USING LARGE LANGUAGE MODELS TO ANALYZE POLITICAL TEXTS THROUGH NATURAL LANGUAGE UNDERSTANDING

Short title: "Using LLMs for Natural Language Understanding"

Kenneth Benoit, Scott De Marchi, Conor Laver, Michael Laver, Jinshuai Ma*

July 20, 2025

* Author affiliations, in order: Singapore Management University and London School of Economics; Duke University; Roux Institute, Northeastern University; New York University and London School of Economics; Data Science Institute, London School of Economics. Author contact details: Kenneth Benoit: kbenoit@smu.edu.sg; Scott De Marchi: demarchi@duke.edu; Conor Laver: conorlaver@gmail.com; Michael Laver: michael.laver@nyu.edu; Jinshuai Ma: J.Ma23@lse.ac.uk.

We thank the Data Science Institute of LSE for material support for this project. We thank Ryan Bakker, Zach Dickson, Gaël Le Mens, Gary King, Lanny Martin, Philip Resnick, Georg Vanberg, and Annika Werner for comments on earlier drafts of this paper.

ABSTRACT

Large language models (LLMs) offer scalable alternatives to human experts when analyzing political texts for *meaning*, using Natural Language Understanding (NLU). Qualitative NLU methods relying on human experts are severely limited by cost and scalability. Statistical text-as-data (TaDa) methods are scalable but rely on strong and often unrealistic assumptions. We propose a systematic, scalable, and replicable method that can extend existing qualitative and quantitative approaches by using LLMs to interpret texts meaningfully, rather than as mere data. Our ensemble means of LLM-generated estimates of party positions on six key issue dimensions correlate highly with equivalent mean ratings by country specialists. When applied to coalition policy declarations, LLM estimates align more closely with standard models of government formation than hand-coded estimates. We conclude with a discussion of the profound implications of modern LLMs for political text analysis.

Verification Materials:

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/XY1FFE.

Word count: 9,654

Keywords: LLMs, political parties, policy positions, AI, text analysis

INTRODUCTION

Reliably extracting information from text is fundamental to social science. This often means identifying authors' stated preferences on key issues such as environmental protection or immigration policy. Written documents and speech transcripts are rich sources of such information, and scholars often systematically describe and compare such preferences using latent issue dimensions. The ideal way to extract preference information from a text involves reading it for meaning and interpreting its issue positions. This fundamentally qualitative approach has been practiced informally by generations of country specialists, using deep familiarity with both the language and the political context of the relevant setting.

Qualitative text labeling is a longstanding research tradition that has produced widely-used datasets in political science, including the Manifesto Project (hereafter MP; Budge, et al. 1987, Budge, et al. 2001, Klingemann, et al. 2006, Volkens, et al. 2013) and the Comparative Policy Agendas Project (CAP; Baumgartner, et al. 2006, Baumgartner, et al. 2019). Both projects spent decades, and millions of dollars, labeling the policy content of political documents—over three million sentences and 5,000 manifestos for the MP alone—using hundreds of trained expert annotators. Despite widespread use in the profession of datasets produced by the MP and CAP, we have no sense of their reproducibility. These datasets are too expensive to replicate. While alternative methods use non-experts to crowd-source sentence labeling (Snow, et al. 2008, Benoit, et al. 2016), these also scale poorly, typically requiring dozens of crowd coders to process single documents, even when the aim is to identify a small target of relevant policy statements. More generally, methods based on human input do not scale.

Seeking ways to extract information from political text at scale, scholars devised and implemented methods that are fully or partially *quantitative*. Automated "text-as-data" (TaDa) approaches had considerable success using quantitative methods to analyze words in a text, despite minimal to zero knowledge of the language or political context involved and remaining completely blind to the text's meaning (see Grimmer and Stewart 2013; Gentzkow, et al. 2019; Benoit 2020). In their crudest forms, TaDa methods use statistical techniques to analyze counts of tokens, while more sophisticated approaches may also rely on vector representations of text units ("word embeddings", e.g., Mikolov, et al. 2013).

The key shortcoming of quantitative methods that ignore a text's meaning is that they struggle to capture its nuances, including the intensity of positions it expresses. Consider two sentences about labor market policy. "Capital is dead labour, which, vampire-like, lives only by sucking living labour, and lives the more, the more labour it sucks." "We will establish an independent review to consult on how to set a genuine Living Wage across all sectors." Both set out pro-worker policies. The first comes from *Das Kapital*[1] and uses only a few words to convey a trenchant far-left position in unambiguous terms. The second comes from the UK Liberal Democrat manifesto of 2019 and sets out a much weaker left-wing position. Any human reader can immediately see the stark difference between these statements. Traditional TaDa methods struggle—if not fail outright—to do this. While it has long been recognized that such methods depend on "wrong" models of language (Grimmer and Stewart 2013, 3), they have been widely accepted, absent a better alternative.

Existing approaches to text analysis therefore face a fundamental trade-off. Quantitative TaDa methods scale to large corpora but treat words as features of the data, often relying on statistical assumptions that abstract away their meaning. In contrast, qualitative human text labeling captures the meaning and intent behind language but lacks scalability. Here, we describe a method of text analysis based on an ensemble of large language models (LLMs). Unlike TaDa models, LLMs are powered by massive pre-training, using effectively the entire internet, of huge models with billions or even trillions of

1 Volume 1, chapter 10. This quote was located in *Das Kapital* by ChatGPT as an example of an extreme anti-capitalist position.

parameters. This gives them extraordinary ability to summarize and draw inferences from large texts, combining the scalability and reproducibility of quantitative TaDa approaches with the interpretative nuance of traditional qualitative methods.

Given this, our method first asks LLMs to summarize what a text has to say about particular issues. We then ask them to answer questions about the author's issue positions. This is a canonical application of "Natural Language Understanding" (NLU), a subfield of Natural Language Processing (NLP) focused on tasks such as text summarization and question answering — precisely what we are using LLMs to do here. Note that "understanding" in NLU is not understanding in the sense that machines use anything approaching human cognition. NLU characterizes a set of machine tasks that can *replicate outputs* of human interpretations of text (Samant, et al. 2022 review the NLU literature).

Using LLMs in this way offers the scalability of quantitative methods applied to interpretative text analysis—the output of which captures some of the depth and richness of traditional qualitative analysis. This has the potential to revolutionize text analysis in political science, offering cheap, agile, but effective ways to generate results which are functionally equivalent to traditional qualitative text analysis, while doing this at potentially massive scale in practically any language. This has led to their increasing adoption in social sciences where qualitative text analysis has traditionally been the dominant paradigm (see Karjus 2023; Linegar, et al. 2023).

In any scientific endeavor, what is crucial is *not any particular dataset*, but a *reliable and replicable method for generating data* that addresses the precise problem at hand. In what follows, we develop protocols for using LLMs in this way and show in our findings that, with scrupulous run design, LLM-based results are reliable, reproducible and valid.

USING LLMS TO ESTIMATE ISSUE POSITIONS IN POLITICAL TEXT

Our goal is to develop protocols for systematically and reliably using LLMs to extract information about political preferences of the authors of political texts. Our workflow is summarized in Table 1 and described in the sections that follow.

[TABLE 1 ABOUT HERE]

Input texts and validation data

Text corpus: party manifestos and coalition agreements

Results we report in this paper concern issue positions in party manifestos and coalition agreements. The same approach can be used for other political texts—for example legislative speeches or social media posts. Substantively, we focus validation of our approach on party manifestos because estimating parties' issue positions is a core research program in political science (Laver 2003), while manifestos are parties' published commitments to these positions. Methodologically, we focus on parties' issue positions because there are authoritative and widely used datasets, in particular the MP and expert survey series we describe below, dealing with precisely the quantities we wish to estimate. This allows us to validate results in a high-information environment, before moving on to analyze the issue positions of coalition cabinets, on which much less research has been conducted.

Our text corpus for party manifestos is the subset of manifestos in the MP collection (Merz, et al. 2016) issued in years for which expert surveys were also fielded. This comprises 235 party manifestos written in 21 different languages (detailed in Supplemental Appendix A1). We downloaded these from the MP collection, cleaned and converted them into plain text format for input into LLMs.[2] Our corpus of coalition policy agreements is the set of 23 coalition agreements, from a collection assembled by Kluwer et al., for multi-party cabinets forming in election years covered by our manifesto corpus (Klüver, et al. 2023).

Validation data: expert surveys and MP issue scales

Whenever we evaluate a new measurement instrument, a standard way to assess the validity

[2] Since we began this work, most commercial LLMs can read most pdf documents.

of its results is to compare these with the judgments of experts who have deep local knowledge of the matter under investigation. As a form of "expert crowd sourcing," published expert surveys are widely accepted as reliable and valid estimates of party policy positions. We therefore assess the *convergent* validity of our LLM-generated data by comparing these with qualitative judgements given by ensembles of experts collected in a widely used expert survey series, which we use as our benchmark. This begins with a survey by Laver and Hunt in 1989, extended by Benoit and Laver in the early 2000s (Laver and Hunt 1992, Benoit and Laver 2006). The Chapel Hill Expert Survey (CHES) series used the same dimensions and vignettes as these in four expert surveys fielded between 2006 and 2019 (Jolly, et al. 2022).[3] We selected six issue dimensions consistently deployed over this series, each concerning a different aspect of policy, for estimation using the LLMs. Table 2 lists these, together with the end-of-scale vignettes offered to experts.

[TABLE 2 ABOUT HERE]

In answering these surveys, expert survey respondents are instructed to rely on their overall knowledge of the parties, of which manifesto content is only a potential component. We also compare our results, therefore, with positions on logit scales[4] referring to the same six issues, derived from published estimates generated by the MP, based on quantitative analyses of the qualitative human labeling of manifesto text units.[5]

To evaluate *predictive* validity, we take as our benchmark a prediction shared by many spatial models of government formation, regardless of the bargaining model on which they are based. This is that the agreed coalition issue position should fall within the range of

[3] As of April 2025, the trend file for 1999-2010 CHES data had 1,328 Google Scholar citations (Bakker, et al. 2015).
[4] As recommended by Lowe et al. 2011, see Appendix A2 for details.
[5] As of April 2025, the MP website listed 785 peer-reviewed publications using these data, over and above 63 publications by the MP core team.

member parties' issue positions. If it were outside this range, then a Pareto improvement would be possible by moving it into this range.

Choosing LLMs and software environment

LLMs fall into three broad categories: proprietary, open-weight, and open-source. Proprietary models—such as GPT-4o, Claude, and Gemini—are closed systems: their training data, model code, and weights are not publicly released. While often the most powerful and user-friendly, they evolve in opaque ways and pose challenges for replicability, transparency, and data control. Open-weight models, including LLaMA 3, Deepseek-V3, and Gemma-3, release their model weights under restrictive licenses. They can be downloaded and run locally or deployed on private cloud infrastructure. This offers greater transparency and control—especially for research requiring data confidentiality or, more crucially for the social sciences, long-term reproducibility (Weber and Reichardt 2023). Unlike proprietary models, open-weight models can be extended by fine-tuning or continued pretraining on domain-specific corpora, enabling users to adapt them more directly to their research needs. However, they do not typically disclose their training data or code, and thus are not fully open-source. Fully open-source models (such as Pythia, Mistral, or Falcon) go further, releasing not only model weights but also training data and training code under permissive licences. While this is useful for understanding and manipulating the detail of how LLMs work, open-source models are much less powerful, with much less language coverage, than the best proprietary and open-weight models, so we do not consider them here.

Privileging powerful and accessible text analysis at scale, and designed to analyze public domain documents, our baseline method uses three proprietary LLMs: GPT-4o (from OpenAI), Gemini 1.5 Pro (from Google), and Claude 3.5 Sonnet (from Anthropic). Free browser-based versions of these are impractical for the text analysis workflow outlined in

Table 1.[6] Subscription versions give access to the model's API, allowing much more powerful analysis to be integrated into the user's software.[7] In light of the increasing performance of the best open-weight models, our replication analysis also used three of these: DeepSeek-V3-0324; Llama-3.3-70B-Instruct (from Meta); gemma-3-27b-it (from Google).[8]

LLMs typically have a *temperature* setting controlling the randomness—or "creativity"—of generated text: higher values increase variability by allowing the model to sample less probable tokens when generating responses. To maximize replicability, we set temperature to zero for all models, which encourages more deterministic outputs but does not fully eliminate variability. We also set the *top-p* (nucleus sampling) parameter to 1.0 to prevent additional sampling constraints, and where available, we fixed the *random seed*. Despite these measures, we observed small residual variation in some models' outputs across runs—likely due to stochastic processes within their generation pipelines or undocumented aspects of inference architecture.

The systematic analysis of large text corpora using multiple prompts and a variety of LLMs requires a robust software environment. This is provided by Langchain (Chase 2022) an open-source package that is a central hub for the development of LLM applications. This allows users to run the same analysis on the latest versions of a wide range of LLMs—proprietary as well as open-weights and open-source. Langchain also requires specification and recording of the precise issue number of each LLM deployed in each analysis—crucial replication information often lost or omitted in reports of more casual LLM-based analyses.

[6] Browser versions are also increasingly "agentic", unlike models accessed via the API, making web calls on the fly rather than relying solely on their training data. This makes their output less replicable.
[7] LLM APIs typically have tiers; higher tiers have increasing costs but much higher volumes and rates of inputs and outputs. It was essential in our work to spend (quite modest) funds to move up several tiers for each LLM.
[8] Given the considerable computing power required, we found it most convenient to run these on virtual machines hosted by the Nebius platform (https://studio.nebius.com/), but with sufficiently powerful hardware, all three LLMs could have been deployed locally. In all three cases, we were able to download and archive the model files for the open-weight LLMs.

We are not the first to use GPT-class LLMs to estimate policy positions from party manifestos. Applying LLMs to tasks from Benoit et al. (2016), both Le Mens and Gallego and Ornstein et al. used LLMs to replicate crowd-sourced and expert classifications of policy content at the sentence level (Le Mens and Gallego 2025, Ornstein, et al. 2025). Bol and Bono simply asked ChatGPT to answer the CHES expert surveys questions but did ground these in manifestos and applied them only to eight French parties for one election (Bol and Bono 2025). Our approach differs significantly. Rather than analyzing individual sentences and aggregating results into an index, we treat texts *holistically*, asking LLMs to "read" and summarize their policy content in its entirety before scoring the summarized content on issue scales. While the LLMs do not replicate human cognition in any mechanistic sense, their capacity to interpret a document as a coherent whole enables a *functional equivalence* to how a human coder might read and assess a text's policy positions on particular issues. Instead of relying on human qualitative coding or using LLMs as a machine learning tool for sentence-level classification, therefore, we exploit their NLU capacities for abstractive summarization, interpretation and question answering (Zhang, et al. 2024).

Summarizing texts using LLMs

The main proprietary LLMs can analyze very long texts. However it is neither efficient nor effective to feed an LLM a large text (some manifestos exceed 150 pages) and ask it to rate positions on an issue discussed in only sentence or two. This is the "needle in the haystack" problem (Wang, et al. 2024, Hosseini, et al. 2024, Chang, et al. 2023). We modularized the huge transition from a large text to a single-integer summary of an issue position by leveraging a capability at which LLMs are highly effective: generating concise summaries of large documents (Hoyle, et al. 2023, Wu, et al. 2024). We therefore introduced an intermediate stage between the full text and the single-integer summary, asking the LLMs to generate a short English-language summary of what the original language manifesto says about the issue in question. We engaged in extensive reading of these summaries, which were typically lucid and on-topic, to check their substantive validity. Value added by the summary stage became clear when we prototyped a method bypassing this, which resulted in substantially worse results.[9]

Our prompt strategy for using the LLMs to summarize long texts was extensively explored during prototyping, described in Supplemental Appendix B. We experimented with multi-issue summaries as well as longer or shorter summaries and settled on prompting for six short (300-400 words) single-issue summaries per document.[10]

We see generating intermediate text summaries as an integral part of our method, even as LLMs continue to advance in capabilities. Improvements in text summarization will naturally accompany these advances, but passing large volumes of text to an LLM still carries the risk of context or attention drift (Liu, et al. 2023). More importantly, even if this issue were completely resolved, using English language summaries provides critical transparency—especially for researchers working with texts written in languages they cannot understand. These summaries offer a traceable representation of the input content, enabling human review and validation in a way that would be lost if the process moved directly from a 150-page original-language document to a single numerical score. Crucially, this transparency also permits us to inspect and validate how the LLMs derived their positions, something we illustrate below through a closer examination of the decentralization and environmental policy domains.

The summarizer performs two tasks simultaneously: it summarizes the original-language manifesto and returns that summary in English. During prototyping, we explored whether performance could be improved by modularizing these tasks—first translating the manifesto, then summarizing the translation. However, we found that correlations with expert

[9] See Supplemental Appendix B, especially Table B5.
[10] For prompt wordings, see Supplemental Appendix A3.

survey scores tended to be lower when a translation stage was inserted before summarization,[11] so we excluded this step from our final design.

Scoring issue positions in text summaries

Concerned with the possibility that small changes in the wording of scoring prompts might have big effects on LLM outputs, following (Battle and Gollapudi 2024), we tested a matrix of nine different wordings for each LLM scoring prompt.[12] We concluded that small changes in prompt wording of the type we investigated are not critical in this application, settling on the prompt wording most closely corresponding to prompts given by CHES to human experts: "You are an expert social scientist with a PhD in political science. Think carefully about your answer." Again responding to the received wisdom that LLMs perform better, the more help they are given, we settled on a standard seven-point scale for recording issue preferences, supplying detailed substantive interpretations of each of the seven points for each of the six issue scales. These are set out in Supplemental Appendix A4.[13]

Few-shot learning

Asking LLMs to score texts without giving examples of what a "good" score looks like is known as "zero-shot" learning. "Few-shot" learning, by contrast, uses longer LLM prompts that include sample text summaries along with associated benchmark scores. We implemented few-shot learning, using benchmark scores from expert surveys. To generate training examples, we selected manifesto summaries—edited for succinctness and clarity—from manifestos associated with expert survey scores that span the issue dimensions of interest. For each issue, our few-shot scoring prompts supplement the zero-shot prompts with three examples—from the right, left and center of each dimension—paired with their

[11] Detailed results in Supplemental Appendix Table B7.
[12] See Supplemental Appendix Table B1.
[13] We conducted an initial prototyping run using the CHES 11-point scales which have two anchors, one at each end. This produced systematically worse results than the seven-point, seven-anchor scales we developed, likely because of the precise meaning given to each scale position.

corresponding expert survey scores.[14]

Run design and ensemble means

Following design decisions discussed above, the final "3x3" run design for our baseline text analysis was as follows. For each of the 235 manifestos in Appendix A1 and each of the six dimensions in Table 2, we first used each of Claude, GPT and Gemini to generate 300- to 400-word summaries (in English) of the issue positions set out in the full native language text. This generated 235 x 6 x 3 = 4,230 issue position summaries. We then used each of Claude, GPT and Gemini to score each of the three summaries per manifesto-dimension on a 1–7 scale, returning "NA" if there was insufficient information. This generated nine LLM scores (including NAs) for each issue dimension for each manifesto—12,690 LLM scores in all. We used this design for two different runs, the first with zero-shot learning and the second with few-shot learning, as specified above. The total output was an ensemble of 18 LLM scores for each manifesto-dimension—25,380 LLM scores in all.

We use the mean of the 18 estimates from zero- and few-shot 3x3 runs to aggregate results for each issue-manifesto generated by our ensemble of LLMs (Sujatha and Nimala 2024; Abburi, et al. 2023). This offers a robust and simple way to compare results of our manifesto analysis, using multiple LLM summarizers and scorers, with the ensemble means in benchmark expert survey data (Yang, et al. 2023).

SUBSTANTIVE VALIDITY OF LLM ESTIMATES OF PARTIES' ISSUE POSITIONS

Benchmarks for substantive validity

Our quantity of interest is the "true" position of party $\boldsymbol{p}$ on issue dimension $\boldsymbol{d}$. Given a set of parties and issues, we want to estimate a matrix of parties' true issue positions, $\boldsymbol{T_{pd}}$. These positions may mean many things. For our purposes they are positions motivating party

[14] Since the LLMs are asked to return integer scores, we rounded the relevant rescaled CHES means.

leaders' elite actions in government, opposition, or government formation. These true positions exist only in the minds of politicians whose observable behaviors may well be strategic. $\boldsymbol{T_{pd}}$ is *fundamentally unobservable*. This means that there is no observable "ground truth" target with which we can compare our estimates—a problem common to many methods of data collection in the social sciences, including all survey research.

We use our ensemble of LLMs to generate $\boldsymbol{L_{pd}}$, a new observable *estimate* of unobservable $\boldsymbol{T_{pd}}$. To validate $\boldsymbol{L_{pd}}$, we compare it with two published data sources, the expert survey series and the MP. Expert surveys ask a collection of experts to score party positions on issue dimensions with substantively defined endpoints.[15] For a given setting and time point, an expert survey generates an observed matrix, $\boldsymbol{E_{pd}}$, of party positions on dimensions of interest. Each cell in the matrix is an ensemble mean of expert scores. There is substantial variation around these means. This arises both from disagreements among experts on the "true" position and the varying frames of reference they might bring to bear on their task.[16] $\boldsymbol{E_{pd}}$ is an observable estimate of $\boldsymbol{T_{pd}}$.

Our conjecture is that observed expert survey and LLM estimates, $\boldsymbol{E_{pd}}$ and $\boldsymbol{L_{pd}}$, are both systematically related to our unobservable target, $\boldsymbol{T_{pd}}$. We therefore expect them to be systematically related to each other. Using $\boldsymbol{E_{pd}}$ as our benchmark, therefore, we assess content validity of $\boldsymbol{L_{pd}}$ using correlations between $\boldsymbol{L_{pd}}$ and $\boldsymbol{E_{pd}}$. It is also possible that, notwithstanding high correlations, $\boldsymbol{L_{pd}}$ is *biased* relative to $\boldsymbol{E_{pd}}$—it is impossible to know whether it is biased relative to $\boldsymbol{T_{pd}}$ since the latter is unobservable. We investigate this by comparing *distributions* of scores generated by $\boldsymbol{L_{pd}}$ and our $\boldsymbol{E_{pd}}$ benchmark.

$\boldsymbol{L_{pd}}$ is generated by using LLMs to analyze party manifestos, to which experts generating the $\boldsymbol{E_{pd}}$ benchmark are never referred. We therefore also compare LLM estimates

[15] The median number of experts per party was 16 in the Benoit-Laver survey and 12 in the CHES series.
[16] The median standard deviation of expert estimates of party positions on the 1-20 scales used in the Benoit-Laver survey was 3.3, and the median standard error 0.75. Given this level of variance in the expert estimates, a test of whether an LLM could have been an expert in an expert survey would present a very soft target.

with data from the MP which, unlike expert survey data, explicitly derive from manifesto content. We computed MP estimates of party positions on each issue dimension using logit scales recommended by Lowe, et al. 2011.[17] These generate a manifesto-based observed data matrix, $\boldsymbol{M_{pd}}$. The value of comparing our LLM estimates with $\boldsymbol{M_{pd}}$ as well as our $\boldsymbol{E_{pd}}$ benchmark, as we do in Figure 3 below, is that there may be issues for which party manifestos *per se* are "biased" relative to expert judgements—biases in parties' *published* issue positions which have nothing to do with measurement instruments.

Upper bound on correlations between LLM and expert ensemble means

Our expert survey benchmark is an ensemble mean of quantities with substantial variation, arising both from sample variability among expert respondents and fundamental uncertainty about the meaning of policy scales and positions (see Benoit and Laver 2003). This variation sets an upper bound on correlations between LLM and expert survey estimates that can reasonably be considered a benchmark for success. To establish this, we simulated correlations between two expert surveys estimating the same issue positions, using the Benoit-Laver replication dataset.[18] For each country, we randomly split the expert pool into two sub-samples. For each sub-sample, we calculated mean expert scores for each quantity of interest. We then correlated the resulting pairs of expert means, simulating the correlation between two expert surveys. Repeating this 1,000 times generates *distributions* of expert-expert correlations, with the means and ranges reported in Table 3. This shows that, given typical variation in expert responses, mean correlations between estimates of the same issue positions by different expert surveys are typically around 0.90—somewhat lower for decentralization and environmental policy. We cannot reasonably expect LLM estimates to correlate with the expert benchmark at a level higher than this upper bound.

[17] For details, see Appendix A2.
[18] Chosen because this involves the largest number of respondents in the expert survey series.

[TABLE 3 ABOUT HERE]

Results: Observed correlations between LLM estimates and expert benchmark

We assess content validity of LLM estimates of party policy positions in terms of their correlations with benchmark expert survey estimates of positions of the same parties on the same issues. Figure 1 summarizes these results, plotting for each issue the relationship between the ensemble mean of LLM scores and the mean expert survey score. Pearson correlations between the LLM ensemble scores and benchmark expert survey means are high for most issues, often ranging from 0.87 to 0.92. Except for decentralization—to which we return in detail below—these results are far better than we hoped for at the beginning of this project. They are at or close to the upper bound, identified in Table 3, of what can reasonably be expected. This is compelling evidence of the convergent validity of our LLM-based approach for estimating party policy positions.

Detailed results, reported in Supplemental Appendix Table C1, show these correlations to be robust between different summarizing and scaling LLMs. Supplemental Appendix Tables C1 and C2 show that correlations between few-shot LLM and expert scores are essentially the same as for zero-shot scores. For many issues, zero-shot correlations are already close enough to the upper bound that there is very little headroom for improvement with few-shot results.

[FIGURE 1 ABOUT HERE]

Decentralization: do experts see issue positions not found in party manifestos?

Figure 1 shows lower correlations between LLM and expert scores for the decentralization dimension. However, we found that the decentralization scores from different LLMs are highly correlated with each other (Appendix Table C3). This suggests that different LLMs consistently identify the same patterns, raising the question of whether experts are basing their judgments on information not present in the published party manifesto.

As we show in Figure 3, expert scores on decentralization span the full range of the policy scale. An informal reading of the manifesto summaries, however, reveals that nearly all party manifestos publicly advocate some form of decentralization, with almost none opposing it. A similar, though less pronounced, pattern occurs with environmental policy: experts score party positions across the full scale, yet most manifestos publicly promote relatively pro-environmental policies; very few take an opposing stance.

MP data add systematic evidence to the casual observation that the manifestos we analyzed *contain almost no anti-decentralization content*. The MP coding scheme has variables for pro- and anti- decentralization.[19] Figure 2 shows box plots of percentages of MP-labeled quasi-sentences favoring and opposing decentralization for the set of manifestos we analyzed with the LLMs, confirming the observation that these have very few anti-decentralization labels. LLM scores have lower correlations with expert judgments on this issue because, while experts judge parties to have a wide range of decentralization policies, there is almost no observable variation in manifesto positions on decentralization. This lack of variation results, as Table 5 and 6 show below, in lower inter-LLM correlations for decentralization than for the other five issue dimensions. In effect, not only are experts likely seeing features of decentralization policy that are not expressed in party manifestos, but also the lack of variation in manifesto discussions of decentralization makes this a more difficult information retrieval task for the LLMs.

[FIGURE 2 ABOUT HERE]

More generally, we can check the extent to which experts and manifestos say different things about party policy by comparing LLM estimates with MP data and expert scores for the same parties' positions on the same policy dimensions.[20] Figure 3 plots the

[19] MP codes, PER 301 and 302

[20] We make direct comparisons between LLM, MP and expert scores by first rescaling expert and MP scores from to the 1–7 scale used by the LLMs.

distributions of these three sets of scores, by issue. The disjunctions between LLM (in green) and expert scores (in red) for decentralization and environmental policy suggest strongly they are not measuring the same thing. The distribution of MP scores (in blue) for these same dimensions suggests that LLMs are capturing what is in the manifestos on decentralization and the environment. Experts must be reading more into party policy on decentralization and the environment than is published in the manifesto. One way of looking at this is to see manifesto content as structurally "biased", systematically downplaying pro-centralization (and anti-environment) positions for strategic reasons, with expert respondents in effect correcting for this. Which measures are biased with respect to unobservable "true" party policy positions is not a question of bias in LLMs, but rather a question of whether parties bias their manifesto commitments on some types of issue dimension, a matter on which Figure 3 offers systematic evidence.

[FIGURE 3 ABOUT HERE]

On the remaining four issues—economic, social, immigration, and EU policy—Figure 3 shows close correspondence between distributions of LLM and expert scores. This suggests both are measuring the same thing, reflected in very high correlations between LLM and expert scores for these dimensions. In contrast, the distribution of human-labeled MP estimates does not match LLM or expert scores for these four dimensions. This reflects not only the structural bias from some "valence" issues as noted above, but also the limitations of the label-counting approach to aggregating issue positions. In results shown in Supplemental Appendix D we note that LLM results correlate much more highly with benchmark expert survey estimates on each of the six policy dimensions than do MP-based estimates.

ROBUSTNESS AND REPRODUCIBILITY ISSUES

"Informative missingness" in LLM summarization and scoring

A crucial feature of our LLM scoring prompt is the instruction to return *NA*, or "not applicable," if the manifesto summary does *not* state a position on the issue in question. This is crucial as both a matter of principle and to mitigate the well-documented tendency of LLMs to produce "hallucinations" (Bruno, et al. 2023, Chrysostomou, et al. 2024)—instances where the LLM generates content that appears plausible but is unsupported by, or even contradicts, the source text. By ensuring the model identifies when sufficient information to assign a score is unavailable, we avoid forcing it to generate spurious scores on dimensions lacking evidence in the manifesto. Conversely, when sufficient information is present, we expect the LLM to return a valid and reliable score.

"NAs" returned by LLMs, expert survey respondents, or MP text coders are highly informative. Knowing what is *not* in a text is as important as knowing what is in it. Expert survey respondents often decline to rate a given party on a given dimension when they lack sufficient information. These "informative NAs", however, are typically not reported in expert survey data comprising ensemble mean scores on party positions. However, they can often be retrieved from replication datasets. For example, 58 experts responded to Benoit and Laver (2006) on issue positions of two UK parties: the well-known Conservative Party; and the less well-known *Plaid Cymru (PC,* Welsh Nationalists). For the Conservatives, the number of experts giving judgments on the various issue dimensions deployed ranged from 54 to 58. For *PC*, this number ranged from 29 (on immigration policy) to 44 (on the economic policy dimension). Does the 50% "expert NA" response on *PC's* immigration policy mean that these experts judged *PC* to have *no immigration policy* in 2002? Or does it mean that *PC* may have had an immigration policy, but the experts do not have the information or ability to make judgements on this? Both are valid and informative expert responses, with very different substantive meanings.

Similarly, LLMs might return an "NA" issue score for one of two reasons: either the manifesto contains no stated position on a particular issue, or the information does exist but the LLM could not retrieve it. We gain insight into this by comparing very different rates of NAs returned when we ask LLMs to generate scores directly from large documents without first summarizing these (the most difficult retrieval task), with those when we give LLMs help with retrieval by first asking them to summarize manifesto content on the issue in question. As reported in Supplemental Appendix B, the more difficult retrieval task generates far higher levels of NAs from the same text, suggesting that many of these result from retrieval problems. This supports a conjecture, reinforced by our extensive reading of manifesto summaries, that NAs returned by the much easier task of scoring summaries imply absent or unclear issue positions.

A systematic feature of our results is considerable variation between LLMs—indeed the only systematic difference we found between them—in the frequency of NAs they return for a given issue. One summarizer (typically GPT) may tell us there is policy content on some dimension while another (typically Claude and/or Gemini) tells us there is not. Reading many summaries, we found a tendency for GPT summaries to say something like "there is no explicit mention of X in the manifesto, however the overall ideological orientation of the manifesto suggests that its policy on X is y". Claude and Gemini summaries were more likely to stop at "there is no explicit mention of X in the manifesto", which then generates an "NA" at the scoring stage. Table 4 summarizes these differences by computing the proportion of missing scores underlying each of our ensemble means of (up to) 18 zero- and few-shot estimates of the same quantity, by issue dimension and summarizer model.[21] This shows both that GPT-4o summaries generate substantially less missingness and that there is more missingness on some dimensions (e.g., immigration) than others (e.g., the environment).

[TABLE 4 ABOUT HERE]

This highlights a crucial trade-off when we use LLMs to analyze political texts, a

[21] For a given summary, there was very little difference between LLMs in missingness at the scoring stage.

trade-off common to most data generation exercises. If we want to be very confident in our estimates, then we may opt to discard estimates for which not all summarizers and scorers are agreed. The price will be a large amount of missing data. If we want to increase the volume of “non-missing” data, we may tolerate less consensual scores, possibly arising because LLMs disagree on whether there is relevant policy content in the manifesto. This is directly analogous to a decision we might make to use only pristine expert survey scores on which *all surveyed experts both make a judgment and agree* or, in common with most of the profession, to use mean scores based on varying numbers of missing expert responses, knowing that some are more reliable than others.

As we have seen, lower levels of missingness generated by GPT-4o arise because it is more likely to use other manifesto text to infer manifesto content on an issue which is not explicitly mentioned. Whether or not this is a good thing is a decision for the end-user, faced with a trade-off between pristine estimates and missing data which is common to most measurement exercises. This is an important feature of research design which we inform here but do not pre-judge. Our best results, however, came from aggregating multiple LLM results into an ensemble mean, just as previous approaches based on experts have found the best estimates to come from aggregating the individual expert scores into a single ensemble mean (see Benoit and Laver 2006, Appendix A).

Reliability of LLM policy scores

Our discussion of LLMs’ varying judgments about whether texts contain specific policy content highlights a key point: just like human experts, LLMs do not always agree. We argue that the workflow outlined in Table 1 produces estimates that are not only valid but also reliable. We examine this claim by analyzing both intra- and inter-LLM agreement.

[TABLE 5 ABOUT HERE]

Leveraging the reproducibility analysis discussed below, Table 5 shows levels of both intra- and inter-LLM agreement for each policy dimension and overall, using standard measures for assessing the reliability of human raters (Burla, et al. 2008). We measure how well each LLM agreed with itself in a repeat scoring of the same manifesto, and agreement between different LLMs on the same issue score. Table 5 shows both the interclass correlation coefficient (ICC), which measures agreement rates for interval data, as well as a version of Krippendorff's alpha designed for ordinal scales (Krippendorff 2018).

Table 5 shows very high reliabilities compared to human raters: on both measures typically 0.90 or above for both intra- and inter-LLM reliability on the first four policy dimensions. The results for the more valence-oriented issues of the environmental and decentralization policy were lower—between 0.76–0.80 (ICC) for environmental policy and 0.66–0.69 (ICC) for decentralization policy—though still high compared to human benchmarks. In experiments with multiple trained coders applying the MP scheme, Mikhaylov et al. (2012, 90) found human coders were highly prone to misclassification and error, with reliability scores typically ranging between 0.3–0.5. Human experts have a "temperature" far from zero; their training data, ability to follow instructions, and internal reasoning processes vary in ways that are wide-ranging and unknowable. LLMs, by contrast, offer remarkably consistent evaluations of textual content which, while not in perfect agreement, are substantially more so than we expect from human readers doing the same job.

Addressing Potential Concerns about "Data Leakage"

A significant concern when using LLMs to estimate policy positions is *data leakage*—which could arise when a model draws on information it memorized during training rather than interpreting the input text. For instance, an LLM might recall a party's EU position from training data instead of inferring it from a manifesto summary, inflating performance and overstating the model's ability to simulate expert judgments. We took several steps to assess and minimize this risk.

First, neither the data we supplied to the LLMs for positional scaling nor the expert benchmarks used for evaluation were part of the models' training data. The policy summaries were created specifically for this study and could not have been seen by the models during training. Our prompts explicitly instructed the LLMs to base their answers solely on these summary texts, with no reference whatsoever to benchmark datasets (CHES or others). While LLMs are trained on a wide range of general content, including academic literature on expert surveys, they do not have access to the expert survey datasets themselves.[22] Furthermore, the coalition policy agreement analyses reported in the next section could not be part of any training data, as the issue position estimates we compute as benchmarks for these have never been published. Finally, the 7-point, 7-anchor, scales we asked LLMs to use differ substantially from those used in the benchmark datasets (11-point for CHES, 20-point for Benoit-Laver and Laver-Hunt, in each case with just two anchors), making direct reproduction of benchmark values unlikely, even if somehow the models had encountered them during training.

Second, if LLMs were simply regurgitating memorized expert survey scores, we would expect strong correlations across all dimensions, including decentralization. Instead, decentralization shows substantially lower correlations with expert benchmark across all six LLMs we deployed. This pattern, along with the alignment between LLM estimates and manifesto-based scores from the MP shown in Figure 3, provides strong evidence that the language models are not retrieving benchmark issue positions from their training data but instead are basing their answers on interpreting the input summaries.

Finally, as a robustness check against possible cueing effects from party

[22] The LLMs themselves support this claim. When we asked Gemini about this, it replied: 'I would not be "drawing on" or "recalling" specific CHES scores for that party from my training data to answer your questions about the manifesto. My analysis would be grounded in the textual evidence of the manifesto itself. The knowledge about CHES and similar frameworks helps me understand the *type* of questions you're asking and the *nature* of the policy dimensions, but the content of my answer about the manifesto's positions would be derived from the manifesto.' See Supplemental Appendix H for full our chats with three LLMs on this matter.

identification, we created an edited version of each summary in which all party names and abbreviations were removed. Since a summary cannot trigger memorized knowledge about a party's policy positions if it is not identifiable, this anonymization provides a strong test for leakage. Just under half of the original summaries contained such identifiers. We used these anonymized summaries to generate a new set of issue position scores (using GPT-4o and zero-shot learning only) and compared these with the originals. The results were virtually identical, a correlation of 0.99, indicating that the presence of party names had no meaningful influence on the model's outputs. (Full results are shown in Supporting Information Appendix I.)

In conclusion, while data leakage is a legitimate concern when benchmarking LLMs' scoring of issue positions, the lack of access of LLMs to expert datasets, the consistently lower correlations for decentralization, and our robustness checks using anonymized summaries combine to provide strong evidence that our results are not meaningfully affected by leakage.

Replicating the LLM analyses

Concern has been expressed (among others, by Barrie, et al. 2024), about the potential for a "replication crisis" in research based on LLMs—that LLMs, for various reasons, do not provide the machinery for researchers to replicate the data they generate. In political science, the term *replication* is often used to describe what other disciplines refer to as *reproducibility*—that is, the ability to obtain identical results by re-running the original code and data through the same computational pipeline. By contrast, the more general scientific usage of *replicability* refers to obtaining consistent results using different tools or data under similar research conditions (National Academies of Sciences, et al. 2019, p6; Breuer and Haim 2024). We assess reproducibility by re-running our original analysis using the same LLMs. We assess replicability by repeating the analysis with three different, open-weight

LLMs not used in the baseline study.

To test the *reproducibility* of our procedure using the same proprietary LLMs, we re-ran our entire analysis following a three-month wait, using the latest versions of the same LLMs (see Supplemental Appendix E for full details). The results, reported in the left-hand panel of Table 6, show reproducibility far superior to any established human standard. In our reproduction run, correlations between pairs of LLM estimates ranged from 0.92 for the Environment to 0.96–0.98 for all other issues except Decentralization, which replicated at “only” 0.69. This is not surprising given the noise in this issue consistent with Table 3 above. These conform to standards for much computational research, when “bitwise reproducibility may be relaxed and reproducible results could be obtained within an accepted range of variation” (National Academies of Sciences, et al. 2019, p8). We conclude our findings are reproducible.

To test *replicability* using a different data production method, we re-ran our entire analysis using entirely different LLMs. We picked three high-performing open-weight models: DeepSeek-V3; Llama-3.3; Gemma-3. The results, reported in the right-hand panel of Table 6, are extremely encouraging. Comparing the two panels, the open weight models correlate with the expert benchmark only slightly less well than the proprietary models, as well as replicating the substantially lower correlations for decentralization. The central panel of Table 6 shows that correlations between the proprietary ensemble and the open-weight ensemble are very high, at about 0.95. Results generated by one class of LLM are replicating results generated by the other.

[TABLE 6 HERE]

This is a particularly important finding for considerations of the costs and benefits of using open-weight or proprietary models. For this application at least, the benefits of using open-weight models—in terms of confidentiality, transparency and replicability—seem to come at very little cost in terms of results. The key question concerns whether, notwithstanding inevitable stochastic components in the data generation process, the data generated from successive replications support the same general inferences. We find this is the case for our LLM-generated data.

PREDICTIVE VALIDITY: LLM ESTIMATES OF COALITION ISSUE POSITIONS

Having substantively validated our LLM method using expert survey data, we use it to generate a dataset for which there is no expert benchmark. This measures policy positions of *coalition governments* on the six dimensions of interest, set out in their agreed policy declarations. Klüver et al. assembled a collection of 229 declarations issued by coalition cabinets in Eastern and Western Europe between 1945 and 2015 (Klüver, et al. 2023). They manually unitized and labeled these using the MP coding scheme. We downloaded these agreements and scored them on our six dimensions of interest, using our 3x3 design for both zero-shot learning as well as few-shot learning using the same prompt examples employed in the party manifesto analysis.

Limited overlap between start and end dates of each project, combined with our selection of party manifestos to coincide with years in which the benchmark expert surveys were fielded, left 23 coalition agreements in the Klüver et al. collection for which we also have LLM estimates of member party positions. Supplemental Appendix Table F1 lists these. We used Kluver et al. labels to generate logit scales of coalition policy on the six dimensions of interest, using the method described in Supplemental Appendix A2. For each coalition agreement, for each policy dimension of interest, we therefore have LLM and Klüver et al. estimates of *coalition* policy, and LLM and MP estimates of *member-party* policy positions.

We have no benchmark estimate of coalition policy positions. Indeed, there is no published estimate of these positions on the six issue dimensions of interest. We do, however, have a well-founded *theoretical* expectation about these. There are many different spatial models of government formation, dating back over 50 years of theoretical and empirical scholarship on coalition government. While these models differ in their precise predictions of agreed coalition policy, to the best of our knowledge they all converge on one general prediction. On any given issue dimension, *coalition policy should be within the range of member parties' policy positions.* This is because, if coalition policy is *outside* the range of members' policy positions, then moving it into that range generates a Pareto improvement for all coalition members.[23]

[TABLE 7 ABOUT HERE]

We compare LLM-generated estimates with Klüver/MP hand-coded estimates of party and coalition policy positions against this theoretical expectation. Table 7 shows clear results. Except for decentralization, LLM-generated estimates of coalition policy positions are significantly more likely to fall within the range of members parties' policy positions than those produced by MP-style hand-labeling.

It is not our purpose to assess content validity of the hand-coding of coalition agreements using the MP scheme, and we cannot say that out-of-range estimates of coalition policy are definitively "wrong." It is possible, for example, to imagine an idiosyncratic situation in which all member parties promise something completely infeasible in their manifestos—"deport all immigrants" for example—that does not make it into a government declaration. But this caveat applies to both types of estimate. We find it encouraging, therefore, that LLM-generated estimates conform much more closely to theoretical expectations than those generated by methods based on labeling and counting sentences.

CONCLUSIONS: THE REVOLUTIONARY POTENTIAL OF LLMs

Our aim is to evaluate a new method for analyzing political text, seeking reliable and valid

[23] Government formation models differ on *where within this range* they predict coalition policy.

inferences about the stated preferences of its author.[24] An ideal method would recruit a large sample of experts with profound knowledge of both the language and the political setting involved. These would be asked to read the relevant text for meaning, then respond to carefully crafted questions about the preferences of the author on issue areas of interest. We'd need a large sample rather than a single expert because we expect experts to differ in their judgements about positions expressed in any given text.[25] This approach is infeasible for even a single text, let alone a large corpus of texts in many different languages dealing with many different political settings.

The current state of the art offers two approaches to reliable and replicable text analysis at scale. Quantitative text-as-data (TaDa) methods are based on supervised machine learning algorithms of ever-increasing sophistication, but are grounded in flawed assumptions about how quantitative patterns in text can be used to infer qualitative meaning. TaDa methods scale to massive text corpora and are now a tried and tested mainstream approach to analyzing political text–until recently, the only option for replicable text analysis at scale.

The new alternative is the portfolio of NLU tasks we assess here, implemented by LLMs. This extends the TaDa paradigm from statistical representation to interpretative modeling, generating results which are functionally equivalent to those produced by human qualitative text analysis—but in many languages and at potentially massive scale. Casual use of LLMs to analyze political texts has increased rapidly with their exploding power and availability. Our concern here is to develop reliable and valid methods for using LLMs *as a scientific tool*. If shown to be reliable and valid, using LLMs to summarize and interpret texts by drawing on the collective wisdom ingested in their massive training data, comes closest to the ideal of drawing on the collective wisdom of experts.

[24] The systematic relationship between "stated" and fundamentally unobservable "true" preferences can only be established using theoretical models of politics which do not concern us here.
[25] Consider the dramatically varying expert interpretations of any classic political text.

Our results demonstrate that LLMs can effectively perform this task. Correlations between LLM ensemble means and benchmark expert surveys are as high as we can reasonably expect for most issue dimensions, matching the upper bound of agreement between different expert surveys. Given a task for which there is no expert benchmark, the LLM ensemble produced coalition policy estimates with strong predictive validity, outperforming human-coded text labeling of the texts of coalition agreements.

Concerns about the scientific use of LLMs—to do with reproducibility, replicability, transparency, bias and "leakage"—are understandable. We address the challenge of *reproducibility* by almost perfectly replicating all our results after a three-month interval, using updated versions of each proprietary LLM. We evaluate *replicability* by rerunning our entire analysis with three completely different LLMs—open weight-models with versions which can be downloaded and will not evolve over time. Again, we get almost identical results. These latter results are particularly significant, since many scholars may have a principled preference for open weights rather than "closed" proprietary models, and we show here they work equally well.

Using "black box" LLMs to replicate qualitative text analysis is less *transparent* than quantitative TaDa methods, which implement explicit algorithmic or statistical procedures, typically using open-source software. Yet the cognitive processes of qualitative human coders and expert survey respondents are arguably even more opaque. While LLMs can be prompted to provide detailed justifications for their outputs, human text annotators are almost never asked to do this, and would likely find it challenging if asked to do so.

A related and distinctive contribution of our approach is the production of concise policy summaries in the first stage of our analysis. In contrast to sentence-level classification approaches, which segment text into isolated labelling decisions, our method begins with a holistic summarization stage that yields interpretable, document-level representations of policy content. These summaries are themselves a valuable research output: they not only enhance interpretability and transparency but also provide a reusable textual representation of policy content that could support further analysis. Our results—and the extensive replication materials accompanying this paper—should encourage future work on using, evaluating, and standardising such summaries.

While LLMs may still operate as black boxes, we have shown that the summary-and-scaling pipeline yields outputs that are systematic, replicable, and substantively valid. An ensemble of models, applied across diverse issue areas, consistently maps textual inputs to policy position estimates in a reliable and interpretable manner.

To evaluate potential *bias*, we need a ground truth against which bias can be measured—something fundamentally unobservable when it comes to the "true" issue positions of politicians. Using ensembles of experts as a surrogate for ground truth, we found strong evidence of "bias" for decentralization, some evidence of this for environmental policy, and none for the other four issue areas. LLMs tended to conclude that party manifestos articulated systematically more pro-decentralization and pro-environmental policies than those identified by experts. However, when comparing LLM estimates to those generated by the MP, this seems to reflect "structural bias" inherent in manifesto content rather than bias attributable to LLMs.

Substantially lower correlations between LLM estimates and the expert benchmark for decentralization policy supplement other strong evidence against the possibility of "leakage", in which LLMs retrieve information from their training data rather than analyzing the manifestos *de novo*. If that were not the case, there would be no reason for the models to perform markedly worse for decentralization than for the five other issue areas we study.

Developing scientifically valid methods for using LLMs in political text analysis is crucial because these powerful tools are widely accessible, easy to use, multilingual, and cheap. Their potential is transformative, enabling motivated scholars, regardless of their geographic locations or funding constraints, to conduct sophisticated, targeted analyses without the massive resources traditionally required for large-scale qualitative text analysis. We have demonstrated how, with minimal resources and in a short timeframe, it is possible to replicate the outputs of extremely expensive and time-consuming projects reliant on human experts, such as the MP and CAP. By leveraging LLMs for automated text analysis, underfunded junior researchers can generate the specific text data they need for their work, rather than relying on legacy datasets of limited relevance.

We offer a demonstration rather than a technical analysis of LLMs and their potential in the analysis of political text. Using state-of-the-art LLMs available at the time of writing, we achieved results comparable to those from projects reliant on human experts. By the time this article is published, ongoing advances in this rapidly moving field will likely have produced a new generation of LLMs that are not only significantly more powerful than these, but also even easier to use. The strong results we report here will only get stronger as LLMs become more powerful. Love them or hate them, LLMs are poised to drive a new measurement paradigm in political science, making it essential to establish clear and robust protocols to ensure their valid and replicable use.

REFERENCES

Abburi, Harika, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. "Generative Ai Text Classification Using Ensemble Llm Approaches." *arXiv preprint arXiv:2309.07755*.

Bakker, Ryan, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2015. "Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999–2010." *Party Politics* 21: 143-52.

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2024. "Replication for Language Models Problems, Principles, and Best Practice for Political Science." *URL: https://arthurspirling.org/documents/BarriePalmerSpirling TrustMeBro.pdf*.

Battle, Rick, and Teja Gollapudi. 2024. "The Unreasonable Effectiveness of Eccentric Automatic Prompts." *arXiv preprint arXiv:2402.10949*.

Baumgartner, Frank R, Christian Breunig, and Emiliano Grossman. 2019. "The Comparative Agendas Project: Intellectual Roots and Current Developments."

Baumgartner, Frank R, Christoffer Green-Pedersen, and Bryan D Jones. 2006. "Comparative Studies of Policy Agendas." *Journal of European Public Policy* 13: 959-74.

Benoit, Ken. 2020. "Text as Data: An Overview." *The SAGE handbook of research methods in political science and international relations*: 461-97.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American political science review* 110: 278-95.

Benoit, Kenneth, and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Bol, Damien, and Pierre-Henri Bono. 2025. "Can ChatGPT Accurately Identify the Position of Parties? A Validation Study with an Expert Survey in France." *Research & Politics* 12: 20531680251335653.

Breuer, Johannes, and Mario Haim. 2024. "Are We Replicating Yet? Reproduction and Replication in Communication Research." *Media and Communication* 12.

Bruno, A, PL Mazzeo, A Chetouani, M Tliba, and MA Kerkouri. 2023. "Insights into Classifying and Mitigating LLMs' Hallucinations." Arxiv.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. 2001. *Mapping Policy Preferences: Parties, Electors and Governments: 1945-1998: Estimates for Parties, Electors and Governments 1945-1998.* Oxford: Oxford University Press.

Budge, Ian, David Robertson, and Derek Hearl. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Burla, Laila, Birte Knierim, Jurgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel. 2008. "From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis." *Nursing Research* 57: 113-17.

Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. "Booookscore: A Systematic Exploration of Book-Length Summarization in the Era of Llms." *arXiv preprint arXiv:2310.00785*.

Chase, Harrison. 2022. "Langchain, October 2022." *URL https://github.com/langchain-ai/langchain*.

Chrysostomou, George, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. "Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization." *Transactions of the Association for Computational Linguistics* 12: 1163-81.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57: 535-74.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* *21*(3), 267-297.

Hosseini, Peyman, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. 2024. "Efficient Solutions for an Intriguing Failure of Llms: Long Context Window Does Not Mean Llms Can Analyze Long Sequences Flawlessly." *arXiv preprint arXiv:2408.01866*.

Hoyle, Alexander, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. "Natural Language Decompositions of Implicit Content Enable Better Text Representations." *arXiv preprint arXiv:2305.14583*.

Jolly, Seth, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. "Chapel Hill Expert Survey Trend File, 1999–2019." *Electoral Studies* 75: 102420.

Karjus, Andres. 2023. "Machine-Assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence." *arXiv preprint arXiv:2309.14379*.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003.* Oxford: Oxford University Press.

Klüver, Heike, Hanna Bäck, and Svenja Krauss. 2023. *Coalition Agreements as Control Devices: Coalition Governance in Western and Eastern Europe*: Oxford University Press.

Krippendorff, Klaus. 2018. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

Laver, Michael. 2003. "Why Should We Estimate the Policy Positions of Political Actors?" In *Estimating the Policy Position of Political Actors*: Routledge. 3-9.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. New York: Routledge.

Le Mens, Gaël, and Aina Gallego. 2025. "Positioning Political Texts with Large Language Models by Asking and Averaging." *Political Analysis*: 1-9.

Linegar, Mitchell, Rafal Kocielnik, and R Michael Alvarez. 2023. "Large Language Models and Political Science." *Frontiers in Political Science* 5: 1257092.

Liu, Nelson F, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. "Lost in the Middle: How Language Models Use Long Contexts, 2023." *URL https://arxiv.org/abs/2307.03172*.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36: 123-55.

Merz, Nicolas, Sven Regel, and Jirka Lewandowski. 2016. "The Manifesto Corpus: A New Resource for Research on Political Parties and Quantitative Text Analysis." *Research & Politics* 3: 2053168016643346.

Mikhaylov, Slava, Michael Laver, and Kenneth R Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20: 78-91.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." Paper presented at the Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies.

Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods*: 1-18.

Samant, Rahul Manohar, Mrinal R Bachute, Shilpa Gite, and Ketan Kotecha. 2022. "Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions." *IEEE Access* 10: 17078-97.

Sciences, National Academies of, Medicine, Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, and Theoretical Statistics. 2019. *Reproducibility and Replicability in Science*: National Academies Press.

Snow, Rion, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng. 2008. "Cheap and Fast–but Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks."

Paper presented at the Proceedings of the 2008 conference on empirical methods in natural language processing.

Sujatha, R, and K Nimala. 2024. "Classification of Conversational Sentences Using an Ensemble Pre-Trained Language Model with the Fine-Tuned Parameter." *Computers, Materials & Continua* 78.

Volkens, Andrea, Judith Bara, Ian Budge, Michael D McDonald, and Hans-Dieter Klingemann. 2013. *Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts*. Vol. 3: OUP Oxford.

Wang, Hengyi, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2024. "Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models." *arXiv preprint arXiv:2406.11230*.

Weber, Maximilian, and Merle Reichardt. 2023. "Evaluation Is All You Need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer Using Open Models." *arXiv preprint arXiv:2401.00284*.

Wu, Patrick Y, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2024. "Concept-Guided Chain-of-Thought Prompting for Pairwise Comparison Scoring of Texts with Large Language Models." Paper presented at the 2024 IEEE International Conference on Big Data (BigData).

Yang, Han, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. "One Llm Is Not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering." *medRxiv*.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. "Benchmarking Large Language Models for News Summarization." *Transactions of the Association for Computational Linguistics* 12: 39-57.

1. **Input texts and validation data**
   a. *Text corpus:*
      i. Party manifestos from the MP
      ii. Coalition agreements from Kluwer et al (2023)
   b. *Positions on six issue dimensions from expert surveys fielded in election years:*
      i. Laver and Hunt (1992); Benoit and Laver (2006)
      ii. Chapel Hill expert survey (CHES)
   c. *Issue positions on same six dimensions* derived from scaling hand-coded labels:
      i. MP data for party manifestos
      ii. Kluwer et al (2023) data for coalition agreements

2. **LLM authorization**
   Establish high-tier API access to three commercial LLMs: OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, Google's Gemini 1.5 Pro.

3. **Summarize texts**
   a. Present original language texts to LLMs for processing.
   b. Prompt LLMs to generate 200-300 word English language summaries of each text's content relating to each of the six issue dimensions.

4. **Scale text summaries**
   Prompt LLMs to locate positions of each summary on specified 7-point scales
   a. for "zero-shot" learning (no examples)
   b. for "few-shot" learning (adding three sample summaries and associated benchmark positions for each scale)

5. **Validation**
   a. *Concurrent validity:* Correlate ensemble means of LLM estimates of party positions with ensemble means of expert judgements from 1b.
   b. *Reliability:* As a benchmark for reliability, compare LLM-expert correlations with expert-expert correlations.
   c. *Predictive validity:* Compare ensemble means of LLM estimates of coalitions' issue positions with analogous hand-coded estimates from 1c, to see which better accords with a prediction common to almost all models of government formation.

6. **Replication and Robustness Checks**
   Repeat steps 2-5a to see if results are stable.
   Replicate results using three open-weight LLMs: Deepseek-V3 (671B), Llama-3.3 (70B), and Gemma-3-27B-it.

Table 1: AI-powered text analysis workflow

ECONOMIC: position on improving public services vs. reducing taxes.
0 = Strongly favors improving public services
10 = Strongly favors reducing taxes

SOCIAL: position on social lifestyle
0 = Strongly supports liberal policies
10 = Strongly opposes liberal policies

IMMIGRATION: position on immigration policy.
0 = Strongly opposes tough policy
10 = Strongly favors tough policy

EU: overall orientation of the party leadership towards European integration
1 = Strongly opposed
7 = Strongly in favor

ENVIRONMENT: position towards the environment.
0 = Environmental protection even at the cost of economic growth
10 = Economic growth even at the cost of environmental protection

DECENTRALIZATION: position on political decentralization to regions/localities.
0 = Strongly favors political decentralization
10 = Strongly opposes political decentralization

Table 2. Expert survey dimensions constant over time.

Issue dimension	Mean	95% CI
Economic	0.88	[0.85, 0.90]
Social	0.91	[0.89, 0.93]
Immigration	0.88	[0.86, 0.91]
EU	0.95	[0.93, 0.96]
Environment	0.84	[0.82, 0.87]
Decentralization	0.78	[0.74, 0.82]

Table 3: Upper bounds of human reproducibility: Pearson correlations between split-sample expert surveys.

	Proportion of Missing LLM Scores			
Issue dimension	**Claude 3.5**	**GPT-4o**	**Gemini 1.5 Pro**	**Overall**
Economic	0.13	0.02	0.08	0.08
Social	0.24	0.21	0.38	0.28
Immigration	0.42	0.24	0.27	0.31
EU	0.16	0.03	0.05	0.08
Environment	0.12	0.01	0.04	0.06
Decentralization	0.20	0.01	0.13	0.11
All Issues	0.21	0.09	0.16	

Table 4: Proportion missing LLM scores contributing to ensemble mean.

	Intra-LLM		Inter-LLM	
Issue dimension	**ICC**	**Kripp α**	**ICC**	**Kripp α**
Economic	0.90	0.90	0.91	0.90
Social	0.95	0.91	0.97	0.91
Immigration	0.93	0.93	0.93	0.92
EU	0.92	0.87	0.92	0.88
Environment	0.80	0.79	0.76	0.74
Decentralization	0.66	0.62	0.69	0.57
Overall	0.91	0.90		

Table 5: Agreement rates of LLM positional ratings.

ICC is the Inter-class correlation coefficient; "Kripp α" is Krippendorff's alpha for ordinal data.

Issue dimension	Reproducibility			Replicability			
	Proprietary LLMs			Open-Weight LLMs			
	Second Ensemble with Original Ensemble	Correlation with Expert Surveys		Open-weight ensemble with proprietary Ensemble	Correlation with Expert Surveys		
		Replication	Original		LLaMA	Deepseek	Gemma
Economic	0.97	0.87	0.87	0.95	0.84	0.84	0.86
Social	0.97	0.92	0.90	0.95	0.87	0.87	0.86
Immigration	0.96	0.89	0.89	0.93	0.86	0.89	0.89
European Union	0.98	0.91	0.91	0.96	0.86	0.86	0.84
Environment	0.92	0.82	0.76	0.92	0.68	0.79	0.86
Decentralization	0.69	0.49	0.40	0.78	0.40	0.45	0.45

Table 6: Reproduction and Replication of LLM-Generated Results.
All values shown are Pearson product-moment correlations. The open-weight models are LLaMa-3-70B, Deepseek-V3 (671B), Gemma-3-27B-it.

	LLM Mean		Kluwer and MP logit Scores	
Issue dimension	**Prop. Inside**	**N missing**	**Prop. Inside**	**N missing**
Economic	0.76	2	0.36	1
Social	0.75	3	0.45	1
Immigration	0.53	4	0.14	9
EU	0.57	0	0.45	1
Environment	0.78	0	0.18	1
Decentralization	0.43	0	0.61	5
Overall	0.64		0.38	

Table 7: Proportions of 23 coalition policy positions within the range of members' positions.[26]

[26] LLM mean is the proportion of coalition documents whose positions lie within between the left-most and right-most coalition members, measured by LLMs; Kluwer and MP scores are the same using logit scores from manifesto scoring counts, from the MP data and Kluwer et al.

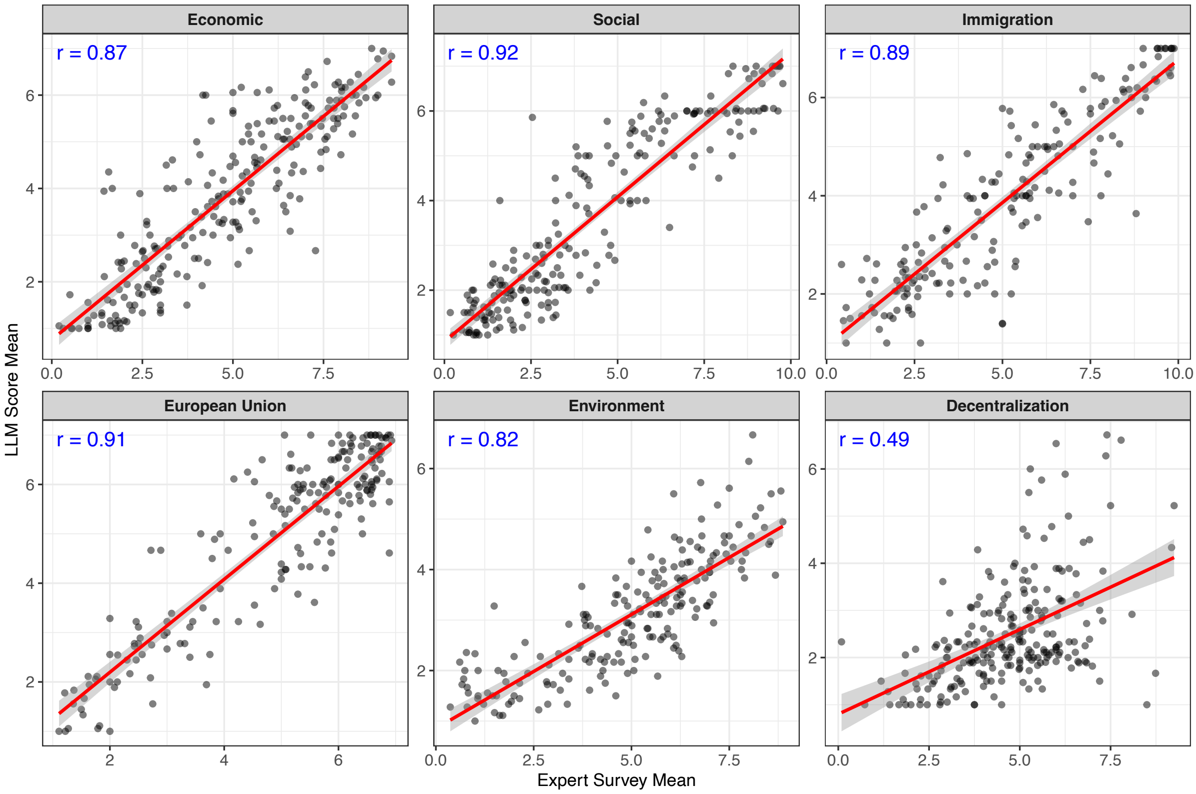

Figure 1: Plot of LLM ensemble of 18 scores vs expert means, by issue.

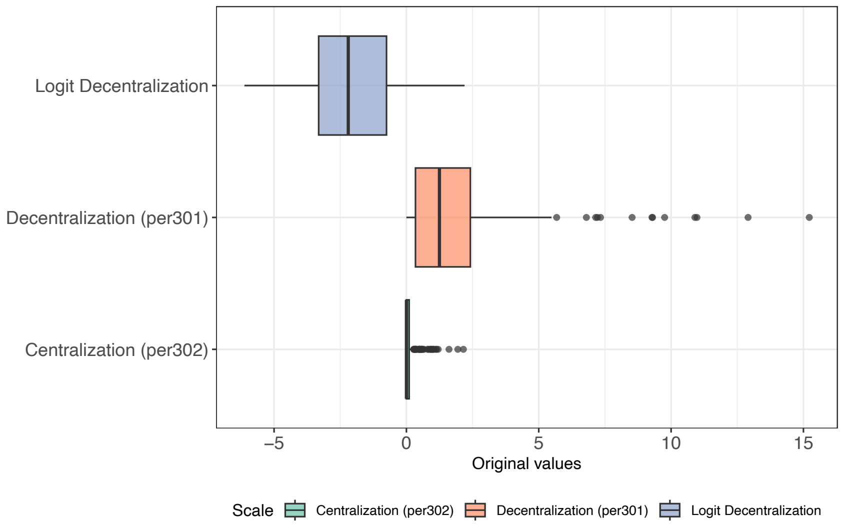

Figure 2: Distributions of MP-estimated manifesto percentages on (de)centralization for manifestos in the LLM corpus.

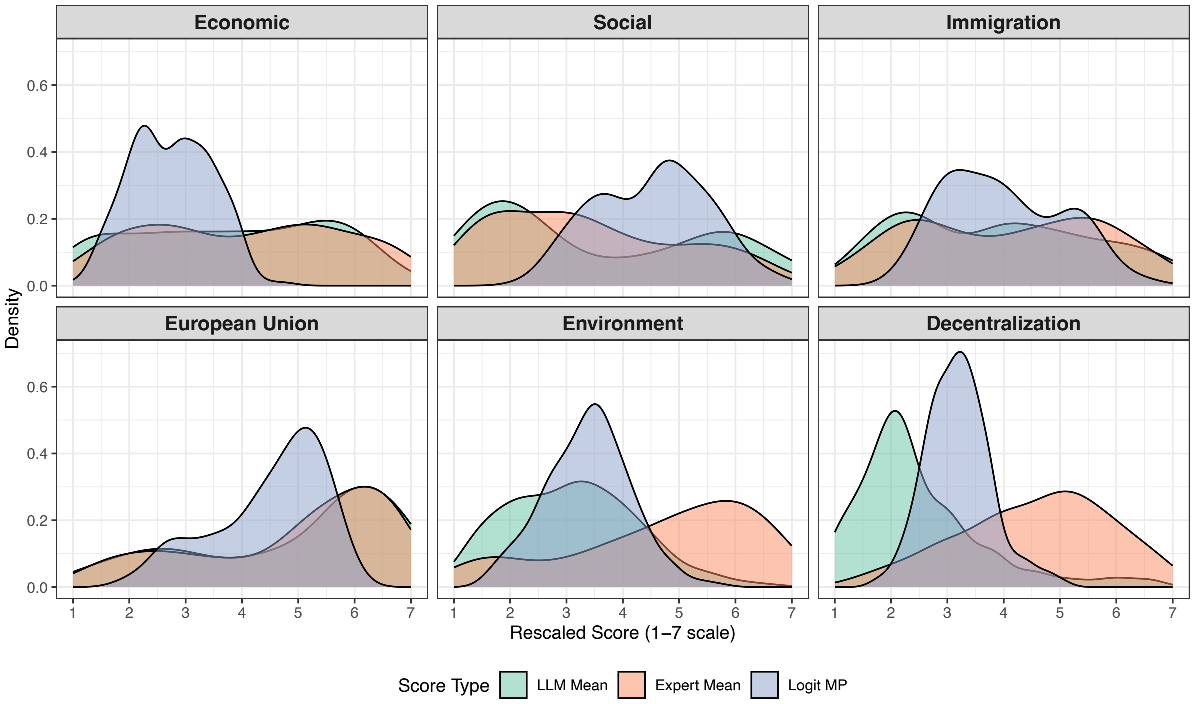

Figure 3: Distributions of LLM 18-score ensemble, expert and MP scores, by issue.

USING LARGE LANGUAGE MODELS TO ANALYZE POLITICAL TEXTS THROUGH NATURAL LANGUAGE UNDERSTANDING

SUPPORTING INFORMATION

Kenneth Benoit, Scott De Marchi, Conor Laver, Michael Laver, Jinshuai Ma

Table of Contents

APPENDIX A1: PARTIES AND ELECTION YEARS INCLUDED IN THE STUDY

We excluded settings where there is no manifesto in the MP collection, or where the only version of the manifesto does not have machine readable text.

Country	Election Year	Expert Survey Source	Expert Survey Year	Number of Parties
Austria	2006	CHES	2006	4
Austria	2019	CHES	2019	5
Belgium	2003	BL	2003	6
Belgium	2010	CHES	2010	10
Belgium	2014	CHES	2014	11
Belgium	2019	CHES	2019	6
Bulgaria	2014	CHES	2014	7
Croatia	2003	BL	2003	7
Czech Republic	2006	CHES	2006	5
Czech Republic	2010	CHES	2010	6
Denmark	2019	CHES	2019	10
Estonia	2019	CHES	2019	5
Finland	2003	BL	2003	8
Finland	2019	CHES	2019	8
Greece	2019	CHES	2019	7
Hungary	2006	CHES	2006	4
Hungary	2010	CHES	2010	5
Hungary	2014	CHES	2014	6
Iceland	2003	BL	2003	5
Ireland	1989	LH	1989	5
Netherlands	1989	LH	1989	7
Netherlands	2003	BL	2003	7
Netherlands	2006	CHES	2006	8
Netherlands	2010	CHES	2010	10
Norway	1989	LH	1989	7
Poland	2019	CHES	2019	6
Portugal	2019	CHES	2019	7
Slovakia	2006	CHES	2006	6
Slovakia	2010	CHES	2010	6
Slovenia	2014	CHES	2014	7
Spain	1989	LH	1989	5
Sweden	2006	CHES	2006	7
Sweden	2010	CHES	2010	8
Sweden	2014	CHES	2014	8
United Kingdom	2019	CHES	2019	6
Total				235

Table A1: Parties and election years included in the study, with corresponding expert surveys. CHES = Chapel Hill Expert Survey (Jolly et al. 2022); LH = Laver and Hunt (1992); BL = Benoit and Laver (2006).

APPENDIX A2: MANIFESTO PROJECT SCALES TO MATCH CHES AND LLM SCALES

We generated scales using MP (PER) coding categories substantively close to scale definitions used in our LLM and the expert survey estimates. We constructed the following as logit scales, based on recommendations in Lowe et al (2010), i.e.: $\theta^{(L)} = \log(R + 0.5) - \log(L + 0.5)$.

Economic
per504 Welfare State Expansion
per505 Welfare State Limitation
per506 Education Expansion
per507 Education Limitation
welfare_expand = log(per505 + per507 + 0.5) - log(per504 + per506 + 0.5)

Social
per603 Traditional Morality: Positive
per604 Traditional Morality: Negative
tradmoral_hi = log(per603 + 0.5) - log(per604 + 0.5)

Immigration (only in MP since 2014)
per601_2 National Way of Life: Immigration: Negative
per602_2 National Way of Life: Immigration: Positive
immig = log(per601_2 + 0.5) - log(per602_2 + 0.5)

EU
per108 European Community/Union: Positive
per110 European Community/Union: Negative
pro_eu_hi = log(per108 + 0.5) - log(per110 + 0.5)

Environment
per410 Economic Growth: Positive
per501 Environmental Protection
envir_v_growth = log(per410 + 0.5) - log(per501 + 0.5)

Decentralization
per301 Decentralization
per302 Centralization
decentral = log(per302 + 0.5) - log(per301 + 0.5)

APPENDIX A3: SUMMARIZER PROMPTS

You are an expert political analyst. Please summarize what the following political manifesto has to say about

[insert issue from list]

You should detect the original language and output a concise summary in English of about 300 - 400 words, dealing only with the manifesto's discussion of this issue.

[Load manifesto]

Issue list

- Taxation, spending on public services, and tradeoffs between these
- Social and lifestyle issues, including, abortion, LBTQ+ issues; support for traditional social values
- Immigration: immigration and border control
- The European Union and European integration
- Environmental protection: tradeoffs between protecting the environment and economic growth
- Decentralization of political power and the role of regional governments

APPENDIX A4: SEVEN-POINT SCALER PROMPTS FOR SIX ISSUE DIMENSIONS

Taxes versus spending

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party position on the tradeoff between higher spending on public services and lower taxation, using the following seven-point scale:

1. **High Public Spending, High Taxation**: The party strongly advocates for high levels of public spending on public services, accompanied by high taxation to fund this expenditure. This position prioritizes extensive public services and welfare programs, emphasizing the role of the government in providing these services.
2. **Moderately High Public Spending, High Taxation**: The party supports relatively high levels of public spending and is comfortable with high taxation to support this. While not as extreme as position 1, it still prioritizes significant public investment and welfare services, albeit with some consideration for tax efficiency.
3. **Moderate Public Spending, Moderate to High Taxation**: The party advocates for a balanced approach with moderately high public spending and corresponding taxation. It aims to maintain robust public services but with a careful approach to avoid excessive taxation.
4. **Balanced Approach**: The party supports a balanced mix of public spending and taxation. It advocates for a pragmatic approach, where public services are funded adequately without excessively high or low taxation. This position seeks to strike a middle ground between providing services and keeping taxes reasonable.
5. **Moderate Public Spending, Moderate to Low Taxation**: The party favors moderately low levels of public spending and seeks to reduce taxation to a moderate extent. This position leans towards reducing the size of public services to decrease the tax burden but still maintains some level of essential public services.
6. **Moderately Low Public Spending, Low Taxation**: The party supports low levels of public spending on public services and correspondingly low levels of taxation. This position emphasizes reducing the role of government in providing services and seeks to minimize the tax burden on individuals and businesses.
7. **Low Public Spending, Low Taxation**: The party strongly advocates for minimal public spending on public services and very low taxation. This position prioritizes individual responsibility and the free market, with minimal government intervention in providing services.

If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

Social lifestyle

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party position in terms of liberalism or conservatism on social lifestyle issues (including abortion, gay rights, and other related matters), using the following seven point scale:

1. **Very Liberal**: The party strongly advocates for liberal social policies, including full support for abortion rights, comprehensive LGBTQ+ rights (including marriage equality, adoption rights, and anti-discrimination laws), and other progressive social issues. It emphasizes individual freedom and equality in all aspects of social life and seeks to advance and protect these rights through robust legislative measures.
2. **Liberal with Progressive Policies**: The party supports liberal social policies, promoting significant rights and protections for individuals in areas such as abortion and LGBTQ+ rights. It advocates for accessible abortion services, strong anti-discrimination laws, and broad recognition of LGBTQ+ relationships and family structures. While not as radical as the most liberal position, it still prioritizes progressive changes and protections.
3. **Moderately Liberal**: The party supports liberal social policies but with some moderate constraints. It advocates for legal abortion with certain regulations, supports LGBTQ+ rights including marriage equality and anti-discrimination measures, and promotes other progressive social policies. This position seeks a balance between individual freedoms and societal norms, aiming for broad but not unrestricted liberalization.
4. **Balanced Approach**: The party takes a centrist position on social lifestyle issues, seeking to balance liberal and conservative views. It supports some liberal policies such as legal abortion with restrictions and basic LGBTQ+ rights while also accommodating certain conservative perspectives. This position aims to find common ground and promote social harmony.
5. **Moderately Conservative**: The party favors conservative social policies with some liberal allowances. It supports more restrictive abortion laws, such as limiting access to later-term abortions, and may support certain LGBTQ+ rights but oppose others like adoption or marriage equality. This position emphasizes traditional values while allowing for limited individual freedoms.
6. **Conservative with Traditional Values**: The party strongly supports conservative social policies, advocating for significant restrictions on abortion (possibly aiming to make it illegal or highly restricted) and opposing many LGBTQ+ rights such as marriage equality and adoption. It promotes traditional family structures and social norms, emphasizing the preservation of established societal values.
7. **Very Conservative**: The party strongly advocates for highly conservative social policies, including making abortion illegal in most or all cases, opposing LGBTQ+ rights (including marriage, adoption, and anti-discrimination protections), and promoting traditional, often religiously-informed, social values. This position prioritizes preserving traditional societal norms and minimizing changes to established social structures.

   If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

Immigration

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party position on immigration, using the following seven point scale:

1. **Very Open Immigration**: The party advocates for highly liberal immigration policies. This includes few restrictions on immigration, support for high levels of immigration, and extensive integration programs for immigrants. The party may support policies such as open borders, easy pathways to citizenship, and generous refugee resettlement programs.
2. **Moderately Open Immigration**: The party supports liberal immigration policies but with some limitations. It advocates for significant levels of immigration and comprehensive integration programs but includes certain restrictions to control the flow and ensure security. Policies may include streamlined visa processes, expanded refugee programs, and proactive immigrant support services.
3. **Balanced Open Immigration**: The party favors relatively open immigration policies with balanced controls. It supports moderate to high levels of immigration and substantial integration efforts while maintaining specific measures to manage immigration levels and ensure societal integration. Policies may include welcoming skilled workers, family reunification programs, and moderate refugee intake.
4. **Balanced Approach**: The party supports a balanced approach to immigration, with a mix of openness and control. It advocates for moderate immigration levels, focusing on both economic needs and social integration. Policies include a points-based immigration system, selective refugee intake, and targeted integration programs.
5. **Balanced Restriction**: The party favors a more controlled immigration policy with moderate restrictions. It supports lower levels of immigration compared to an open approach, emphasizing security and economic impact. Policies may include strict visa requirements, limited refugee intake, and prioritizing skilled immigration.
6. **Moderately Restrictive Immigration**: The party advocates for restrictive immigration policies with significant limitations. It supports low levels of immigration and emphasizes national security, cultural integration, and economic self-sufficiency. Policies may include stringent border controls, capped immigration quotas, and limited pathways to citizenship.
7. **Very Restrictive Immigration**: The party strongly supports highly restrictive immigration policies. This includes very low levels of immigration, strict border controls, and minimal refugee intake. The party may advocate for policies such as severe visa limitations, extensive vetting processes, and a focus on repatriation and deportation of illegal immigrants.

If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

European Union

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party position on the European Union, using the following seven point scale:

1. **Strong Pro-EU Integration**: The party advocates for deep integration with the EU, including support for further political and economic union. This position favors policies like adopting the euro (if not already adopted), further ceding national sovereignty to EU institutions, and strong support for EU-wide policies and regulations. The party is likely to support increased EU powers and closer cooperation in areas such as defense, foreign policy, and fiscal matters.
2. **Pro-EU with Moderate Integration**: The party supports the EU and further integration, but with some reservations. It advocates for increased cooperation and integration in certain areas but prefers to retain more national control in others. This position supports the EU's current framework and seeks to enhance it, but without fully endorsing all aspects of deeper integration.
3. **Pro-EU with Limited Integration**: The party supports EU membership and the benefits it brings, such as the single market and free movement. However, it is cautious about further political integration and prefers to maintain significant national sovereignty. This position supports cooperation on shared issues but is wary of ceding more power to EU institutions.
4. **Balanced Approach to the EU**: The party has a balanced view on the EU, recognizing its benefits but also its drawbacks. It supports continued membership and cooperation while advocating for reforms to address specific concerns. This position seeks to strike a balance between participation in the EU and retaining national autonomy.
5. **Skeptical but Cooperative with the EU**: The party is skeptical of the EU and its current direction but supports remaining a member. It advocates for significant reforms and increased national control over certain policies. This position favors a more intergovernmental approach, with less power for EU institutions and more for member states.
6. **EU-Skeptical**: The party is highly critical of the EU and its impact on national sovereignty and economic policies. It supports a significant reduction in the powers of EU institutions and advocates for repatriating certain competences to the national level. This position might include support for reducing the country's involvement in certain EU programs and seeking opt-outs from specific EU policies.
7. **Strong Anti-EU/Stance for Withdrawal**: The party advocates for leaving the EU entirely or drastically reducing its influence. This position supports policies like exiting the single market and customs union, rejecting EU regulations, and regaining full national sovereignty. The party is likely to campaign for a referendum on EU membership and emphasize national control over laws and borders.

   If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

Environment

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party's environmental policy, using the following seven point scale:

1. **Very Strong Environmental Policies**: The party advocates for aggressive and comprehensive environmental policies. This includes ambitious targets for reducing greenhouse gas emissions, transitioning rapidly to renewable energy, implementing strict regulations on pollution, and investing heavily in sustainable practices. The party supports international environmental agreements and prioritizes environmental concerns above economic growth.
2. **Strong Environmental Policies**: The party supports robust environmental policies, with a focus on significant reductions in emissions and substantial investments in renewable energy and sustainability. While not as extreme as position 1, it still emphasizes environmental protection and includes strong regulatory frameworks and incentives for green practices.
3. **Moderately Strong Environmental Policies**: The party advocates for considerable environmental measures with a balance between environmental protection and economic considerations. It supports meaningful emission reductions, promotes renewable energy, and enforces environmental regulations, but allows for more flexibility and gradual implementation compared to stronger positions.
4. **Balanced Environmental Policies**: The party supports a balanced approach to environmental policy, integrating environmental concerns with economic growth and development. It advocates for moderate emission reductions, encourages renewable energy, and enforces essential environmental regulations, seeking to find a middle ground between sustainability and economic interests.
5. **Moderately Weak Environmental Policies**: The party favors more limited environmental policies with an emphasis on economic growth. It supports some environmental measures but prioritizes economic considerations and seeks to minimize regulatory burdens. Policies may include modest emission reductions, limited promotion of renewable energy, and relaxed environmental regulations.
6. **Weak Environmental Policies**: The party advocates for minimal environmental measures, prioritizing economic growth and development over environmental protection. It supports low levels of emission reductions, minimal investment in renewable energy, and very limited environmental regulations, emphasizing the need for economic flexibility and reduced regulatory impact on businesses.
7. **Very Weak Environmental Policies**: The party strongly supports minimal to no environmental measures, focusing entirely on economic growth and development. It opposes significant environmental regulations, does not prioritize emission reductions, and invests little to nothing in renewable energy or sustainability, arguing that environmental concerns should not hinder economic progress.

If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

Decentralization

You are conducting research on the policy positions that European parties in parliamentary democracies take in their political manifestos. Political manifestos are documents parties produce to explain their policy positions to voters in an election. For the following text of a party manifesto, please classify the party position on decentralization of political power to the regions, using the following seven point scale:

1. **Strong Pro-Decentralization**: The party advocates for extensive decentralization of political power to regional and local governments. This includes significant devolution of legislative, fiscal, and administrative powers, allowing regions to have substantial autonomy in decision-making. The party may support federalism or even a confederal structure, where regions have powers comparable to those of a national government.
2. **Pro-Decentralization with Moderate Autonomy**: The party supports substantial decentralization, giving regions considerable control over various policy areas. This includes granting regional governments significant fiscal autonomy and legislative powers, but with some key national policies still controlled centrally. The party emphasizes the importance of regional self-governance while maintaining a unified national framework.
3. **Moderate Decentralization**: The party favors a balanced approach to decentralization, supporting the transfer of some powers to regional and local governments while retaining significant national oversight. This includes limited legislative and fiscal autonomy for regions in specific areas such as education, health, and transportation, but with key national policies remaining under central control.
4. **Balanced Approach to Decentralization**: The party supports a pragmatic approach to decentralization, advocating for the devolution of certain powers to regional governments where it makes practical sense. This position seeks to balance regional autonomy with the need for national coherence, promoting cooperation between regional and central governments in shared policy areas.
5. **Moderate Centralization**: The party favors maintaining a strong central government while allowing limited decentralization. It supports giving regional governments some administrative and fiscal powers but retains most legislative and policy-making authority at the national level. This position emphasizes the importance of national unity and consistency in key policy areas.
6. **Pro-Centralization with Limited Regional Powers**: The party advocates for a highly centralized government structure, with minimal devolution of powers to regional governments. It supports keeping most legislative, fiscal, and administrative powers at the national level, allowing regional governments only minor administrative responsibilities. The party emphasizes national control and uniformity in policy implementation.
7. **Strong Centralization**: The party strongly supports a centralized government, with almost all political power retained at the national level. It opposes significant decentralization and advocates for a uniform policy approach across all regions. This position emphasizes the importance of a cohesive national strategy and minimizing regional disparities in governance.

   If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable).

APPENDIX B: PROTOTYPING

Manipulations of scoring prompts

Concerned with reports that small changes in prompt wording may have big effects on LLM outputs we developed, following Battle and Gollapudi (2024), a matrix of nine different wordings for each LLM scoring prompt. These systematically made small changes in prompt wording that should not on the face of things have any effect on results. Table B1 shows how we did this, summarizing our systematic manipulation of scoring prompts. The first column simulates three different "personas" for the LLM. The second column simulates three different types of "encouragement". The third column specifies six different seven-point scales for issue positions.

<table>
<tr><td>Pair each of these ...</td><td>with each of these ...</td><td>followed by ...</td></tr>
<tr><th>Persona</th><th>Encouragement</th><th>Issue prompt</th></tr>
<tr><td>You are an expert social scientist with a PhD in political science.</td><td>This will be fun!</td><td rowspan="3">One of the six dimension-specific prompts listed in Appendix A

Each issue prompt ends with the instruction "If the text of the manifesto does not provide a clear position on this issue, return the result of NA (meaning non-applicable)."</td></tr>
<tr><td>You are highly intelligent.</td><td>Think carefully about your answer.</td></tr>
<tr><td>You are a professor of economics.</td><td>I really need your help.</td></tr>
</table>

Table B1: Systematic variation of prompt wording for a single dimension

The prototyping exercise fed each of the three different LLMs with nine different prompts for each of six dimensions per document, for each of 16 documents. This generated 864 prompts per LLM and 2592 prompts in total for the three LLMs under investigation. We used three variables to code results in terms of the nine prompt manipulations set out in the first two columns of Table B1: persona; encouragement; persona + encouragement. These in turn can be used to generate three binary output variables to assess effects of prompt manipulation.

The first is *prompt anomaly*: the result using this prompt differs from those using the majority of the nine prompts for this score. Nearly all such anomalies were a single point difference in scores. Estimates with prompt anomalies can be classified into two types:

- *split score:* a "split verdict" (6-3 or 5-4) among the nine prompts.
- *odd prompt:* only one or two prompts differed (1-8 or 2-7) from all the others.

There were 157 prompt anomalies out of 2592 scores generated in our prototyping run. Of these 82 (3 percent of the total) were odd prompts. The rest arose from split scores, which might be taken to represent tiny differences of opinion when the "true" score lies between two integers. Table B2 breaks these down by LLM. Claude had fewer odd prompts but more split scores than the other models, with essentially nothing to choose between Gemini and GPT.

Model	**Prompt anomaly**	**Odd Prompt**	**Split score**
Claude 3.5	64	19	45
Gemini 1.5 Pro	44	30	14
GPT-4o	49	33	16
Total	157	82	75

Table B2: Prompt anomalies, by model

Table B3 shows little difference between dimensions in the frequency of odd prompts, though immigration policy does generate fewer of these.

	Odd Prompt		
Issue	**Yes**	**No**	**Total**
EU	13	419	432
Economic	14	418	432
Social	18	414	432
Immigration	6	426	432
Environment	13	419	432
Decentralization	18	414	432
Total	82	2510	2592

Table B3: Odd prompts, by dimension

Turning to features of the prompt manipulations, Table B4 shows that, barring two prompt wordings (1-3 and 2-1) that generated substantially more odd prompts, there is little difference between the others.[1] Results reported in Tables B2- B4, suggest that odd prompt responses are infrequent and depend relatively little on prompt wording. When they do arise, they almost always amount to a single point difference on the seven-point scale. We therefore concluded that detailed changes in prompt wording of the type we investigate here are not critical in this application. In all subsequent runs, we therefore use the prompt wording most closely corresponding to the prompts given by CHES to human experts: "You are an expert social scientist with a PhD in political science. Think carefully about your answer." (Prompt 1-2 in Table B4.)

	Odd Prompt		
Issue	**Yes**	**No**	**Total**
1-1	9	279	288
1-2	8	280	288
1-3	19	269	288
2-1	14	274	288
2-2	6	282	288
2-3	5	283	288
3-1	4	284	288
3-2	8	280	288
3-3	9	279	288
Total	82	2,510	2,592

Table B4: Odd prompts, by prompt features. Codes run 1, 2, 3 in the order of Table B1.

Wording of summarizer prompts

Issue order. Our initial summarizer prompt generated a single "six-issue" 500-1000 word summary of manifesto content concerning each of the six issues of interest, using short issue descriptions. On inspecting the resulting summaries, one possible reason emerged for poor

[1] The "underperforming" prompts were, puzzlingly: "You are an expert social scientist with a PhD in political science. I really need your help." And, less surprisingly: "You are highly intelligent. This will be fun!"

correlations between LLM and expert scores on decentralization. The issue was originally listed last in the order of issues to be summarized, resulting in shorter and sometimes truncated summaries of manifesto positions on this issue. We changed the order of issues in six-issue summaries and found that results depended somewhat on the order of issues, which is clearly undesirable.

Prompt detail. We prototyped the use of more detailed summarizer prompts which much more closely approached the scoring prompts for each dimension set out in Appendix A1. These tended to generate lower correlations with expert scores, perhaps because the more detailed issue descriptions were leading the LLMs to ignore potentially relevant text not explicitly covered by these.

Single-issue summaries. All of this implied using six single-issue summaries rather than a single six-issue summary and reverting to a polished version of the short summarizer issue prompts. The final summarizer configuration, set out in Appendix A4, therefore used short prompts to generate 300-400 word single-issue summaries for each of the six issues.

Bypassing manifesto summaries: scoring native language manifestos in a single step

We assessed effects of bypassing the intermediate summarization of manifesto content on issue areas of interest and moving straight from the original language full-text manifesto to a score on a seven-point scale for each issue. We encountered a series of practical difficulties in doing this, mostly arising from LLM behaviors with the largest full-text manifestos. The LLMs claim in their publicity to be able to "ingest" very large documents (Claude's term). It is less clear, however, whether they can fully "digest" these without outside help. One notable problem was a tendency to "forget" the original scoring instructions when these were followed by a very long manifesto text, which never seemed to happen when using short summaries. We addressed this by repeating the same scoring instructions at the end of the prompt.

Even with this workaround, the LLMs' performance was substantially worse when the summarization stage was bypassed. This manifested in the much higher proportion of NA responses – on issues where including the summarization stage had resulted in high correlations with expert scores. This suggests that there is a retrieval problem with large native language manifestos when these are analyzed in a single stage. Table B5 reports rates of NA responses for the three LLMs: with a summarizer stage (left panel); and bypassing the summarizer (right panel). *Bypassing the summarizer substantially increases the proportion of NAs for all LLMs.*

	With summaries		Bypassing summary	
	Total	Prop. NA	Total	Prop. NA
GPT	18/96	0.19	26/84	0.31
Claude	5/96	0.05	44/96	0.46
Gemini	23/96	0.24	48/96	0.50

Table B5: NAs, all dimensions, preferred prompt, with and without summarizer

Not only were there many more NA responses when the summarization stage was bypassed but, for those scores which were not NA, *correlations with CHES were systematically lower.* Table B6 shows these correlations.

Correlation with CHES	With summaries	No summary
GPT	0.79	0.52
Claude	0.67	0.59
Gemini	0.76	0.75

Table B6: Correlations with expert scores, all dimensions, (incl. decentralization)

Translation

The default summarizer does two jobs at once: it summarizes the original language text and returns the summary in English. Given the success of adding the summarization module, we experimented with further modularization which separated the functions of translation and summary. The resulting workflow is then:

a. **Translate**: the full native language manifesto into English. We know that LLMs are now exceptionally good at this.
b. **Generate a text summary** of manifesto content relevant to specified policy dimension(s). We also know that summarization of large documents is a tried, tested and successful use case for LLMs.
c. **Summarize** manifesto content relevant to specified policy dimension(s) using an **integer on a seven-point scale**, or a missing data code.

We therefore added the Google translate API to the workflow and inserted a stage before the summarizer which translated the original language manifestos into English. We reran the full nine-prompt, three-model, six-dimension run for the 15 Non-English language manifestos used in prototyping.

Comparing pairwise correlations between LLM scores, with and without translation, and corresponding expert survey scores, the systematic pattern is that *correlations with expert survey scores tend to be a little lower when the translation stage is inserted before the summarizer.*

Dimension	Translation /not	EU	Environ ment	Immigr-ation	Decent-ralization	Social lifestyle	Taxes vs spending
Claude 3.5	No trans.	**0.89**	**0.93**	0.73	0.26	**0.93**	**0.91**
	Translation	0.85	0.76	0.72	0.27	0.89	0.88
Gemini 1.5	No trans.	0.76	**0.71**	**0.85**	**0.69**	0.59	**0.92**
	Translation	**0.84**	0.60	0.77	0.18	0.96	0.86
GPT 4o	No trans.	0.80	**0.84**	**0.84**	**0.38**	**0.93**	**0.91**
	Translation	**0.89**	0.50	0.73	0.12	0.89	0.85

Table B7: Correlations between LLM and expert scores, with and without a translation stage.

APPENDIX C: CORRELATIONS BETWEEN LLM AND EXPERT SCORES

Details on correlations between LLM and expert scores

Summ-arizer	Scaler	Run	Decent-ralization	Environ-ment	EU	Immig-ration	Social	Taxspend	All Issues
Claude	Claude	Zero-shot	0.50	0.66	0.87	0.88	0.87	0.82	0.72
Claude	Claude	Few-shot	0.50	0.64	0.88	0.89	0.89	0.85	0.75
Claude	GPT-4o	Zero-shot	0.49	0.68	0.90	0.88	0.91	0.83	0.75
Claude	GPT-4o	Few-shot	0.48	0.67	0.90	0.88	0.89	0.82	0.73
Claude	Gemini	Zero-shot	0.47	0.65	0.88	0.89	0.89	0.85	0.73
Claude	Gemini	Few-shot	0.41	0.65	0.89	0.88	0.90	0.86	0.73
GPT-4o	Claude	Zero-shot	0.25	0.69	0.90	0.89	0.89	0.79	0.67
GPT-4o	Claude	Few-shot	0.28	0.70	0.90	0.90	0.90	0.85	0.69
GPT-4o	GPT-4o	Zero-shot	0.25	0.65	0.90	0.88	0.90	0.80	0.69
GPT-4o	GPT-4o	Few-shot	0.21	0.67	0.89	0.88	0.90	0.81	0.67
GPT-4o	Gemini	Zero-shot	0.25	0.63	0.87	0.89	0.90	0.82	0.68
GPT-4o	Gemini	Few-shot	0.29	0.65	0.89	0.89	0.90	0.84	0.68
Gemini	Claude	Zero-shot	0.37	0.73	0.86	0.83	0.88	0.82	0.71
Gemini	Claude	Few-shot	0.34	0.69	0.87	0.84	0.92	0.84	0.73
Gemini	GPT-4o	Zero-shot	0.39	0.74	0.85	0.85	0.89	0.83	0.73
Gemini	GPT-4o	Few-shot	0.32	0.71	0.85	0.86	0.89	0.84	0.72
Gemini	Gemini	Zero-shot	0.33	0.70	0.79	0.86	0.91	0.88	0.72
Gemini	Gemini	Few-shot	0.35	0.66	0.84	0.85	0.92	0.87	0.72
Overall	**Overall**	**Zero-shot**	**0.34**	**0.67**	**0.86**	**0.87**	**0.89**	**0.82**	**0.70**
Overall	**Overall**	**Few-shot**	**0.31**	**0.65**	**0.88**	**0.87**	**0.90**	**0.83**	**0.71**

Table C1: Correlations between LLM and expert scores by summarizer and scaler

Calculation	Run	Decent-ralization	Environ-ment	EU	Immig-ration	Social	Taxes v. Spending	All issues
Summary	Zero-shot	0.34	0.67	0.86	0.87	0.89	0.82	0.71
LLM Ensemble 9	Zero-shot	0.42	0.76	0.91	0.89	0.90	0.87	
Summary	Few-shot	0.31	0.65	0.88	0.87	0.90	0.83	0.71
LLM Ensemble 9	Few-shot	0.34	0.75	0.91	0.90	0.90	0.87	
Summary	Zero + Few	0.32	0.66	0.87	0.87	0.89	0.82	0.71
LLM Ensemble 18	Zero + Few	0.38	0.75	0.91	0.89	0.90	0.87	

Table C2: Correlations between "ensemble" zero-shot LLM and expert scores.

Correlations between different zero-shot LLM Scores

The first three variables show the mean score for each scaler model across all summaries. The second three variables show the mean score for each summarizer model across all scalers. All correlations extremely high, though somewhat less so for decentralization.

Economic

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.9435	1.0000				
mean_Gemin~e	0.9484	0.9464	1.0000			
mean_Claud~y	0.9288	0.9405	0.9314	1.0000		
mean_Gemin~y	0.9581	0.9602	0.9467	0.8983	1.0000	
mean_GPTsu~y	0.9517	0.9418	0.9471	0.8736	0.9080	1.0000

Social

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.9657	1.0000				
mean_Gemin~e	0.9756	0.9713	1.0000			
mean_Claud~y	0.9465	0.9619	0.9448	1.0000		
mean_Gemin~y	0.9385	0.9548	0.9488	0.8978	1.0000	
mean_GPTsu~y	0.9719	0.9755	0.9740	0.9187	0.9213	1.0000

Immigration

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.9658	1.0000				
mean_Gemin~e	0.9684	0.9709	1.0000			
mean_Claud~y	0.9708	0.9676	0.9700	1.0000		
mean_Gemin~y	0.9599	0.9655	0.9601	0.9155	1.0000	
mean_GPTsu~y	0.9723	0.9741	0.9747	0.9492	0.9365	1.0000

EU

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.9711	1.0000				
mean_Gemin~e	0.9715	0.9571	1.0000			
mean_Claud~y	0.9520	0.9608	0.9392	1.0000		
mean_Gemin~y	0.9531	0.9471	0.9509	0.8826	1.0000	
mean_GPTsu~y	0.9672	0.9602	0.9650	0.9216	0.9018	1.0000

Environment

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.9227	1.0000				
mean_Gemin~e	0.8878	0.9365	1.0000			
mean_Claud~y	0.8951	0.9175	0.8821	1.0000		
mean_Gemin~y	0.8829	0.9034	0.8964	0.7711	1.0000	
mean_GPTsu~y	0.9200	0.9060	0.9002	0.8068	0.7944	1.0000

Decentralization

	m~Tscore	mean_C~e	m~iscore	mean_C~y	mean_G..	mean_G..
mean_GPTsc~e	1.0000					
mean_Claud~e	0.8585	1.0000				
mean_Gemin~e	0.8662	0.8598	1.0000			
mean_Claud~y	0.7833	0.8396	0.8746	1.0000		
mean_Gemin~y	0.8337	0.8406	0.8289	0.6914	1.0000	
mean_GPTsu~y	0.7478	0.7556	0.7216	0.5998	0.4950	1.0000

Table C3: Inter-correlations between LLM scores, by dimension

APPENDIX D: CORRELATIONS OF LLM-BASED ESTIMATES WITH MP ESTIMATES

Applying the MP data as an additional convergent validity check, we note that these correlated the least with both LLM-based and expert survey-based estimates. These measures use the logit scales described in Appendix A2.

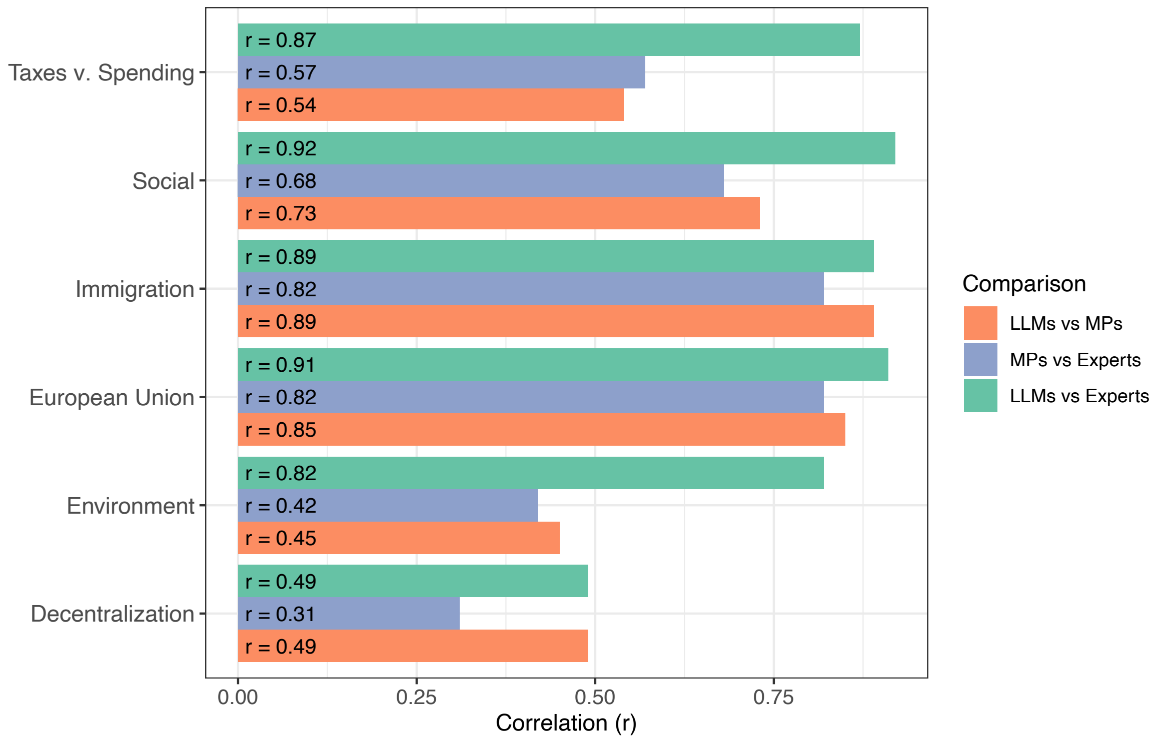

Figure D1: Correlations between LLM-, MP-, and expert survey-based estimates.

APPENDIX E: REPLICATION RESULTS

We ran the first set of results in August and September of 2024, and the second set (reported in the main results) in December 2024. The versions we used for the first run were claude-3-5-sonnet-20240620, gemini-1.5-pro-002, and gpt-4o-2024-08-06, and for the second, claude-3-5-sonnet-20241022, gemini-1.5-pro-002, and gpt-4o-2024-11-20.

Table E1 reports the correlations between the LLM-based estimates of policy between the first and second runs.

Issue dimension	Correlation
Economic	0.97
Social	0.97
Immigration	0.96
European Union	0.98
Environment	0.92
Decentralization	0.69

Table E1: Correlations of LLM-based estimates across an initial and a replication run.

APPENDIX F: COALITION AGREEMENT OVERLAP WITH LLM MANIFESTO CORPUS

The list of coalition agreements collected by Klüver et al. (2023) that overlap with the set of member party manifestos in the LLM corpus is as follows. Missing manifestos for coalition members are highlighted in bold.

Cabinet	Member Parties
AT_2007_Gusenbauer_I	SPÖ, ÖVP
BE_2003_Verhofstadt_III	VLD, SP, PS, PRL
BE_2011_Di_Rupo_I	PS, CVP, PRL, SP, LP/PL, PSC
BE_2014_Michel_I	VU, CVP, VLD, PRL
BG_2014_Borissow_II	GERB, RB
CZ_2007_Topolanek_II	ODS, KDU-CSL, SZ
CZ_2010_Necas_I	ODS, TOP09, VV
FI_2003_Jäätteenmäki_I	ZE, SDP, SW
HU_2006_Gyurcsany_II	MSZP, SZDSZ
HU_2010_Orban_II	Fidesz-MPS, KDNP
IE_1989_Haughey_IV	FF, PD
IS_2003_Oddsson_IV	IP, PP
NL_1989_Lubbers_III	CDA, PvdA
NL_2003_Balkenende_II	CDA, VVD, D66
NL_2007_Balkenende_IV	CDA, PvdA, CU
NL_2010_Rutte_I	VVD, CDA
NO_1989_Syse_I	H, KRF, SP
SE_2006_Reinfeldt_I	KO, ZE,CD,LI
SE_2010_Reinfeldt_II	KO, LI, ZE,CD
SE_2014_Löfven_I	SD, G
SI_2014_Cerar_I	SMC, DeSUS, SD
SK_2006_Fico_I	S, SNS, LS-HZDS
SK_2010_Radicova_I	SDKU, SaS, KDH, MH

Table F1: Overlap between Kluwer coalition and LLM manifesto corpora

APPENDIX G: COST AND TIME COMPARISONS TO GENERATE POLICY POSITIONS

Method/Run	Financial Cost	Time Required
3 x 3 x 2 proprietary LLMs (Summary: GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro) x (Scaling: GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro) x (Zero-shot, Few shot)	Approximately **$1,100**	3 days
3 x 2 open-weight LLMs (Summary x Scaling: LlaMA-3.3-70B) + (Summary x Scaling: Deepseek-V3-671B) + (Summary x Scaling: Gemma-3-27B-it)	Approximately **$300** (Nebius.com hosting costs)	1 day
Expert surveys Mean number of respondents was 11.6 per CHES expert survey and 25.2 for Benoit and Laver)	Assuming (conservatively) that the cost of expert time was $150/hr, and each expert survey took 45 mins to complete and 15 mins to check and process: 178 CHES x 11.6 x $150 = $30,9720 + 57 BL/LH x 25.2 x $150 = $215,460 = **$525,180** (15 mins processing time is very conservative and excludes the time spent finding experts and sending reminders.)	If each expert survey takes one month, then repeating the expert surveys applied here would take **235 months** total time In practice these spanned **30 years**, as expert surveys must be conducted contemporaneously with the period being rated.
Manifesto coding Each of the 235 party manifestos required one expert coder, plus preparation time to recruit and train expert coders and to check and publish the data.	Each manifesto coder is paid approximately $170 (€150) to code the manifesto, or $170 * 235 = $39,950 Each election can involve recruitment and training of up to 8 hrs per new coder, approx. $200/hr * 8 * 35 elections = $56,000 **$95,950**	Each manifesto takes approximately 20 hours to code and another 5 for auxiliary processes, total time = 235 * (25 hours) / 8 hrs/day / 365 days/yr = **2 person-years**

Table G1: Cost and Time Comparison of LLM-based Methods versus Expert Survey and Manifesto Coding Methods.

Notes: *MP estimates based on correspondence with Annika Werner, a member of the Manifesto Project Steering Committee and former member of the MP Core Group.

APPENDIX H: QUERYING LLMS ABOUT DATA LEAKAGE VIA CHAT

Claude 3.5 Sonnet

I want to ask you a question about a political data project on party policy positions, and I do not want you to access the web for generating your answers.
Are you familiar with the Chapel Hill Expert Survey project? And are the specific party policy positions in your training data, in a way that you might draw on them if I were asking you to answer questions similar to the CHES policy scales, but were instructed to base your answer only on a party manifesto that I supplied to you?

Claude: Yes, I am familiar with the Chapel Hill Expert Survey (CHES) project as a key expert survey measuring party positioning in European countries. The project began in 1999 and collects expert assessments of national political party positions on European integration, ideology, and various policy dimensions like economic policy, social policy, and environmental policy using standardised numerical scales.

Regarding your second question: While I understand the CHES methodology and scaling approaches, I cannot say with certainty whether I have reliable training data on specific party position scores from CHES surveys. However, I am capable of:

1. Understanding and applying CHES-style policy scales (e.g. the 0-10 scaling on economic left-right)
2. Analysing manifestos using similar dimensional frameworks
3. Extracting policy positions from text using comparable methodological approaches

If you were to provide a party manifesto, I could analyse it using frameworks similar to CHES dimensions, but I would be transparent about:

- Basing assessments solely on the manifesto content you provide
- Using CHES-style scales as an analytical framework rather than claiming to replicate CHES expert scores

If I asked you to base your answers to the policy scale questions only on a summary of the text that I supplied, and this summary lacked any identifying information about the party, would that eliminate the risk of data leakage?

Claude: Yes, that approach would effectively eliminate the risk of data leakage. By providing an anonymised summary of manifesto content, you would ensure that my analysis would be based solely on the textual evidence you supply, rather than potentially being influenced by any pre-existing training data about specific parties.

This methodology would be analytically sound because:

1. It forces a purely text-based assessment using only the supplied content
2. It prevents any potential anchoring effects from party identity or reputation
3. It aligns with blind coding practices used in content analysis
4. It enables focus on specific policy signals in the text rather than broader contextual knowledge

The key would be ensuring the summary retains sufficient policy-relevant content to make meaningful assessments on the dimensions you're interested in measuring.

Gemini 2.5 Pro

I want to ask you a question about a political data project on party policy positions, and I do not want you to access the web for generating your answers.
Are you familiar with the Chapel Hill Expert Survey project? And are the specific party policy positions in your training data, in a way that you might draw on them if I were asking you to answer questions similar to the CHES policy scales, but were instructed to base your answer only on a party manifesto that I supplied to you?

Gemini: Yes, I am familiar with the Chapel Hill Expert Survey (CHES) project. It is a well-known and significant academic project that collects data on the policy positions, ideological stances, and other characteristics of political parties across various European countries (and increasingly beyond). The surveys are typically completed by country-specific experts who provide quantitative assessments of parties on a range of predefined scales.

Regarding your second question:

My training data includes a vast amount of text, which encompasses a wide array of academic literature, including political science research. Therefore, it is highly probable that information *about* the Chapel Hill Expert Survey, its methodology, the policy dimensions it measures (e.g., economic left-right, social traditionalism-liberalism, views on European integration), and published analyses using CHES data are part of my training.

However, it's important to distinguish between being trained *on information about CHES* and having the specific, granular CHES dataset of party policy scores directly encoded in a way that I could "look them up" and reproduce them. I don't "store" or "access" datasets in that manner.

If you were to provide me with a party manifesto and instruct me to base my answers *only* on that manifesto for questions similar to CHES policy scales, here's how my training would likely interact with that task:

1. **Adherence to Instructions:** My primary directive would be to use *only* the supplied manifesto as the source for my answers. I would focus on analyzing the text of that document.
2. **Conceptual Understanding:** My training has equipped me with a broad understanding of political concepts, policy areas, and ideological dimensions that are also central to frameworks like CHES. For example, I can understand what "economic policy," "social policy," "environmental policy," or "foreign policy" refers to. I can also understand concepts like "left-wing," "right-wing," "liberal," "conservative," "pro-EU," or "eurosceptic" as they are generally used and discussed in political discourse and academic literature.
3. **Analytical Capability:** I would use my language processing and analytical capabilities to identify statements, proposals, and priorities within the manifesto that pertain to the policy dimensions you specify (which might be similar to CHES scales). I would then synthesize this information from the manifesto to characterize the party's stance on those dimensions.

4. **No Direct Recall of CHES Scores:** I would not be "drawing on" or "recalling" specific CHES scores for that party from my training data to answer your questions about the manifesto. My analysis would be grounded in the textual evidence of the manifesto itself. The knowledge about CHES and similar frameworks helps me understand the *type* of questions you're asking and the *nature* of the policy dimensions, but the content of my answer about the manifesto's positions would be derived from the manifesto.

In essence, while my general knowledge base (which includes information about CHES) provides the conceptual tools to understand and interpret your request and the policy dimensions involved, my answers regarding a specific party's policy positions would be based on the information extracted directly from the manifesto you supply, as per your instructions. I would be performing a text-based analysis rather than trying to replicate or draw upon pre-existing CHES expert placements.

If I asked you to base your answers to the policy scale questions only on a summary of the text that I supplied, and this summary lacked any identifying information about the party, would that eliminate the risk of data leakage?

Gemini: Yes, providing a summary of the text that lacks any identifying information about the party would **substantially reduce, and in many practical senses, aim to eliminate** the risk of what you're calling "data leakage" – that is, me drawing upon specific knowledge about a known political party from my training data, rather than solely relying on the information you've provided.

Here's why this approach is effective and what it addresses:

1. **No Direct Identifiers:** If the summary contains no party names, country specifics, names of leaders, or other uniquely identifying information, I would have no direct cues to link the provided policy positions to any specific party entity that exists in my training data. My analysis would, by necessity, be confined to the text of the summary itself.
2. **Focus on Policy Content:** The task then becomes purely an analysis of the stated policy positions within the summary. I would evaluate these positions against the policy scales you define, based on the semantic content of the summary.
3. **Breaks the Link to Specific Examples:** My training data contains information about many parties and their documented positions. Without identifiers, the summary becomes a more abstract set of policy statements, making it very difficult, if not impossible, for me to consciously or unconsciously associate it with a specific, pre-existing party profile in my knowledge base.
4. **Mitigation of Unconscious Bias:** While my general understanding of political concepts and how certain policies are typically categorized (e.g., "left-wing," "environmentalist") is derived from my broad training, the absence of identifying information prevents this general knowledge from being inappropriately channeled towards a specific, known party. The risk of me saying "This sounds like Party X, and therefore its position on Y is likely Z (even if Z isn't in the summary)" is greatly diminished.

Regarding "eliminating" the risk entirely:

"Eliminate" is a very strong term. While this method is highly effective and is standard practice in many research designs to prevent bias:

- If the *combination* of policy positions described in the anonymized summary is extremely unique and happens to be famously associated with only one widely documented party (even if that party isn't named), it's theoretically possible that my general pattern recognition capabilities might find a strong resonance with archetypes I've learned. However, if I am strictly instructed to use *only the summary*, I would still base my answer on the explicit text. The influence here would be more about the *interpretation* of those uniquely combined policies rather than "leaking" factual data about a specific party that wasn't mentioned. The key is that I wouldn't be able to attribute it to a specific party.
- The effectiveness hinges on how thoroughly the summary is anonymized and how generic or specific the policy descriptions are.

In conclusion:

Supplying an anonymized summary, with the instruction to base answers *only* on that summary, is a very strong method to ensure that my analysis focuses on the provided text and does not draw on specific party information from my training data. It would **drastically reduce the risk of data leakage** concerning specific party identities and their known policy scores to a minimal, likely negligible, level for the purpose of policy scale assessment. It forces the evaluation to be based on the intrinsic content of the summary you provide.

ChatGPT 4o

I want to ask you a question about a political data project on party policy positions, and I do not want you to access the web for generating your answers. Are you familiar with the Chapel Hill Expert Survey project? And are the specific party policy positions in your training data, in a way that you might draw on them if I were asking you to answer questions similar to the CHES policy scales, but were instructed to base your answer only on a party manifesto that I supplied to you?

ChatGPT said: Yes, I am familiar with the Chapel Hill Expert Survey (CHES) project. CHES provides expert-coded estimates of party positions across a range of policy dimensions in European countries, typically covering dimensions like left–right ideology, EU integration, economic and social policy, among others. The data is based on expert surveys rather than party documents directly.

Regarding whether specific CHES party policy positions are in my training data: while I do not have direct access to the CHES dataset or its proprietary content, it is possible that summaries or discussions of CHES results (e.g. from published papers or publicly accessible academic articles that cite CHES data) were part of the training data. Therefore, there's a **risk of "data leakage"** if I appear to answer CHES-style questions about a party using prior knowledge rather than only the manifesto text you provide.

Even if you explicitly instruct me to base responses only on a manifesto, I cannot guarantee that I would fully suppress prior knowledge of party positions learned during training. This is particularly relevant for high-salience, well-known parties (e.g. CDU in Germany, or the UK Conservatives or Greens), whose positions may be implicitly encoded in the model.

To avoid this, a better-controlled approach would be:

- Using an LLM fine-tuned only on manifesto texts without external political knowledge.
- Masking party names and references in the manifesto to reduce anchoring effects.
- Comparing outputs from several models (especially open-weight models) to assess consistency.

If I asked you to base your answers to the policy scale questions only on a summary of the text that I supplied, and this summary lacked any identifying information about the party, would that eliminate the risk of data leakage?

ChatGPT said: It would substantially *reduce* the risk of data leakage, but not *completely eliminate* it.

Here's why:

✅ **Why the risk would be greatly reduced:**

1. **No party name or country reference**: If the summary lacks explicit identifiers (e.g. party name, country, election year), it removes key cues that might trigger retrieval of prior associations.
2. **Text-based reasoning**: By focusing the prompt on "based only on the following policy summary," you're steering the model toward grounded inference rather than recall.
3. **Anonymous content**: Without metadata or well-known slogans or priorities, it's harder for the model to match to memorised entities.

⚠️ **Why a *small risk remains*:**

1. **Stylised or unique phrasing**: Some parties have distinctive policy language or issue combinations (e.g. Euroscepticism + green economics), which might still cue memorised associations.
2. **Latent embeddings**: During training, the model might have encoded generalised associations between policy patterns and particular types of parties, which could surface even without explicit names.
3. **Involuntary pattern completion**: LLMs are optimised to complete patterns. If the summary closely resembles real-world material seen during training, even without identifiers, the model might still "fill in" based on prior associations.

Implication:

While anonymising and instructing the model to rely *only* on the summary reduces the likelihood of leakage-induced answers, you cannot guarantee that the model isn't drawing on prior learned associations. This is especially true if you're using a general-purpose model like GPT-4.

For maximum confidence, a better approach would be to:

- Run the same summary through multiple models (especially open-weight LLMs with known training data).
- Compare LLM outputs to human-coded CHES values to identify outliers.
- Apply controlled interventions like prompt randomisation or text scrambling to detect anchoring effects.

Note: all conversations took place on May 7, 2025.

APPENDIX I: ROBUSTNESS TESTS FOR IDENTIFIED PARTIES

Robustness test design

Our examination of the party summaries indicated that in many cases, these contained a brief reference that identified the party by name or abbreviation, for instance, "Based on the provided manifesto from Austria's FPÖ party (2006), here is a summary of their positions on taxation and public spending:…" Using a procedure described below to mask the party names and abbreviations, we saved a full set of the policy summaries generated by GPT-4o, replacing the party identifiers with a mask, and then repeated the scaling stage (using GPT-4o and zero-shot learning). The motivation for this robustness test is that data leakage would be impossible if the summary contained no overt reference to the party for which the LLM might contain information about policy positions from its training data.

In our test, the scores generated from the identified and the non-identified summaries correlated at 0.99, detailed by policy issue in Table I1, and plotted in Figure I1. In sum, it makes no difference to the scoring stage whether the LLM can identify the party—compelling evidence that leakage is not present.

Issue	Correlation	N
Environment	0.97	442
Economic	0.99	284
Social	0.99	295
Immigration	0.98	402
European Union	0.95	428
Decentralization	0.97	447
Overall	0.99	2,298

Table I1: Correlations between scale positions from original and de-identified policy summaries. Note: GPT-4o, zero-shot; includes coalitions and some parties not linked to expert surveys in the original sample.

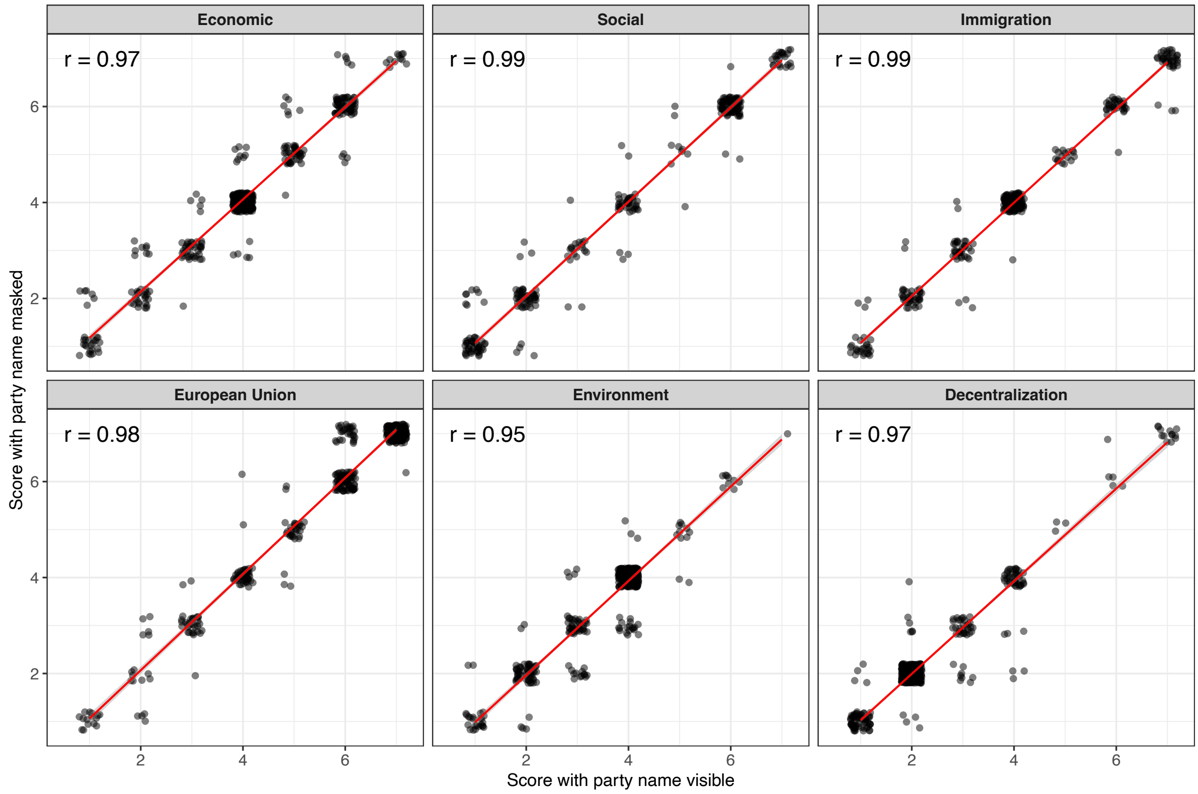

Figure I1: Correspondence between scale positions from original and de-identified policy summaries. Data comes from Table I1.

Detection and anonymization method

We used Claude to extract party names from the original summaries using a zero-shot learning approach to extract all party names or identifier (abbreviations). After extraction, an exact text matching process was performed to verify that the extracted names appeared verbatim in the original summary, thereby ensuring the accuracy and presence of the identified entities. One limitation of this approach is that, although the extracted party names can be verified to appear in the summary text, it cannot be guaranteed that all instances of party names are captured. In other words, high precision can be ensured, but high recall is not necessarily achieved. However, if the LLM fails to extract certain party names, it likely would not identify the party using those names either, suggesting that missed extractions may have limited impact.

We then replaced these party names, as well as any abbreviations with a mask, for instance, “The manifesto of the <PARTY> (<PARTY>)…”. The mask was only applied in about half of the cases (49% or 1,412 from 2,892, where each party or coalition has six summaries each, one for each policy issue).

The prompt used for party name extraction was:

Given the following summary of a political manifesto, please extract ALL mentions of:

Political Parties: Extract every political party name exactly as it appears in the text. Party references may appear in various forms including:

- Full official names (e.g., "Freedom Party", "Social Democratic Party")
- Abbreviations (e.g., "FPÖ", "SPD", "CDU")
- Colloquial names (e.g., "Tories", "Labour")
- Language-specific variations or translations
- Historical party names
- Coalition or alliance names
- Any other form that clearly identifies a political party or politically organized group

Do not modify, standardize, or infer any party names.

Return your answer as a clean JSON object without any markdown formatting, code blocks, or explanation. The response should be ONLY the raw JSON that can be directly parsed. The JSON should have this structure:

{"country": ["country1", "country2", ...], "party": ["party1", "party2", ...]}

If no countries or parties are explicitly mentioned in the text, return an empty array for that field.

Do not infer or add any countries or parties that are not explicitly mentioned in the text.

Extract the names exactly as they appear - do not modify capitalization, abbreviations, or translate them.

Here is the manifesto summary: [SUMMARY TEXT]

REFERENCES

Battle, Rick, and Teja Gollapudi. 2024. "The Unreasonable Effectiveness of Eccentric Automatic Prompts." *arXiv preprint arXiv:2402.10949*.

Benoit, Kenneth, and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Jolly, Seth, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. "Chapel Hill Expert Survey Trend File, 1999–2019." *Electoral studies* 75: 102420.

Klüver, Heike, Hanna Bäck, and Svenja Krauss. 2023. *Coalition Agreements as Control Devices: Coalition Governance in Western and Eastern Europe.* Oxford: Oxford University Press.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition*. New York: Routledge.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36: 123-55.

Merz, Nicolas, Sven Regel and Jirka Lewandowski. 2016. “The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis.” *Research and Politics* (April-June): 1–8. doi: 10.1177/2053168016643346.