2: <CONTEXT 2>
Statement 2: <STATEMENT 2>

Question: Which statement is expressed with higher confidence?
Responses must be a single choice.
```

Contrast pairs are constructed by appending the following text to the above prompt: *"Between Statement 1 and Statement 2, the more confident choice is Statement <CHOICE>"*, with <*CHOICE*> varied between '1' and '2'.

C.4 Linear Probe Training

We follow a similar approach as described in Appendix A.2, training again on 100 ground truth samples. For this experiment we choose 25 examples from each of the third, fourth, fifth, and sixth assessment cycles for a balanced dataset. Probes are trained for a maximum of 1000 epochs (we find allowing for 10000 epochs results in overfitting).

C.5 Visualising Why Probe Performance is High

We can visualise why a linear classifier performs well on this task in Figure C.1, which shows three different angles of a 3D projection of generated embeddings. We follow the methodology of PPAIRS and construct contrast pair embeddings for each pairwise comparison. After centering to remove the syntactical contrastive feature we take the embedding difference and project this dataset onto its first three principal components (principal component analysis [34]). When colour-coding each projected point with its ground truth pairwise comparison we see the first principal component roughly classifies these data, with a positive projection indicating the first claim in a given comparison is expressed with higher confidence, and a negative projection favouring the second claim. The distinction between these classes is rough, reflecting the imperfect classification accuracy of our probe.

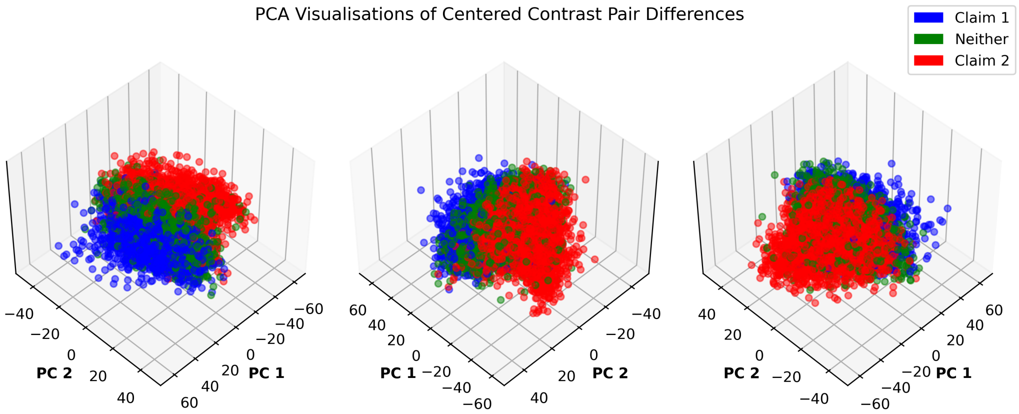


Fig. C.1 PCA visualisations of the centered differences between contrast pair embeddings for pairwise comparisons of claims. Each sub-figure shows a different angle of the same visualisation. Points are colour-coded according to the ground truth of which claim is expressed with higher confidence.

C.6 PPAIRS_INSERT Hyperparameter Settings

For our set of ground truth claims on which we perform a (modified) binary search, we randomly select 100 claims from the sixth assessment cycle of the IPCC. We select a random sample of size 10 when performing pairwise comparisons (Step 2 in Algorithm 1, see Appendix D).

Appendix D

PPAIRS_Insert

In Chapter 4, Experiment 3, we use PPAIRS to automatically evaluate single text samples in an online fashion. This is achieved through a modification to a binary search algorithm, which we call PPAIRS_INSERT.

Consider the general set-up of an automatic evaluation task: a dataset of text samples $T = \{t_i\}_{i=1}^{N}$ is to be scored according to an ordered set of k possible scores $S = \{s_1, s_1, \ldots, s_k\}$ where s_1 represents the lowest score and s_k represents the highest. Given a set of ground-truth labels for text samples $G = \{g_i\}_{i=1}^{M}$, we store a mapping $f : S \rightarrow G$ from each possible score to all labeled text samples whose ground-truth matches. The key aspect of our insertion algorithm is that we determine whether a given text sample t has a higher, lower, or equal score to a given score s by evaluating a trained linear probe on a *subset* of $f(s)$, and calculating the average preference probability.[1] The pseudocode for this process in Algorithm 1 is that of a binary search accounting for this subset operation.

[1] In the case of ternary classification (higher, lower, equal) the mode can also be used.

Algorithm 1: PPAIRS_INSERT

```
input :
        single text sample $t \in T = \{t_i\}_{i=1}^{N}$,
        possible scores $S = \{s_1, s_1, \ldots, s_k\}$,
        ground-truth text samples $G = \{g_i\}_{i=1}^{M}$,
        mapping from scores to ground-truth samples $f : S \to G$,
        number of samples to subset $N_{samples}$,
        linear probe $P : T \times G \to [0, 1]$,
        $low = 1$,
        $high = k$
output :
        predicted score $tag \in S$

$mid \leftarrow (low + high)/2$;
$subset \leftarrow \text{sample}(f(s_{mid}), N_{sample})$;          /* random sample of size $N_{sample}$ */

initialise $predictions$ as empty array;
for $g^* \in subset$ do
    $predictions \leftarrow predictions \cup \{P(t, g^*)\}$;
$\mathbb{P}(t \succ s_{mid}) = \overline{predictions}$;

if $\mathbb{P}(t \succ s_{mid}) > 0.5$ then
    if $mid = high$ then
        return $tag$
    else
        return PPAIRS_INSERT($t, S, G, f, N_{samples}, P, mid + 1, high$)

if $\mathbb{P}(t \succ s_{mid}) < 0.5$ then
    if $mid = low$ then
        return $tag$
    else
        return PPAIRS_INSERT($t, S, G, f, N_{samples}, P, low, mid - 1$)

if $\mathbb{P}(t \succ s_{mid}) = 0.5$ then
    return $tag$
```