Working Paper Series

SoDa LABS

Social [science] insights from alternative data

Est. 2018

Tracking Policy-relevant Narratives of Democratic Resilience at Scale: from experts and machines, to AI & the transformer revolution

Simon D Angus

SoDa Laboratories Working Paper Series

No. 2024-07

REF

Simon D Angus (2024), SoDa Laboratories Working Paper Series No. 2024-07, Monash Business School, available at http://soda-wps.s3-website-ap-southeast-2.amazonaws.com/RePEc/ajr/sodwps/2024-07.pdf

PUBLISHED ONLINE

10 December 2024

SoDa Laboratories

https://www.monash.edu/business/soda-labs/

MONASH University | MONASH BUSINESS SCHOOL

AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

ABN 12 377 614 012 CRICOS Provider Number: 00008C

(2024), 1–34
doi:10.1017/xxxx

RESEARCH ARTICLE

Tracking Policy-relevant Narratives of Democratic Resilience at Scale: from experts and machines, to AI & the transformer revolution

Simon D. Angus[1]

[1]Dept. of Economics, Monash University, Clayton, 3800, VIC, Australia. [1]SoDa Laboratories, Monash Business School, Clayton, 3800, VIC, Australia. .
Corresponding author. E-mail: simon.angus@monash.edu

Received xx xxx xxxx

Keywords: Computational linguistics; Political discourse analysis; Natural Language Processing; Quantitative Social Science; AI in policy research

Abstract

Democratic resilience is as much about the narratives of our nation we affirm, as the institutions that enshrine our values and laws, a fact re-affirmed by scholarship across many branches of social science in recent decades. For policymakers and quantitative social scientists, analysing or tracking public discourse through the lens of narrative and framing has historically involved the annotation of texts by hand, placing severe limitations on the scale and modality of discourse under inquiry. In this study, we consider a variety of tools from the field of computational linguistics, which either automate the standard approach to textual annotation, or introduce entirely new ways of conceptualising 'text as data', opening up new horizons for the tracking of public narratives of democratic resilience. In particular, we assess the regime-shift occurring in natural language processing and artificial intelligence brought about by the advent of the transformer architecture. By undertaking two distinct empirical analysis tasks with recent transformer methods we provide practical demonstrations that these new tools offer, perhaps for the first time, the 'holy grail' of the quantitative social scientist: the ability to identify, accurately, and efficiently, nuanced narratives in text at scale. We conclude by contributing data and research recommendations for public stakeholders who wish to see these opportunities realised.

Policy Significance Statement

Good policy starts with good information. In the case of policy relating to democratic resilience, whilst there is a very large set of choices available to the policymaker who desires to strengthen democracy, there is a remarkably slender set of rich information on citizen narratives of democracy on which to draw. Whilst surveys will always be important, we are now in an age where voluminous alternative data forms are arising in social media, digital media, and political discourse in real time. These new sources present a huge opportunities for policymakers to identify, track, and analyse rich belief structures – how individuals frame issues and events – in near real time. In this paper, we ask whether advances in natural-language-processing (NLP) and in particular, recent breakthroughs in generative AI technologies, can be applied to democratic narrative analysis at scale. We also provide practical demonstrations of these tools on two discourse areas related to democratic resilience: populism and immigration. We further identify the centrality of harmonised data sources to any future endeavour, and make recommendations for policymakers to foster these methods and tools.

1. Introduction

Story-telling has always been a powerful form of human expression: we venerate our literary giants; and we encode our culture, history and identity in the stories that we tell our children and each other in so many private moments. Yet for democracies, story telling, or more formally 'narratives', seem to take on a particularly potent guise. Each election cycle, those of us fortunate enough to live in an open, democratic society, will encounter political leaders who's chief form of political speech will be in narrative form: a vision for a better future; a depiction of an opponent's 'world'; a personal life 'story', connecting the leader to the lounge-room.

Policy experts, economists, sociologists and political scientists have long theorised over the role of public narratives in our national political and policymaking journey. For some, narratives have the power to both initiate and catalyse major economic events, from the Great Depression to financial crises (Shiller, 2019), whilst others articulate the common finding that public narratives are at their most powerful when they resonate with personal convictions *already held* (Polletta and Callahan, 2017). Indeed, reflecting on the potency of Trump's campaign tactics, Polletta and Callahan note that it was the 'small stories' heard through a complex web of radio, TV, social media, and personal networks which (p.13), 'meshed with the experiences of others in a way that made them all seem personal.' More forceful of all, narratives have been rediscovered by some as providing a kind of ontology of self – public narrative as 'identity compass' to the inner self. According to Somers (1994) (p.606) (emphasis added),

> It matters not whether we are social scientists or subjects of historical research, but that all of us come to **be** who we **are** (however ephemeral, multiple, and changing) by being located or locating ourselves (usually unconsciously) in social narratives **rarely of our own making**.

In other words, we now recognise that public narratives matter, not only for how public debate proceeds, but also, for the composition and sense-making of individuals themselves.

Yet if the power of public narratives to shape individuals and society was already recognised in the 90s, how much more should we pay attention to them today, given the unprecedented acceleration in the availability, volume and velocity (Mauro et al., 2016) of public discourse surging through our democracies? Parallel mega-trends including the ubiquity of the internet, social media platforms, the 24/7 news cycle, the spectre of mis-information, 'echo-chambers' and geo-political interference; each contribute to a sense that for those who carry the fire for our democracies, characterising, mapping, and tracking the maelstrom of competing narratives has become an *essential* task.

But are our methodological tools up to the task? Traditional political narrative discovery and analysis methods (e.g. focus-groups, surveys, human labelling of texts) will continue to be part of the democratic resilience analysis toolbox, yet they are ill-equipped to handle the scale or immediacy of the modern narrative tracking challenge. By contrast, modern machine-learning, natural language processing (NLP) and artificial intelligence (AI) technologies present compelling opportunities to bridge this gap.

In this paper, we ask,

> How can advances in NLP and AI technologies, particularly large language models, enhance our ability to track and analyse narratives related to democratic resilience at scale?

To tackle this question, we first present a framework (§2) that connects the democratic resilience analysis challenge with traditional narrative discovery and analysis techniques, and then introduces automated methods as automated companions to these techniques. In so doing, the framework introduces the contention of this paper that modern, transformer-based (AI) technologies have arrived at a level of capability such that human-level narrative discrimination is now available at scale, ready for application to tracking policy-relevant narratives of democratic resilience.

We use this framework to structure the rest of the paper. Section §2.1 positions narrative tracking within the broader question of democratic resilience, drawing on theories of communication to

introduce the 'framing' task as the common object of analysis, and in §2.2 we provide examples of pernicious narratives that can be pursued by this analysis. Section §3 then considers the evolution of narrative analysis via framing, from human-labelling to automation and the recent revolution brought about by transformer architecture that is driving the current wave of AI transformation. We then provide, in section §4 two distinct applied demonstrations of these technologies to fix ideas and ground our analysis of these tools. In so doing, we shall encounter the enormous potential inherent in the latest wave of language modelling technology, opening up new research frontiers to harness this potential in ways that explicitly seek to strengthen democracy. Given that we see few, if any examples of this research to date, we conclude by proposing (§5) a new computational research agenda that takes advantage of these tools for strengthening democracy.

Together, this study makes a number of contributions to the data—policy nexus in relation to democratic resilience: a *conceptual framework* to orientate and rationalise frontier transformer-based computational methods as revolutionary technologies in the democratic resilience toolset; a *multi-disciplinary synthesis* of theories of resilience, framing, and implementation to situate these technologies in a non-trivial intellectual context; a set of *demonstration exercises* to exemplify, in an accessible way, the power of these new technologies relative to former approaches; and a *computational research agenda* with recommendations for data controllers and research funding bodies alike.

2. A framework for analysing narrative aspects of democratic resilience with machines

Conceptually, we propose a three tier framework, as illustrated in Fig. 1 and expanded in following sections, to situate the tracking of narrative aspects of democratic resilience via computational approaches:

Tier 1: The Narrative aspect of Democratic Resilience

Conceptualising democratic resilience as a dynamic concept (Holloway and Manwaring, 2023) the related analysis objective is to explore how a democratic system withstands dynamic shocks and threats. Among a constellation of threats, pernicious narratives feed into most of these (Biddle et al., 2025). For instance, the Australian Government's own identification of eight threats to democracy lists six which have a narrative element: 'foreign interference', 'social media and platforms', 'misinformation and disinformation', 'polarisation and division', 'discrimination and intolerance' and 'dissatisfaction with government and governing' (Strengthening Democracy Taskforce, 2024). By example, populism (covered in more detail in §2.2) is a commonly studied narrative which most directly undermines democratic resilience as it seeks to delegitimise state institutions including the executive, legislature and judiciary.

Tier 2: Narrative analysis to identify and quantify threats to democratic resilience

Within narrative analysis, the objective of study is typically to discover and decompose narratives arising in political discourse. Whilst surveys and focus-groups continue to be the core methodological approaches for many organisations and government efforts, expert-labelling of texts from a variety of sources (e.g. social-media, op-eds, news, speeches etc.) has the power of converting unstructured, 'messy' text data into structured, tabular data forms which lend themselves to the kind of quantitative analysis we are concerned with. In this realm, narrative labelling is often operationalised via framing analysis, either via specific frames (e.g. 'immigrants are victims') (Mendelsohn et al., 2021) or more generic approaches (e.g. 'fairness and equality' vs. 'economic' vs. 'political') (Boydstun et al., 2013) (covered in 2.1 below).

Tier 3: Computational methods to scale narrative analysis

Third, computational methods can be employed to automate and scale a variety of useful 'text-as-data' tasks, supporting quantitative insights on narratives related to democratic resilience. There has been a marked evolution of these technologies over the last three decades (§3), with traditional NLP

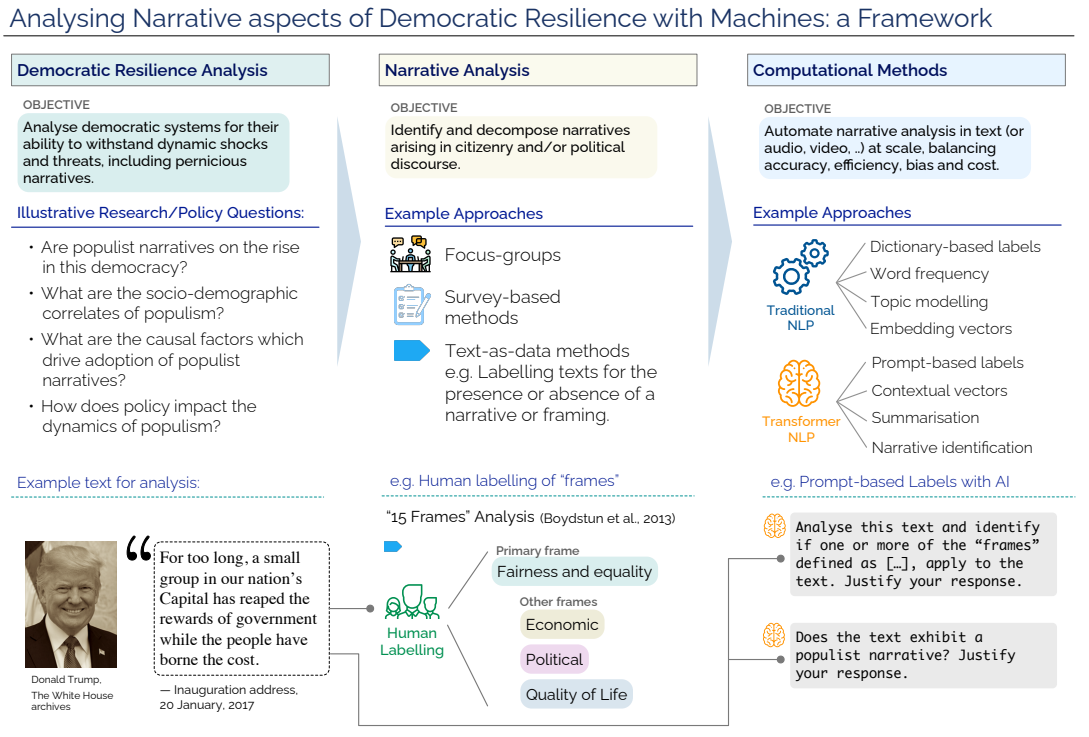

Figure 1. **A framework to connect the democratic resilience challenge to narrative analysis with transformers.** *Left panel: Democratic resilience analysis requires a dynamic approach, one aspect being that of narrative tracking, especially pernicious narratives such as populism (see §2.1). Centre panel: Narrative analysis seeks to discover and decompose narratives arising in texts including public discourse, typically through focus-groups, surveys, or human labelling using framing analysis. Right panel: Computational methods can scale up narrative analysis through a variety of means, including traditional NLP approaches, and lately, transformer-based techniques such as direct prompting..*

methods treating text simply as a collection (a 'bag') of disconnected words without context, compared to subsequent waves of NLP technology which are able to handle semantic components of words, and even sentences in their machinery. However, with the advent of the transformer revolution, powering amongst other technologies the wave of generative AI chat bots, we are entering an analytical world where models can be asked direct, nuanced questions about narratives related to democratic resilience and in response provide human-level intelligence at a fraction of the cost. It is this revolution that is the focus of our technological study (§3.3), drives our demonstration exercises (§4), and motivates our research agenda and data recommendations (§5).

2.1. Connecting Democratic Resilience to Framing Analysis

What are the narrative aspects of democratic resilience? Here, we integrate widely used definitions of both democratic resilience and 'framing' to narrow in on the focus of our study. First, according to Holloway and Manwaring's definition of democratic resilience, a population riven by discordant narratives is a threat to democracy itself. For them, democratic resilience is that quality of a democratic system which can stand despite the narrowest intersection of agreement on the issues that matter. Namely, 'democratic resilience' is (p.76),

> A system's capacity to withstand a major shock such as the onset of extreme polarization and to continue to perform the basic functions of democratic governance – electoral accountability, representation, effective restraints on excessive or concentrated power, and collective decision-making.

As such, and integrating Snow and Benford's decomposition of framing in social movements, a resilient democracy is one which can hold on to its foundational apparatus, even when the society at large is captured by highly opposing diagnoses, prognoses, or motivations regarding the issues of the day. Holloway and Manwaring (2023)'s wide-lens review of literature on democratic resilience finds that (Table 2, p.72) 'diversity (of groups, individuals)', 'learning', and 'community participation and inclusion' are amongst the common characteristics of resilient systems. So democratic resilience is about finding ways for diverse narratives to *co-exist* – i.e. to challenge the status-quo, and/or seek inclusion – whilst at the same time, being wary of narratives which foment excessive polarisation, distrust of electoral institutions and representation, and undermine social cohesion. The most resilient democracy may be able to withstand the worst of these narratives, but recent international experience suggests that democratic systems can be surprisingly susceptible to hitherto marginal narratives.

Second, a common approach to classifying narratives for quantitative social science, is to draw on theories of *issue-framing* (or just 'framing'). Here, one or more positions on any given topic or issue can be identified by examining how a communicator crafts their message on the topic, i.e. how they 'frame' the topic. Entman (1993) defines framing thus (p.52) (emphasis added),

> To frame is to **select** some aspects of a perceived reality and make them **more salient** in a communicating text, in such a way as to **promote** a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described.

Key in Entman's definition is *selection* and *saliency*, to serve the objective of *promotion* of the speaker's particular messaging on the topic. As such, issue framing can be seen as a specific aspect of narrative. Where narrative involves characters, events, and often a temporal component, framing is concerned with the specific treatment of an issue, event, or topic by the speaker who wishes to convince the receiver of the same. Chong and Druckman (2007) makes this connection in their more recent theorising (p.104) (emphasis added),

> The major premise of framing theory is that an issue can be viewed from a variety of perspectives and be construed as having implications for multiple values or considerations. Framing refers to the process by which people develop a **particular conceptualisation** of an issue or **reorient their thinking** about an issue.

Critically then, issue-framing applies equally to the speaker *and* receiver. The speaker seeks to promote a particular framing of an issue (*ala* Entman), with the objective of bringing the receiver to the same perspective (*ala* Chong & Druckman).

But framing is not merely an individual consideration, in democracies, individuals freely join with others to form *groups*, groups join again to form *movements*. And here, at these higher orders of organisation, social theorists see that framing applies equally well. Indeed, there is an important *interplay* between the framing of the group or movement and the individuals who compose it. Snow and Benford (1988) write (p.198),

> Movements function as carriers and transmitters of mobilizing beliefs and ideas, to be sure; but they are also actively engaged in the production of meaning for participants, antagonists, and observers. ... They frame, or assign meaning to and interpret, relevant events and conditions in ways that are intended to mobilize potential adherents and constituents, to garner bystander support, and to demobilize antagonists.

So together, we find that narratives matter to democratic resilience (at least) because of the power of framing in the hands of political leaders to coerce reality to their messages, because of the way that framing can form and locate political identities, and because framing can mobilise the masses.

For these reasons, we should not have been surprised by our earlier observation that the majority of threats to democracy identified by the Australian Strengthening Democracy Taskforce have a narrative component.

2.2. So which Frames Matter to Democratic Resilience?

The literature offers up a number of recognisable narratives that any democratic society may wish to be wary of. We briefly consider three.

First, the **'Us vs. Them'** narrative is the most transparently polarising, reducing complex, multi-dimensional issues down to a single dimension, and at its worst, encouraging a kind of 'purity test' of belief for those who adhere to one 'side' of the political landscape or the other (McCoy et al., 2018; McCoy and Somer, 2019; Hetherington and Rudolph, 2015; Iyengar et al., 2019). Despite wide variation in how a specific country becomes extremely polarised, authors contend that the subsequent undermining of democracy follows a very common path: political opponents dis-engage from meaningful communication and compromise; government reform agendas are gridlocked; each side dismantles the other's legislative agenda immediately on gaining power; and in the end-state, democratic transfer of power falls under restrictions of civil liberties and sees the military's entry to replace government.

Relatedly, narratives of nationalism can also work to undermine a plural, inclusive democratic society. As argued by Mylonas and Tudor (2021), we should distinguish between nationalism which draws on ideals of freedom and justice (e.g. the American and French Revolutions), and (p.111), "**exclusionary nationalism** (also called ethnic or essentialist nationalism)". Exclusionary nationalism shares rhetorical approaches with the 'us vs. them' narrative in its reductionism, i.e. driving a hierarchical narrative along racial, ethnic or religious grounds, 'othering' some group or groups in society to achieve some idealised, patriotic objective.

Third, as Algan et al. (2017) ominously begin, 'a spectre is haunting Europe and the West – the spectre of **populism**'. Populism is perhaps the narrative which is most directly opposed to the democratic project, for its proponents set themselves against the very institutions of democracy itself – established political parties, established institutions of state, even the judiciary, are all said to be held captive by 'elites', set against the best interests of the common people. Here, former president Donald Trump's dark framing of modern America, promising to 'drain the swamp' (i.e. of elites), and 'make America great again', spoke to the felt experiences of the American people and dove-tailed with right-wing programming across print, radio, TV and online (Polletta and Callahan, 2017; Norris and Inglehart, 2019). The message was particularly appealing to those who felt like 'outsiders' in their own country due to social and economic marginalisation (Gidron and Hall, 2020).

But how prevalent are these narratives in a modern democracy? Are narratives of populism or exclusionary nationalism already making themselves heard in the United Kingdom, Australia, or the United States? Has there been a change over time? Where? Is this only for the online 'echo-chambers' or do we see growing evidence of these narratives in our public discourses arising in Parliament, print or on the air-waves?

The answer to all these questions requires empirical tools.

2.3. Aside: Social Listening & Surveillance

Before moving to empirical methods, it is worth spending a moment to distinguish present considerations from 'surveillance' orientated methods, or what is sometimes referred to as 'surreptitious social listening'. Listening – the active or passive act of receiving/collecting information from people in a given context – can take a variety of forms, best considered as sitting on a spectrum (Martino, 2020). According to such typologies, tracking of public narratives related to social cohesion and democratic resilience, sits in the 'active listening' category, and, according to Martino seeks to *enhance* democratic and foreign policy objectives by promoting trust and understanding. In a word, the government or actor who actively listens to citizens and counterparts, including those who hold opposing views,

demonstrates a long-term commitment to dialogue, relation-building, and engagement. Importantly, such active listening, or 'public discourse tracking' as we consider here, is far away from *surreptitious* listening, or what is otherwise known as surveillance or spying. Here, there are spaces for legitimate uses of such methods in democracies, but such uses must carefully balance personal rights and freedoms with policing and counter-terrorism objectives (Hagen and Lysne, 2016; Moore, 2011). Importantly, it should be acknowledge that all surveillance carries harm (Richards, 2013) either because of a chilling effect on civil liberties (Penney, 2016), or by introducing unhealthy power dynamics due to information asymmetries. Nevertheless, tracking public narratives related to democracy, where it seeks to understand, characterise, build trust, and engage, aims to support democratic values, not undermine them, and should be kept distinct from states or actors who seek to infiltrate the private sphere to advance coercive, controlling, or chilling purposes.

3. From experts to transformers: the evolution and revolution in narrative analysis methods

With the advent of more advanced natural-language-processing (NLP) tools over the past two decades, and in particular, with the rise of powerful large-language-model (LLM) generative artificial intelligence (AI) systems, there has been a rapid growth in methodologies which seek to automate labelling tasks, at scale, with near human levels of accuracy.

The ultimate aim of such efforts is to generate a structured dataset where each record represents a given *text*, *meta-data* or *attributes* related to that text (e.g. the date, speaker, party affiliation, location, medium etc.), and a series of *labels* which indicate whether a given framing relating to some issue exists or does not exist in that text. From there, the wide array of standard empirical tools can be applied to generate analytical objects (e.g. time-series, histograms), statistical tests, and causal inferences.

We focus here on non-survey methods, charting the pathway from labor-intensive labelling to automated approaches.[1]

Traditionally, labelling texts for the presence of a frame has followed a standard procedure by human experts which we cover briefly to set the scene to ground our discussion of automated approaches. In their review of framing theory, Chong and Druckman (2007) note that the identification of frames in communication by domain experts has become a, 'virtual cottage industry' (p.106). And, whilst they note that no uniform definition of how to proceed exists, the 'most compelling studies' follow six standard steps: 1) identify the issue (e.g. 'immigration'); 2) isolate an attitude (dimension) (e.g. 'posture towards refugees'); 3) define frames (e.g. 'humanitarian approach' versus 'border security approach'); 4) select sources; 5) develop a validated coding scheme; and 6) undertake the coding.

Examples of the standard approach in practice include Terkildsen and Schnell (1997) who analyse weekly news coverage of the women's movement over four decades for the presence of frames related to sex roles, feminism/anti-feminism, political rights, and economic rights, and Jennings and John (2009) who code discourse frames including social inclusion/exclusion in successive releases of the poverty reduction strategy in Ontario, Canada. The approach is also used at the level above frames, i.e. to identify the presence or absence of a given set of issues in texts such as in Smith-Carrier and Lawlor (2016) who code the Queen's speech to Parliament for policy areas and join this to Gallup surveys on 'the most urgent problem' to examine whether the government agenda responds to public opinion.

3.1. 'Traditional' NLP approaches to automation

Whilst the standard approach is still widely used and accepted in rigorous analysis of public discourse, the recent digitisation of all areas of human social, economic and political life, and the massive quantity of discourse artefacts this has created (Mauro et al., 2016), has seen a move to automation of the standard approach, in particular, Step 5 – the coding of texts. Yet, as will be shown below, automating, and so scaling, the standard approach to framing analysis is just the starting point for automated methods with new technological frontiers opening up new ways of analysing and tracking public discourse.

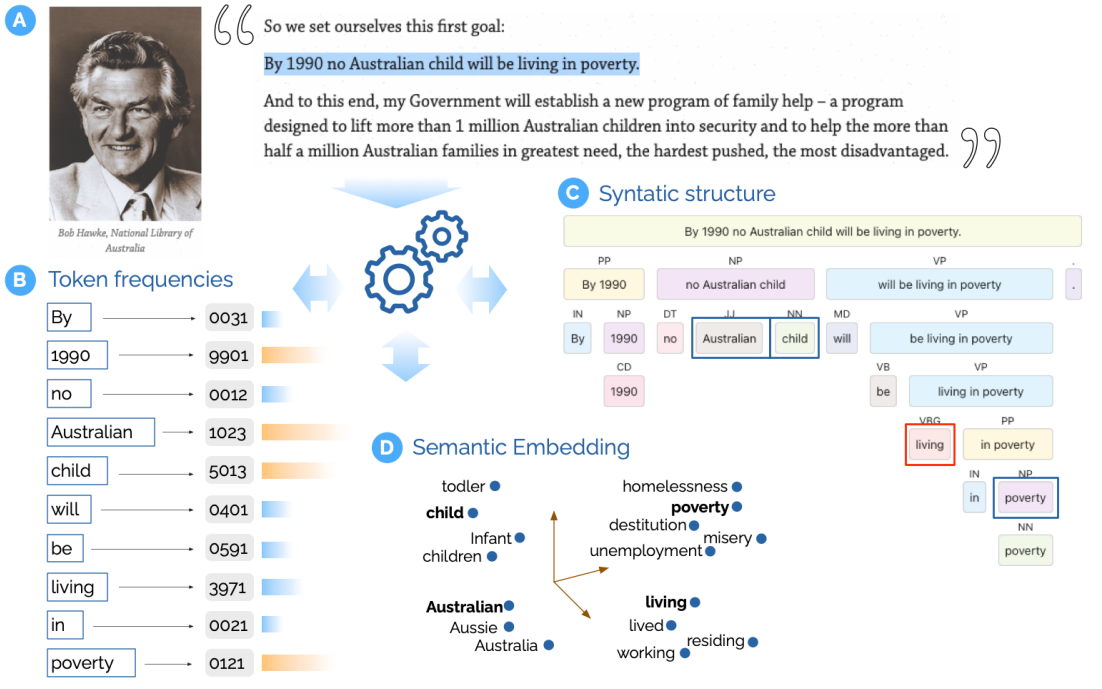

Figure 2. **Traditional (pre-transformer) NLP approaches to analysing discourse.** *Given some text (A), unique words (tokens) can be counted and turned into frequencies (B), syntactic structures can be identified and named-entities can be labelled (C), and words can be associated with their location in high-dimensional, pre-trained, semantic embedding spaces (D). .*

Here, the fundamental challenge of computation lies at the heart of technological advance: how to encode the richness of human semantics in machine-readable format? Computational linguistics is the field of computer science perhaps most associated with attempting to answer this question, and multiple techniques have arisen over recent times which are still in use by computational social-, political-, and economic- scientists alike, to study public discourse (Bail, 2014; Gentzkow et al., 2019).

Indeed, we can chart the trajectory of NLP research which, starting with proximate abstractions of narrative framing analysis (e.g. word occurrences, topic modelling, semantic analysis), has been narrowing closer towards the object itself – trying to identify accurately elements of speech which encode one framing or another (rich, nuanced and contextual) in text, at scale. Along the way, several ingenious approaches to quantifying discourse dynamics aside from simply 'coding' has arisen, yielding insights derived from much larger volumes of text than could ever have been feasibly handled by human means alone.

In figure 2 we summarise the fundamental technologies used in traditional (pre-transformer) NLP research applications in the social science. We discuss each in turn.

Token frequencies

– A classic NLP approach is to define a unique vocabulary for a corpus of texts, typically dropping short, non-meaningful words - 'stop-words' ('a', 'to', 'is', 'the', etc.), and reducing words to their root or stem form (e.g. 'running' → 'run') so as to create a minimal set of meaningful word forms, known as 'tokens', by which any text can be represented. The computational task is then to count up the number of times each token appears in a given text ('token frequencies'). Importantly, such frequency representations lose any *contextual* information about the token. As such, these representations are called 'bag-of-word' approaches, since it is as if all of the words have been cut out of their context

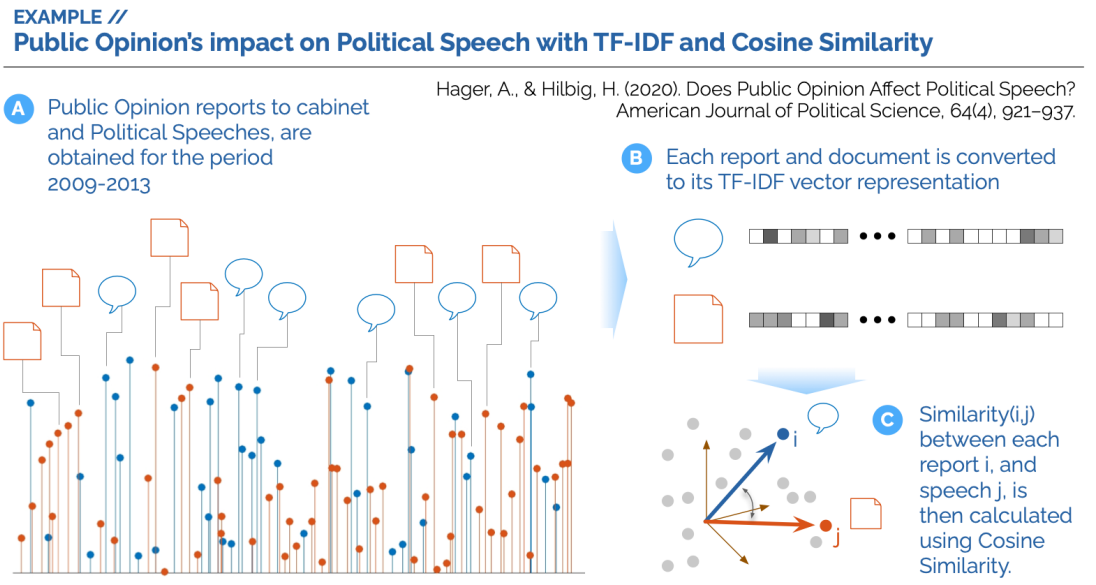

Figure 3. **Causal analysis of public opinion's impact on political discourse.** *In Hager and Hilbig (2020) a large corpus of German cabinet public opinion surveys were obtained after release, and, together with political speeches, press releases and other announcements of cabinet members (A), a regression discontinuity causal analysis framework was undertaken. After standard pre-processing, 3,860 length tf-idf vectors for each report and speech were calculated (B), enabling the calculation of semantic similarity between each report—speech pair via cosine similarity (C)..*

and tipped into a bag, regardless of order. Whilst such an approach may seem overly reductive, further processing can recover remarkable insights. For instance, by converting token frequencies ('tf') for a text into a vector, and then modifying token frequencies by a scarcity weighting for that token across all texts ('idf', inverse document frequency), researchers can then compute semantic distances between texts in this vector space.

An example of such an approach is found in Hager and Hilbig (2020) who analyse public opinion survey texts along with political speeches to examine the causal relationship between public opinion and political discourse by calculating tf-idf vector distances (see Fig. 3). Alternatively, tf-idf vectors can be fed into topic-modelling algorithms which perform unsupervised clustering[2] over the vectors typically using latent Dirichlet allocation (LDA) methods, such that topic dynamics can be followed over time (Jelodar et al., 2019). Examples of this strategy include Vidgen and Yasseri (2020b) who study the impact of petitioners on policy direction in the United Kingdom, Bonilla and Mo (2019) who analyse discourse topics in newspapers related to human trafficking, and Goyal and Howlett (2021) who study Covid-19 policy responses.

Other extensions to token frequency approaches include counting specific tokens which are said to encode attributes of language like sentiment. So-called 'dictionary-based' sentiment classifiers such as the Lexicoder Sentiment Dictionary (Young and Soroka, 2012) or VADER - Valence Aware Dictionary for Sentiment Reasoning (Hutto and Gilbert, 2014) are two examples of such approaches. Pauwels (2011) applies such an approach to analyse populism in Belgian party speeches and manifestos. The method introduced simply calculates the fraction of all words in a document appearing in the populism dictionary – effectively contending that populist word usage is a tell-tale (i.e. *definition*) for populist narratives[3].

Whilst tf-idf methods are reasonably suited to topic modelling, the fact remains that since tf-idf vectors treat language use as entirely context-free, bag-of-word approaches are inherently limited in their ability to accurately identify nuanced semantic patterns which are inherent in narrative discourse analysis. Sentiment analysis being a good example of such limitations – dictionaries of terms associated with positive or negative sentiment are very labour-intensive to build, are domain specific, and are entirely context free, making sarcasm or irony a major stumbling block in their application (Grimmer and Stewart, 2013).

Semantic embeddings

– A major revolution occurred in NLP at the start of the second decade of the millennia, with the development of massively trained neural word-embedding models. Here, neural network architectures are trained to guess the correct token from surrounding tokens, or alternatively, the surrounding tokens from a given token. For the first time, computational representation of human language went beyond mere frequencies of unique index values, to actually associating words with their meaning, as revealed by their (local) context of use. Such models are referred to as 'word embedding' models (or sometimes, simply, 'embeddings'), with Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) among the most popular and widely used embeddings. By passing these models billions of sequences of human text (typically harvested from online sources), the models are able to accurately represent each word with a vector location in high-dimensional space. For instance, the vector for 'child' will be close - in vector space - to 'toddler', 'infant', and 'children'.

For the practitioner, word embeddings are very easy to accommodate, since any word (including stop-words; without stemming) can be passed to a pre-trained embedding (essentially a neural model), which deterministically produces a vector for that word. By repeating this process for a sequence of words in a text, many vectors are obtained, which can then be averaged to obtain a document location in 'semantic space'. From here, down-stream tasks like topic-modelling can be performed as introduced earlier.

However, there are more elaborate uses of word embeddings. A researcher can take the generically trained word-embedding model, and then further train the model on a sufficiently large sample of study texts, a process known as 'fine-tuning'. In this way, the word-embedding model itself starts to represent more closely the way that words encode semantics as revealed by the language in the study materials itself. An example of this approach is found in Kozlowski et al. (2019)'s ground-breaking study of class, where millions of book texts from each decade spanning the 1900s to the 1990s were each fed into a initialised Word2Vec model to fine-tune it to the particular semantic relationships of that decade. The authors then looked at semantic similarity between certain words which denote class and other words which denote dimensions of culture (see Fig. 4). Whilst not directly assessing framing within narrative discourses, one can see how deep cultural associations – the underpinning of narrative framing (ala Polletta and Callahan (2017)) – can be quantified and tracked over time using such an approach.

Syntactic structure

– Whilst grammar, on its own, does not closely associate with one narrative or frame, syntactic analysis empowers research methods to create more nuanced representations of narratives. Today, the field is almost entirely dominated by transformer methods, but there are still use-cases for traditional approaches, especially where data are scarce or computational resources are constrained. Here, hidden Markov Models or their derivatives represent pre-transformer approaches to labelling parts of speech McCallum et al. (2000); Brants (2000). Alternatively, other methods have been developed to correctly identify who is speaking in a sentence where pronouns replace names. This is especially important when conducting sentence-level analysis (a common approach in NLP), especially in political analysis, since knowing that the 'they' refers to the government, or opposition, or 'she' refers to a shadow-minister or the Prime Minister are critical pieces of information for any quantitative assessment. In NLP this task is referred to as 'coreference resolution' and fast, accurate methods are in common use (Clark and Manning, 2016a,0).

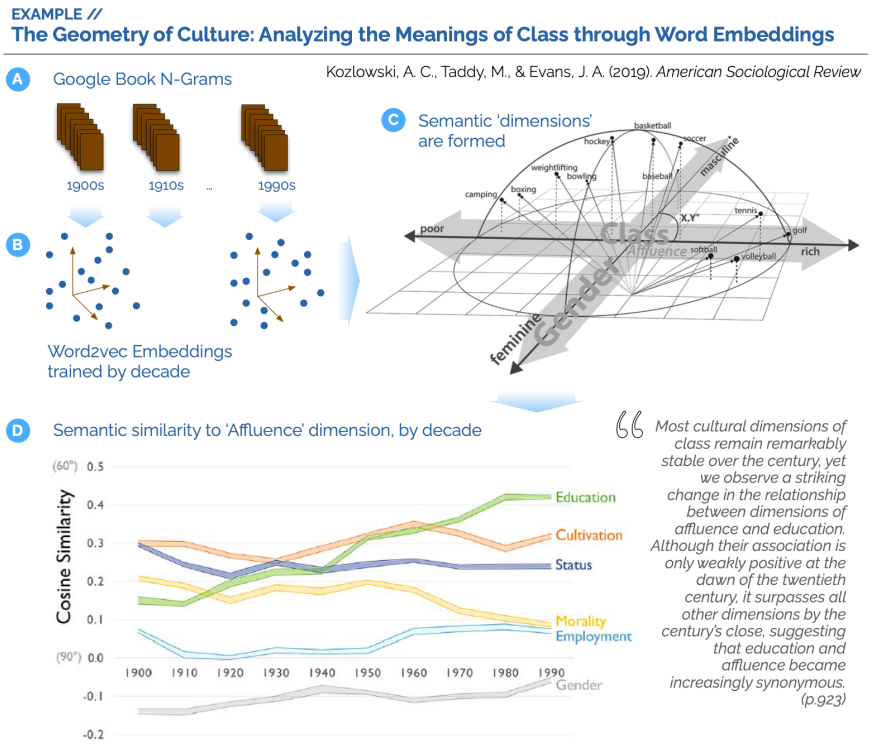

Figure 4. **Using word-embeddings to analyse the foundations of narrative framing in millions of texts.** *Kozlowski et al. (2019) take US provenance Google Books n-gram texts, per decade (A), to fine-tune Word2Vec neural embedding models (B), such that distances between semantic dimensions along word-pairs lines in each decadal model can be calculated (C), to track the cultural association of e.g. 'Affluence' with these dimensions over time (D)..*

3.2. Ensemble approaches

In practice, traditional NLP workflows typically combine tools together in a sequence of steps, known as a 'pipeline' – a *combination* of transformations is used to achieve higher accuracy on a given research objective. This is the approach taken to a series of studies in computer science literature which define the task as a multi-class classification 'framing' problem[4]. However, this literature is of limited application to tracking conceptual framing (in the spirit of Entman, i.e. narratives to *persuade* or *promote*) since 'framing' is used in this literature to denote a *dimension* of an *issue* rather than what might be called the conceptual framing itself. For example, Boydstun et al. (2013)'s 15 generic policy 'frames' include 'Economic', 'Public opinion', and 'Cultural identity' – these are generic dimensions of an issue and do not carry enough meaning to denote a specific narrative or conceptual framing. In the literature these 'frames' are then instrumented across multi-issue labelled datasets such as the Media Frames Corpus (MFC) (Card et al., 2015) and the Gun Violence Frame Corpus (GVFC) (Liu et al., 2019a) and analysed using a variety of multi-class classification algorithms (Field et al., 2018; Zhang et al., 2023).

However, not all computational literature follows this path, with several studies taking the more nuanced approach to framing analysis required for tracking narratives of democratic resilience. For example, Vidgen and Yasseri (2020a) offer a promising approach to tracking nuances in Islamic hate-speech in 4,000 social media posts. Namely, they distinguish between 'strong' and 'weak' islamophobia – a distinction which can be considered as two frames within the general attitude of islamophobia. In

the study, they trial seven different pipelines, leveraging all of the traditional NLP approaches mentioned thus-far: tf-idf vectors, word embeddings, fine-tuned word-embeddings, sentiment, and syntactic features. They also trial a number of different classification algorithms from logistic regression, Naive-Bayes, through random forests, and support vector machines (SVM) and even deep learning (neural methods). The most accurate approach they identify combines features, and uses the SVM classifier, obtaining 72.17% accuracy on their training data ($n = 1,341$), and 77.3% accuracy on a further set of 300 hand-labelled test tweets. Likewise, Morstatter et al. (2018) study narratives of support and opposition (polarity) across 10 framings related to the balistic missile defence (BMD) issue in Europe. However, through hand-coding, and with training a multi-class NLP classifier, accuracy is low (less than 0.5 across classes). Cocco and Monechi (2022) pursue populism narrative analysis via adding machine learning to a bag-of-words approach, processing over 240,000 sentences from 268 electoral manifestos across 99 parties using a random forest (RF) classifier. Here, rather than a list of populist words, the method mechanically labels (i.e. *defines*) as 'populist' all sentences that arise from texts of parties labelled populist in the PopuList database (Rooduijn et al., 2024). The RF tool is trained to pick sentences from these texts (represented by cleaned and stemmed word occurrence), as opposed to non-populist texts in the training data.

Closer still to the narrative analysis task are studies which seek to track the incidence of 'micro-narrative' phrases within a large body of texts. Here, the work of Ash et al. (2024) and Anantharama et al. (2022) seek to identify 'Entity – Verb – Entity' (EVE) triplets, through a multi-step pipeline using word-embeddings (Ash et al.) or transformer-based sentences embeddings (Anantharama et al.), together with syntactic parsing and co-reference resolution. These approaches enable the tracking of phrases like 'God–bless–America' or 'indigenous voice–enshrine–constitution'. By clustering a set of these micro-narratives and associating with a particular framing, narrative components can be tracked across speakers, and through time.

So how well do non-transformer methods perform at traditionally hand-coded approaches to tracking discourse? A study by Nelson et al. (2021) considers three traditional methods (dictionary-based, tf-idf supervised, unsupervised topic discovery) applied to a large number of hand-coded news articles with a hierarchical identification schema relating to inequality. They find that tf-idf style classifiers were the most accurate approach in both identifying articles and tracking themes over time. However, with a combined precision/recall (f1) score of around 0.8, it would be hard to conclude that 'human-level accuracy' was obtained by the best method. However, the authors took methods 'off the shelf' and it is likely that more performance could be obtained by skilled practitioners.[5]

3.3. Harnessing the AI–transformer revolution

Sometime during 2016 or 2017 work began in the Google Research and Google Brain research groups on a revolutionary neural approach to the fundamental machine—human language problem. Known as the 'transformer' architecture, and introduced by the now famous paper entitled, 'Attention is all you need', Vaswani et al. (2017) demonstrated how, with vast quantities of human language training data, the new architecture could be taught, via the simplest of tasks – guess the next word in the sequence – to apparently infer complex linguistic and semantic properties of human speech. Subsequent work followed at Google and other labs, extending, refining, and enlarging the scale of the approach, resulting in breathtaking performance by transformer models on standard NLP benchmarks (Devlin et al., 2019; RaffelColin et al., 2020; Brown et al., 2020).

In Fig. 5 a generalised schematic of the transformer model architecture is shown. As neural models (i.e. large matrix multiplication models), transformer models take input 'at the left' and pass the transformed matrices of information from 'left to right' to arrive at a statistical prediction for the most likely word (token) that might come next. The model is trained by passing it billions of tokens from real sequences of human authored text obtained from online repositories of news, books, and social media, but hiding (or 'masking') the next word from the model. The model's guess for that next word (token)

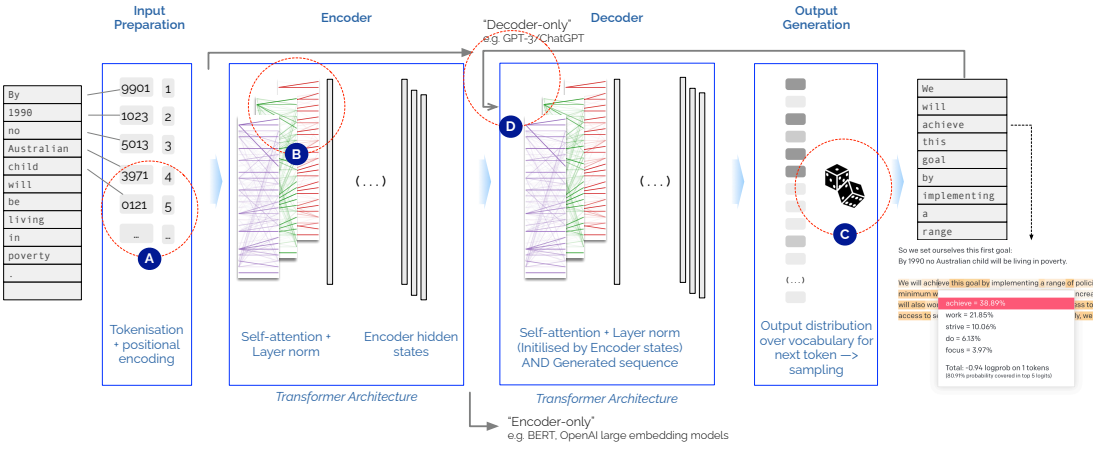

Figure 5. **Schematic of the large-language-model (LLM), encoder-decoder architecture.** *For the first time in NLP technology, text is fed into the model as a* sequence *(A), with positional encoding ensuring that subsequent steps retain knowledge of this ordering; transformer blocks are then applied in encoder, and/or decoder layers which enable the model to simultaneously attend to many aspects of the text (e.g. pronouns, verbs, nouns, punctuation) (B); the final layer is a statistical prediction for the next word (token) (C); and, each new token can be added to the input sequence, such that the model is able to keep generating new words (tokens) based on the generative steps already undertaken (D)..*

is then compared to the real (ground-truth) word (token) and the internal model weights are adjusted accordingly to nudge it towards more accurate next word prediction.

Importantly, these models can be auto-regressive in nature, that is, they can read their own produced tokens as input. This means that rather than a single next word prediction, they can be set to repeatedly generate next words, feeding the expanding sequence back into their input at each turn. This makes them truly 'generative' models, and has led to a large number of applications to traditionally complex NLP tasks. What once required a complex, multi-step pipeline to achieve, can now be achieved by simply describing the task to the model in plain English, and subsequently passing the textual record (phrase, sentence, paragraph, or even whole document) to analyse to the model, and having the model 'generate' the information, label, or decision as desired.

Perhaps most remarkably, recent LLMs have been shown to exhibit a kind of *emergent* intelligence, performing tasks that cannot possibly have been part of their training data, even on complex logical problems hitherto thought of as well beyond the reach of 'thinking machines'. For example, Trinh et al. (2024) demonstrate 'AlphaGeometry' – a LLM trained on millions of examples of Euclidean plane geometry – which is able to perform close to gold medal (human) performance outcomes on the latest International Maths Olympiad problem set. Similarly, Bubeck et al. (2023) demonstrate OpenAI's GPT-4 (General Pre-trained Transformer, 4) ability to stack physical objects, draw unicorns (via coding), compose music and write complex code. Whilst still a topic of debate in the field of computer science, there are indications that recent LLMs have some kind of 'world model', that is, despite "simply" next token predicting, the model's training has somehow conferred some degree of knowledge of how physical objects and abstract concepts relate in real worlds, perhaps even encoding so-called 'common knowledge' (Mitchell and Krakauer, 2022; Mitchell, 2023).

Can these new technologies be turned to the challenges of tracking narratives in public discourse at scale? We consider several promising applications.

Transformers as (better) encoders

– From the perspective of today's remarkable generative models, one can easily forget that the *non-generative*, encoder-only function of transformers was their first astounding achievement; a capability still used actively by researchers and engineers alike. With reference to Fig. 5, 'encoder-only' models are comprised of the left-hand encoder transformer block only, and do not pass their output to the generative, decoder block to form next word/token outputs. Instead, the output from the encoder-only system is a real-valued vector, its length being dependent on the LLM in use, representing the rich, high-dimensional, latent semantic content of the input sequence. Part of the reason for the early, rapid adoption of encoder only models has been that the vector-representation produced is a 'drop-in' replacement for either a tf-idf or word-embedding vector as introduced earlier. Hence, any ensemble pipeline NLP method using these earlier representational vectors would likely be improved by the richer, contextual and sequential representation of the input text encoded in the LLM encoder-vector. Popular transformer-based embedding models for NLP applications include BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) (an optimised version of BERT), Sentence-BERT (Reimers and Gurevych, 2019) (a sentence-to-sentence similarity optimised version), and BERTopic (Grootendorst, 2020) (a system for clustering texts using BERT embeddings). Examples of these technologies applied to related topics include Mendelsohn et al. (2021) who fine-tune the RoBERTa model to predict human-labelled tweets related to immigration. They find a significant uplift in performance from fine-tuning to their dataset, with an average F1 score above 0.65 on a variety of framing tasks relative to 0.61 for RoBERTa without fine-tuning, or just 0.3 for a simple logistic regression on tf-idf vectors. Bonikowski et al. (2022) similarly employ a fine-tuned RoBERTa model with active labelling to train a neural layer to identify paragraphs of text as populist. Again, they find a RoBERTa backed classifier superior to ensemble (random forest) models backed by word-embeddings or tf-idf vectors.

Transformers as labellers

– Perhaps the most obvious first application of the power of LLMs to understand and encode language, is to apply them to the labelling task ('step 5' in the Standard Approach, above), at scale. Here, it has become very apparent that LLMs are extremely good at this task, outperforming both expert and crowd sourced human labelling. Törnberg (2023)'s 2023 study shows that OpenAI's ChatGPT-4 achieved not only higher accuracy (>0.90 vs. 0.82), but also much higher inter-rater reliability (>0.95 vs. 0.65) than expert human coders, yet with statistically similar levels of bias. Rathje et al. (2024) find similar success in a related task in the psychological sciences, demonstrating convincingly the power of generative transformer (decoder) models to undertake such tasks, using prompting directly with, or without definitions of concepts (e.g. 'moral foundations') to label over 16,000 posts on Reddit with high accuracy. And most recently, applied to labelling party positions on a Left–Right spectrum, Mens and Gallego (2025) label positions of thousands of tweets by US congress members, British party manifestos (sentence at a time), and EU policy speeches in 10 languages.[6] Through ground-truth comparison they find the strongest language models perform at or above 0.90 correlation with human annotation, and out-perform examples of all prior NLP approaches (tf-idf vectors, word-embeddings, and BERT fine-tuned). Indeed, small amounts of human-labelled data can now be used to directly train LLMs towards higher accuracy in labelling large datasets very cheaply (Wang et al., 2021), or to 'teach' simple models to do the same, lifting their performance in a process likened to slingshotting a space probe around a larger planet to go further, faster (O'Neill et al., 2023).

Transformers as synthetic data generators

– Alternatively, instead of dramatically reducing the cost of labelling real data, LLMs can be used to create large, synthetic textual datasets, with a highly controllable form. The benefit here is that LLM systems can be evaluated for accuracy, bias, and reliability on large amounts of labelled synthetic data where either real data are unavailable, or are very costly to curate, meaning that a system can be deployed on real data after significant refinement has already occurred on synthetic data. Techniques

here focus on enforcing structured ouputs (Willard and Louf, 2023), or reducing hallucinations (Ayala and Béchard, 2024). However, recent advances by LLM platforms have meant that model control is now even more precise, with the ability for models to provide exact structured outputs fitting a researcher's schema (Pokrass, 2024).

Transformers as direct framing analysers

– Finally, recent research has shown that LLMs can be used directly as highly capable 'language models'. That is, that LLMs can be used to identify issue-framing in text directly, with low cost, and low bias. Angus and O'Neill (2024) introduce 'paired completion', a technique whereby LLMs are primed with examples of statements from one framing or another on a particular issue, and then the text to be analysed is fed as a 'completion' to the model. The method collects the likelihood that the model would have generated the completion text, conditional on already having 'said' the priming text. By comparing the completion likelihoods, a quantitative metric is obtained for the text's *textual alignment* with one frame or another. The study reports balanced accuracy scores as high as 0.93 across three synthetic narrative datasets covering climate change, domestic violence and misogyny.

Alternatively, LLMs can encode an entire 'world' of a given source such as a particular newspaper or online media outlet (Guo et al., 2022). The approach here is to *fine-tune* a LLM, a method which trains a LLM to learn the particular patterns of speech of a given source corpus, by feeding it (say) news from one outlet only. In this way, a series of fine-tuned LLMs can be trained, one per news source, and then each LLM can be analysed for how it might complete a sentence, or use a key word in a sentence (e.g. complete the sentence, 'Immigrants travelling by boat to Australia are a ___ to our society.')

And, we note that whereas our emphasis in this study is the tracking of specific narratives related to democratic resilience at scale (i.e. 'supervised' analysis), one can also employ transformers to assist with narrative discovery (i.e. 'unsupervised' analysis). Here, the "Concept Induction" tool, LLooM, of Lam et al. (2024) represents a good example of the possibilities. With LLooM, a user can input a large body of text and have the sysem iteratively uncover latent concepts, at a variety of levels of analysis (e.g. 'subconcepts' such as "Women's responsibilities," "Traditional gender roles," and "Power dynamics and women" lead to an emergent meta-concept, "Criticism of traditional gender roles"). It is likely many other applications will emerge in the near future along these lines.[7]

3.4. Risks associated with using Transformers

Using LLMs for framing analysis at scale is not without concerns. Generative LLMs present outstanding opportunities for quantified approaches to tracking narratives at scale, but their effective use requires a robust understanding of a suite of risks (Bommasani et al., 2021). Attention must be paid to the provenance of training data, bias in generated text, processing cost, and open or closed access, given that many of the most performant models are closed-source, run by large platforms. As mentioned, the training of LLMs (next word prediction) does not guarantee correctness or completeness in responses obtained, and even the most frontier LLMs can still be tripped up by seemingly trivial completion tasks[8]. In the computational social sciences recent reviews of LLM opportunities and risks catalogue these issues in more detail (Thapa et al., 2025), with some authors arguing that the impact of LLMs will go beyond efficiency and opportunity to start re-shaping the scientific method itself (Fecher et al., 2023), and exacerbate existing resource and access inequalities in the political economy of ideas (Luitse and Denkena, 2021). On the former, work has already been done to create entire *ecosystems* of LLMs (Agentic-AI systems) which operate as virtual, automated scientific laboratories reading each other (AI) lab's work, and improving upon it (Schmidgall and Moor, 2025). For now, these studies are largely confined to computer science domains, where algorithmic optimisation is well handled by such systems, nonetheless, the question of how far science is willing to bring LLMs into the scientific-method is open and dynamic.

Direct assessments of political bias, perhaps most relevant to the present study, in LLMs is important and requires constant updating due to the rapid developments in model releases and capabilities.

For instance, Choudhary (2024) put four recent, major commercial "models" ('ChatGPT-4, Perplexity, Google Gemini, and Claude' - unfortunately, we do not learn which exact model variant in the paper) through major political orientation questionnaires such as the Pew Research Centre's Political Typology Quiz, PoliticalCompass.org, and ISideWith political party quiz, and find that all models lean left of centre, with "ChatGPT-4" and "Google Gemini" aligning with "establishment liberals" (three categories to the left, of four), whilst "Claude" and "Perplexity" aligned just left of centre, as "outsider left" (one category to the left, of four). Political Compass results were similar. These results appear robust to model choice and language (Rettenberger et al., 2025; Pit et al., 2024). It is important to note that such outcomes are more likely when models are used in a 'zero-shot' manner, that is, when they are asked, without examples or further definitions, to answer a direct question (see a demonstration of prompt variation below §4). Further, we note that human annotation in politically relevant domains is itself subject to annotator bias (Plisiecki et al., 2025), as such, it is unlikely that LLMs (which rely on human guidance) will obtain entirely 'bias-free' behaviours. For now, these considerations, like any research involving judgements (human or machine) require thoughtful design and interpretation decisions with replication across models/labs/contexts, as ever, the most likely protection against false-hood. In this light, AI documentation literacy is key for success in this direction (Smith et al., 2024).

4. Demonstrating Tracking Public Narratives with Transformers

We conclude this section by conducting demonstration exercises related to tracking narrative aspects of democratic resilience with transformer methods. Our aim is not to conduct a thorough empirical study with these demonstrations, but rather, to provide tangible examples of the new possibilities of these technologies: the kinds of implementation setup involved; how model choices impact insights; what a fully fledged 'tracking' project would demand.

In the first example, we exemplify the most direct use of LLMs, namely leveraging their ability to make human-level judgements regarding semantic content in natural language through 'just asking the model' in the spirit of Törnberg (2023), Rathje et al. (2024) and Mens and Gallego (2025) introduced above. By applying nine different models from three major providers, across five text excerpts across the populism—democrat spectrum, under two prompt treatments (90 model calls in all), we demonstrate both the efficiency with which AI judgements can be achieved, and some of the considerations practitioners and readers alike need to keep in mind when undertaking these exercises or consuming their findings.

In the second example, we move to a more scaled 'tracking' demonstration, by applying the quantitative LLM *paired-completion* approach (Angus and O'Neill, 2024) to the democratic resilience relevant issue of immigration. By using the method to analyse over 2,800 speeches from 11 prominent speakers appearing in the Australian (Federal) Parliamentary Hansard over 15 years we are able to generate time-series position statements on the 'border security'—'humanitarian' spectrum revealing quantitively the dynamics of this issue, including patterns related to the election cycle.

4.1. Identifying Populism with Transformers

Populism is a common object of study for the political analysis of text, and as noted earlier, is one of the key pernicious narratives related to democratic resilience. That said, defining the populism narrative is the first methodological challenge. Early studies typically looked to Mudde (2004)'s succinct rendering,

> "An ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' versus 'the corrupt elite', and which argues that politics should be an expression of the volonté générale (general will) of the people."

This definition focuses on the core idea of a binary struggle between (a monolithic) 'the people' and 'the elite'. However, more recent authors contend that populism is a latent construct which can only be truly identified via a multi-dimensional tool. Meijers and Zaslove (2021) make this argument and

ZERO-SHOT PROMPT

```
Does the TEXT exhibit a populist narrative?
Provide your answer from the set: [strong yes, weak yes, unsure, weak no, strong no], and a
brief (up to 10 word) justification for your response.
TEXT: {text}
```

EXTENDED PROMPT

```
Does the TEXT exhibit a populist narrative?
In providing your response, apply the following 5-dimensional framework of populism. A text
is considered populist if it fulfills all five dimensions.
 1. Manichean worldview: "Politics is a moral struggle between good and bad"
 2. Indivisible people: "The ordinary people to be indivisible (i.e., the people are seen as
homogenous)"
 3. General will: "The ordinary people's interests to be singular (i.e., a 'general will')"
 4. People-centrism: "Sovereignty should lie exclusively with the ordinary people (i.e., the
ordinary people, not the elites, should have the final say in politics)"
 5. Anti-elitism: "Anti-elite dispositions"
Provide your answer from the set: [strong yes, weak yes, unsure, weak no, strong no], and a
brief (up to 10 word) justification for your response.
TEXT: {text}
```

Figure 6. **Two prompt variants employed in the populism labelling demonstration.** *Both prompts contain the same task: to respond with a label from a five-item ordinal set (and brief justification), as to whether the TEXT exhibits a populist narrative (blue text in both prompts). The extended prompt (bottom) adds a five-dimensional definition (Meijers and Zaslove, 2021) to fix ideas and effectively raise the threshold for positive identification..*

develop a five-dimensional survey tool for validation with an expert judgement panel comprising 294 respondents 'with expertise in party politics' (p.382) across 28 countries[9]. Using exploratory factor analysis (EFA) and item response theory (IRT) they conclude that their five-dimensions of populism are a valid operationalisation of the latent concept. The five dimensions are given as (p. 384),

1. Manichean worldview: "Politics is a moral struggle between good and bad"
2. Indivisible people: "The ordinary people to be indivisible (i.e., the people are seen as homogenous)"
3. General will: "The ordinary people's interests to be singular (i.e., a 'general will')"
4. People-centrism: "Sovereignty should lie exclusively with the ordinary people (i.e., the ordinary people, not the elites, should have the final say in politics)"
5. Anti-elitism: "Anti-elite dispositions"

Given the more robust validation approach of Meijers and Zaslove (2021), in what follows, we shall adopt this definition in our second, expanded, prompt strategy.

Approach

To provide a concise demonstration of how generative transformers perform, we shall conduct a small text ($n = 5$) corpus evaluation, leveraging variation across commercial platforms (Anthropic, OpenAI, Google), model flavours (chat models vs. so-called 'reasoning' models) from these providers, and prompt design. On the latter, we will demonstrate two prompting strategies, starting with a 'zero-shot' style prompt, which leans on the particular model's 'world-knowledge understanding' (from training and alignment) of "populist narrative" compared to an extended prompt, in which we explicitly define the concept, and require the model to only label a text as exhibiting a populist narrative if it fulfils all five dimensions. Both prompts variants are shown in Fig. 6[10]. The extended definition flows from the 'ideational approach' to defining populism as a latent concept taken from Meijers and Zaslove

(2021) introduced above. For the purposes of the demonstration, the extended prompt treatment represents a more controlled setting (since we explicitly define the labelling policy) whilst simultaneously raising the bar for the positive labelling of the text as populist since the prompt requires the text to 'fulfil all five dimensions'.

Corpus

The small text corpus was constructed to contain a core of three texts representing populist politicians or populist political parties (Trump, Vlaams Belang, Left Front), plus two strongly pro-democracy politicians (Albanese, Obama). Of the two pro-democracy politicians, one text (Obama) gives a full throated affirmation of democratic institutions, whilst the other text (Albanese) represents a tricky case, since the (then) Australian Opposition Labor Leader leader accuses the incumbent government with quasi-elitist language ('[they have] treated .. public money as a political slush fund'), calling for a vote to oust them, yet steering well clear of explicit populist rhetoric. We include the text as a mid-point between the explicitly populist texts of Trump, Vlaams Belang and the Left Front on the one hand, and the explicitly anti-populist text of Obama on the other . The texts have the following notes:

- TRUMP – Excerpt from US Presidential inauguration speech, 20 Jan 2017. In one sentence, perhaps the 'textbook' populist position, juxtaposing the 'small group' in the 'nation's Capital' (i.e. the elites) profiting whilst 'the people' (homogenous, united) 'bear the cost'[11].
  - Hypothesis: We expect direct (zero-shot) calls to identify the text as populist, but it is an open question whether models will find this single sentence ticks all five dimensions of the extended prompt requirement.
- VLAAMS BELANG (VB) – Excerpt from the VB Membership Magazine, Oct 2008. This text arises from Pauwels (2011)'s study of Belgian parties, wherein they find that not only is VB 'often labelled as populist by both commentators and scholars' (p.98) but according to the study results, VB is 'by far the most populist of all the examined parties' (p.101).
  - Hypothesis: Strong anti-elite messaging should satisfy zero-shot prompt, satisfies three of five dimensions of the extended prompt, but is not explicit about the indivisible people and their will.
- LEFT FRONT ("Front de Gauche") – Excerpt from party manifesto, Jun 2012. The manifesto was extracted from the Manifesto Project[12] and has been translated from the original French to English using Anthropic's Claude 3.7 Sonnet model [13]. The Left Front represents one of the most prominent left-leaning populist movements today, scoring -47.925 on the rile score (extreme left) per the Manifest Project's own data.
  - Hypothesis: This extended text encapsulates well the multi-dimensional populism concept including strong phrasing around the 'the people', their will, their return to power through a 'citizen revolution' of the incumbent 'greed of a few'. Likely to pass both the zero-shot and extended prompt version.
- ALBANESE – Excerpt from campaign speech delivered 1st May, 2022 in Perth, Australia[14]. Anthony Albanese was at that time the leader of the federal opposition Labor party of Australia. The text was selected as a potentially challenging case for automated methods since in isolation, it has some flavour of anti-elite sentiment – accusing the incumbent leadership as shirking responsibility, and mis-using public money for their own interests – but a human expert would be unlikely to strongly identify this text as populist since there is little contrast of 'the people' with the 'elites' and his call for an anti-corruption commission supports the rule of law and democratic institutions in general.
  - Hypothesis: Mixed outcomes. Zero-shot outputs could fall either side of populist or not, but unlikely strongly so. However, should fail the extended, multi-dimensional prompt, with models unable to find convincing evidence across all aspects of populism required.
- OBAMA – Excerpt from US Presidential inauguration speech, 21 Jan 2013[15]. This text was delivered by Obama near the top of his speech. It affirms strongly the founding democratic principles

of the United States Constitution, it uses inclusive language for the peoples of the country and notes protections provided each of them under the constitution and the rule of law.

– Hypothesis: Neither the zero-shot or extended prompts should return any positive identification of populism. The text is anti-populist from start to finish.

Models

Models were included to present a spectrum of capabilities and varieties. From each provider, a strong 'reasoning' model was included, i.e. one which undertakes chain-of-thought reasoning by design (Wei et al., 2022) with reasoning effort or length set to 'high', a small/fast model, i.e. one which is intended for simpler tasks emphasising low latency and low cost, and an intermediate model, typically the strong model without reasoning, or an equivalent recommended model from the provider[16]. All calls were made via API with default API parameter settings. It is important to note that models are trained only up to a certain cut-off date. These dates ranged from the earliest in Oct 2023 (OpenAI's gpt-4o) to the most recent in Jan 2025 (Google's Gemini 2.5). As such, all texts were within the training context window of each model, bringing in the potential (or advantage), that the model may bring prior training world knowledge to bear on the task at hand (e.g. the model may recognise an excerpt is from a broadly known populist party or figure and so, be biased by the 'reputation' of the source in its adjudication of the given text). Whilst the researcher could attempt to redact identifying characteristics from the text, the reality is that today's models will likely recognise the surrounding text anyway (see below), hence, this consideration is more of a source of potential bias to be assessed through validation against ground-truth rather than something that can be completely avoided.

Findings

Figure 7 presents results across the 90 model calls. Results are presented to facilitate ease of comparison: across prompt strategies (zero-shot, then extended prompt: main panels), across model families (groupings within panels), and across model strengths within each family (stronger to weaker models are presented left to right within each family). We also provide a majority outcome for the zero-shot and extended prompt variants, weighting models equally.

From this exercise we draw the following insights:

1. **Labelling with transformers is very flexible.** It is hard not to overstress this point, though many practitioners may take for granted the new 'chat-based' world of AI, the ability to take the full text, of varying lengths, and even languages (all models considered in this demonstration are multilingual), and obtain valid labels in fractions of a second demonstrates much of the remarkable opportunity these methods bring to the table.
2. **Labelling with transformers is very cost effective.** Model calls were of the order of USD 0.01-0.02 for the most expensive (strongest) models (e.g. Claude 3.7 Sonnet, OpenAI's o3(high)) and fractions of a cent for the weaker models. Together, doing an ensemble call over the nine models cost on the range USD 0.05 to USD 0.10. Human labelling would take around 30s of reading time (e.g. Left Front, Obama), and perhaps 30s of judgement time, in all around 60s, at a typical data labelling rate of USD 25/hour, or USD 0.40/minute, we have an approx. 4 to 8x cost saving. However, this is a lower bound, given that one would rarely call three 'reasoning' models in practice, so the real cost saving is likely on the order of 12 to 24x.[17]
3. **Zero-shot prompting comes with model risks.** The most striking pattern across the two panels is the uniformity of findings under the zero-shot scenario. Here, the three populist texts (Trump, VB, Left Front) were universally found to be populist by each of the nine models, Albanese's tricky case leaned populist, whilst Obama's anti-populist case leaned strongly anti-populist. Whilst on face value this seems reasonable, the Albanese results are troubling given our prior hypothesis that Albanese's text should not be confused as populist despite the inherent government critique. Without defining terms, the models must leverage their own (hidden, 'black-box') understanding of the term "populist", which poses a control risk to any study design.

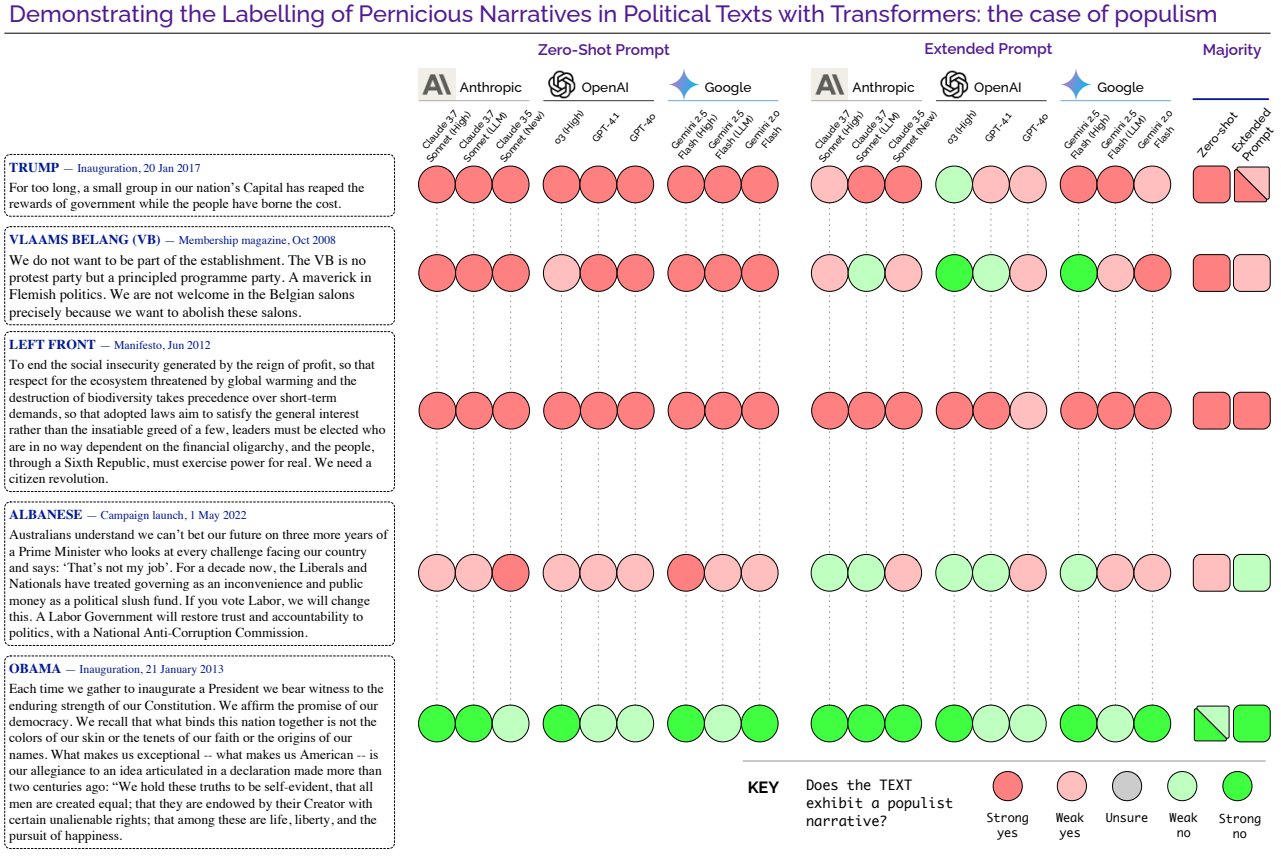

Figure 7. **Results for the labelling of populism in the small political texts corpus.** *Each column indicates the model response, colour coded according to the Key (bottom): more red colours indicate a finding of populism, more green colours indicate a finding against populism. Results are grouped according to model family (Anthropic, OpenAI, Google) with stronger to weaker variants of each family presented left to right within a family, and prompt in use (Zero-shot prompt: first 9 cols; Extended prompt: last 9 cols). The final two cols give a majority rating under each prompt..*

4. **Extended prompting is more robust, but results need careful interpretation.** Extended prompting has confirmed Obama's anti-populism, and moved the dial towards the same for Albanese, especially so in five of the six stronger model calls. However, we also see that positive populism labels are harder to obtain *in general*, with only the Left Front universally receiving the populist tag. Aligning with our hypotheses, Trump's short, sharp text, and VB's less explicitly five-dimensional excerpt both receive mixed reviews. We return to this point below.
5. **Ensembling models can recover signal from noise.** The pattern of the final columns in the figure, where we take the majority view from across the nine models deliver an almost exact imprint of our grounded pre-trial expectations for each text. Ensembling is well known in machine-learning to be a helpful technique for distinguishing signal from noise, and the same appears true from our demonstration in the world of transformer-based labelling.

We close this exercise with one final observation around transformer behaviour. Recall that our two prompts (Fig. 6) both asked the model to provide a label *and* a brief *justification* for the model response. These justifications are instructive but need to be taken with a grain of salt.[18] We focus on the anomalous cases, specifically the differential ways that models handled the VB text under the extended prompt (where most variation in outcome was observed). The three strongest models, across each provider gave the following:

- Anthropic :: Claude 3.7 Sonnet (High): "weak yes" ('Only demonstrates anti-elitism, missing other required populist dimensions.')
- OpenAI :: o3 (High): "strong no" ('Only anti-elite language; no unified people or general will.')
- Google :: Gemini 2.5 Flash (High): "strong no" ('Lacks explicit indivisible people, general will, and people-centrism.')

As expected, we can see that with the extended definition, each model struggled to endorse populism due to the requirement that all five dimensions should be present. It would be wrong therefore to conclude that the stronger models have made a 'mistake' here, instead, it is crucial to understand that as models have grown in capability, they have also grown in their ability to more closely follow the task given to them. We can contrast these explanations with those of the weaker models. Google's Gemini 2.0 Flash commented, "Anti-establishment, anti-elite sentiment is very strong" on the way to a 'strong yes' to populism. Whilst OpenAI's GPT-4o labelled 'weak yes', saying, 'Anti-elitism and people-centrism are evident,' focussing on what was evident, not what was missing.

4.2. Tracking Narratives of Immigration with Transformers

In our second demonstration, we use the novel *paired-completion* approach detailed in Angus and O'Neill (2024) to analyse texts related to immigration. Our aim is to provide an example of what can be done with paired-completion by carrying out a small longitudinal study on Australian Parliamentary Hansard during a particularly turmultuous time for the issue of asylum seeker policy in Australia.

Approach

We follow the method of paired-completion as provided in the reference. Paired-completion requires two key inputs: 1) a set of opposing 'seed' statements which define the opposing framings on the topic under study; and 2) a corpus of texts to apply the method to. On the first, OpenAI's GPT-4 (gpt-4-turbo) model was used to generate five statements to represent two perspectives on asylum seekers in the Australian context.[19] This produced the following seed statements for use as priming texts in the paired-completion method:

Perspective 1: Humanitarian Approach:

1. Human Rights Focus: "We must remember that asylum seekers are escaping dire situations where their basic human rights are often violated; it's our duty to offer them sanctuary."
2. International Obligations: "As a signatory to the UN Refugee Convention, Australia has a legal and ethical obligation to protect those who arrive on our shores seeking refuge from persecution."

3. Moral Duty: "It is our moral obligation as a prosperous and stable nation to extend a helping hand to those in need, showing compassion beyond our borders."
4. Community Benefits: "Asylum seekers bring a wealth of diverse cultural perspectives and skills that can greatly benefit our society and economy, particularly in sectors with labor shortages."
5. Legal and Fair Processing: "We must ensure our asylum processing system is fair and transparent, adhering to international laws and respecting the dignity of each individual."

Perspective 2: Border Security Approach:

1. National Security: "It is crucial to maintain stringent border controls to prevent the entry of potential threats to our national security, ensuring the safety of all Australians."
2. Economic Concerns: "The economic burden of supporting a large number of asylum seekers could strain our public services and welfare systems, impacting the financial stability of our country."
3. Legal Immigration Channels: "We need to uphold the integrity of our legal immigration system; allowing people to bypass it by arriving irregularly sets a precedent that could undermine the entire framework."
4. Sovereignty and Control: "Australia must have the sovereign right to control its borders and decide who can enter and stay in our country, to preserve the order and effectiveness of our immigration policies."
5. Deterring Smuggling Operations: "Implementing strict asylum policies can act as a deterrent to human trafficking and smuggling operations, which often exploit vulnerable individuals seeking refuge."

To proceed, the method requires to interact with a chat-completion (generative) LLM from which completion log-probabilities are available.[20] Since many commercial models do not support this feature, the open-source model 'llama-2' from Meta was employed (Touvron et al., 2023), using the $k = 2$ variant of the paired-completion approach.[21] This means providing the model with two randomly selected seed sentences on a given perspective, concatenated together, as a primer, and then appending the text under study "as if" the model had generated the whole sequence itself (primer(s) plus completion). One can read off the log-probabilities (i.e. the likelihoods) of the completion text, given the priming text(s) as explained earlier. The 'diff' metric is then computed for each text, i.e. the difference in likelihoods between the text being a completion of perspective 1 vs. a completion of perspective 2. Each text thus receives a textual alignment score on the humanitarian – border-security dimension, with higher (positive) scores associating most strongly with the humanitarian framing, and lower (negative) scores associating with the border-security framing.

Corpus

To illustrate the method, we emphasise the longitudinal dimension whilst reducing complexity by focussing only on 7 Prime Ministers along with four minor party or independent prominent speakers on the topic over this period. In alphabetical order the selected speakers were:

- Andrew Wilkie, Independent (Prominent speaker);
- Anthony Norman Albanese, Australian Labor Party (PM 2022-);
- John Winston Howard, Liberal Party of Australia (PM 1996-2007);
- Julia Eileen Gillard, Australian Labor Party (PM 2010-2013);
- Kevin Michael Rudd, Australian Labor Party (PM 2007-2010, 2013);
- Malcolm Bligh Turnbull, Liberal Party of Australia (PM 2015-2018);
- Nick McKim, Australian Greens (Prominent speaker);
- Richard Di Natale, Australian Greens (Prominent speaker);
- Sarah Hanson-Young, Australian Greens (Prominent speaker);
- Scott John Morrison, Liberal Party of Australia (PM 2018-2022); and
- Tony John Abbott, Liberal Party of Australia (PM 2013-2015)

In terms of hypotheses, the debate in Australia during the boat arrival period was highly polarised between the coalition (liberal/national) 'stop the boats' policy framing and the greens and independents who articulated a strongly humanitarian critique of the coalition's approach. Labor notoriously tried

to pursue a more humanitarian approach under Rudd/Gillard, seeking to accommodate and/or process asylum seekers in a mix of onshore and (eventually wholly) offshore detention centres, which in the eyes of the Greens and independents created more humanitarian disasters. Ultimately, with the "success" of the strengthened (military run) turn-back policy of the coalition, Labor in recent years has been accused of effectively adopting the identical policy formulation they previously rejected in opposition. Taken together, we might expect to see:

- Hypothesis 1: Polarisation between the Greens/Independents and the Liberal party on this topic;
- Hypothesis 2: Labor's policy position shifting from somewhat humanitarian to strongly border security based in recent years;

Hansard speeches for the Australian House of Representatives (Australia's lower, legislative, house in its bicameral system) were obtained from the Open Australia project API [22] covering Jan 2006 to Mar 2023[23], a period of particular interest in Australian politics due to the outsized increase in asylum seeker boat arrivals (typically from waters North of Australia) over this period. By illustration, just one boat arrived in 2003, carrying 53 asylum seekers, by 2008 these numbers rose to 161 people on seven boats, increasing rapidly to 6,555 arrivals on 134 boats in 2011 and an astonishing 20,587 people on 300 boats by 2013[24]. Asylum seeker policy of the major parties was a key concern during this period with the tragic loss of life at sea from failed crossings, coupled with images and stories of immigration detention (including children), intermingling with the wars and political hardships being fled by the occupants of these boats, weighing heavily on the Australian psyche.

Following the paired-completion method, speeches (texts) were first selected into the study through classical strict (literal) keyword matching.[25] In addition, speeches were restricted to those having lengths of between 150 and 1200 characters to ensure speeches were long enough for accurate analysis (i.e. not interjections), and not too long that the speech may cover multiple topics. Together, these steps resulted in a corpus comprising 2,856 speeches related to the asylum seeker issue from the 11 party leaders.

Findings

Figure 8 presents longitudinal results of the diff scores arising for the 11 speakers in the corpus. Time was dynamically binned based on the pooled count of speeches in the corpus to adaptively focus on the rhythm of the intensity of debate. Markers are included for a speaker if, for that time-period, they had at least three speeches from which to draw inference. Speaker diff average scores are plotted, with 90% confidence intervals (bootstrapped) provided for reference.

We draw the following insights with respect to the paired completion (P-C) analysis:

1. **P-C is well suited to identifying debate polarisation** – as a method which, by construction, exploits the differential textual alignment between one framing and another, by placing speakers on a Humanitarian—Border Security axis, the method is well suited to debate polarisation analysis. In this case, P-C strongly finds two 'camps' emerge in the debate, with all PMs drawn from the lower score, 'border-security' camp, with clear day-light to the higher score, 'humanitarian' camp of the independents and greens. Thus, confirming Hypothesis 1.
2. **P-C can pick up significant shifts in framing** – since P-C assigns a real-valued score to every text, it is possible to bring statistical machinery to the results. Here, bootstrapped confidence intervals reveal that the gap between camps, or even speakers are significant from a statistical standpoint, thus P-C lends itself to stronger forms of hypothesis testing. Here, we can see that Albanese's trajectory charts a course which begins within the distribution of Independent/Greens speakers, but ends statistically different from them, and thus indistinguishable from the Liberal position of the Abott/Turnbull years, thus supporting Hypothesis 2;
3. **P-C reveals interesting cyclical dynamics** – an unexpected result that the method reveals is the way that the independent/greens show cyclical framing dynamics around elections. We can see that in the case of Hanson-Young, Wilkie, and Natale speeches just after an election are found to be most aligned with the humanitarian framing, but as the next election approaches, their speeches shift towards the centre. The same dynamics (but from the opposite direction) for leaders such as

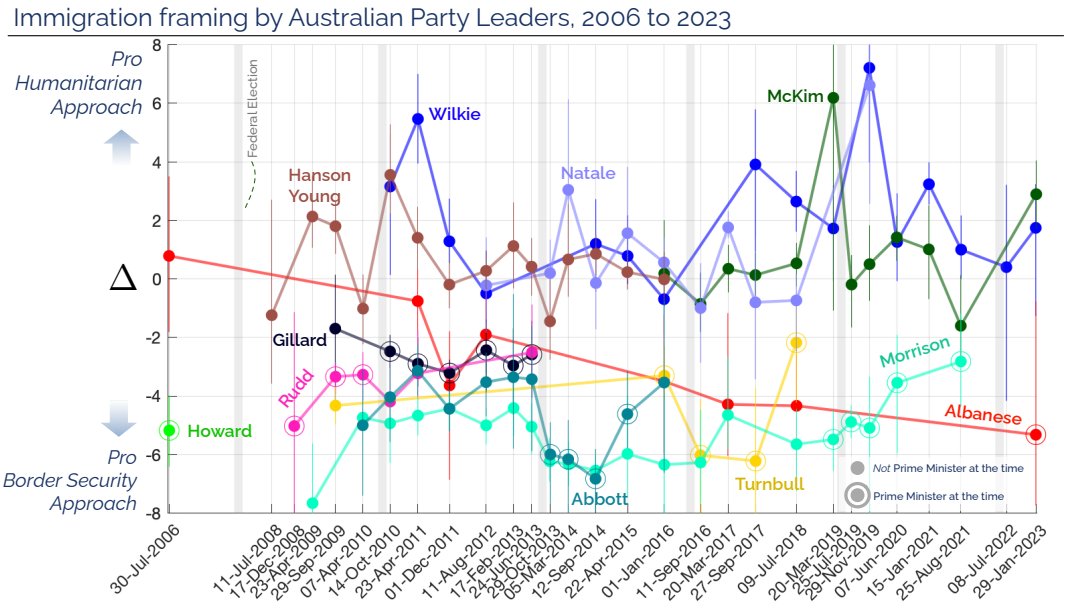

Figure 8. **Framing analysis with LLMs: 'paired-completion' (Angus and O'Neill, 2024) applied to asylum seeker debate framing in Australian Federal Parliament (2006-2023).** *Each line represents paired completion scores on a scale from negative (pro Border Security Approach) to positive (pro Humanitarian Approach) for the speaker shown. Rings around markers indicate that the speaker was Prime Minister at the time. Federal election dates are indicated with vertical grey lines for context..*

Rudd, Abbott, Turnbull and Morrison is evident; further analysis would need to be done to confirm these dynamics, but centre-shifts in political framing prior to elections are unsurprising.

As a complement to the time-series analysis, we can also inspect the most 'extreme' speeches in the corpus, as scored by paired-completion on this topic. The highest three scored speeches (i.e. most aligned with the Humanitarian Approach) unsurprisingly come from the Green/Independent speakers[26]:

Speaker: Sarah Hanson-Young (19 June 2014) (diff: 14.49)

> As one asylum seeker put it to me recently, 'I either decide to die on Manus Island or to die at home,' and that is the choice that this government is putting to these men, women and, indeed, children. So how many Iraqi asylum seekers have we deported in the last month back to a conflict zone? On what basis has it been determined that it is safe to do so? And, for the man in question, what is the government doing to ensure his safety is protected, now that he has been dumped back in the middle of a brutal, violent war, as the Prime Minister himself has identified? It is unconscionable that we are sending Iraqi asylum seekers back to these war zones. This coming Monday in this place we will vote to suspend the deportation of those people—

Speaker: Andrew Wilkie (21 Oct 2019) (diff: 14.42)

> The physical and mental suffering of these people cannot be understated. Indeed, under the medivac legislation, where people can only be transferred for urgent medical treatment to save their lives, over 130 applications have already been approved, with many more applications pending. That's a lot of gravely-ill people, and many of them sick because of the way they have been treated in Australia's offshore detention. Twelve people have died. In this context, it simply beggars belief that the government is ignoring the advice of Australian medical professionals and wants to repeal medivac.

Speaker: Sarah Hanson-Young (14 July 2014) (diff: 14.36)

> Rather than putting forward these ridiculous, fearmongering, insulting vilification-driven code of behaviour protocols, this government should be focussing on what is wrong with a policy that pushes an individual to the point of setting themselves on fire to take their own life so they do not have to be subjected to the awful tortures and persecution that they believe waits for them at home. It is beyond me why this government thinks it is important to continue to peddle the myths and fears around asylum seekers and refugees.

These speeches are clearly in the Humanitarian camp, with each text arguing strongly against the 'unconscionable', 'ridiculous', 'fearmongering' stance of the government of the day. As with prompt-based methods in our earlier demonstration, paired-completion can flexibly handle text of any length.

Turning to the other framing, the lowest three (i.e. most negative, most aligned with the Border Security Approach) scored speeches all come from the main party leaders:

Speaker: Scott John Morrison (2 Nov 2011) (diff: -23.42)

> Our amendment seeks to maintain those protections through one simple measure, which the government stubbornly refuse to adopt. Rather than have that matter accepted and the bill passed in this parliament, rather than return to the proven measures that they abolished, they have not only decided to continue their current policy of onshore processing; they have processed nobody offshore. We have had onshore processing ever since they abolished offshore processing. They would seek to further soften the policies by adopting those of the Greens, refusing, out of stubborn pride, to accept the coalition's amendment and return to policies that are proven and work, and, indeed, even pick up some of their own policies. Our amendment would enable the government to establish offshore processing again at Manus Island, which is their policy, yet they stubbornly refuse our amendment, as they have done consistently. Therefore, refusing to look at a simple, straightforward, modest amendment, the government deny themselves the opportunity to restore policies that are proven and work.

Speaker: Kevin Michael Rudd (26 Oct 2009) (diff: -19.99)

> In the period that the government has been in, we have had some 85 disruptions of planned smuggling adventures since September 2008, involving nearly 2,000 persons. That has been with the assistance of the Indonesian national police. The government has already returned almost 100 people because their asylum claims have been refused. Since September 2008, 15 people have been convicted in Australian courts and eight separate people are facing people-smuggling prosecutions. Further, currently before the Australian courts are 38 defendants in 14 separate matters, with three defendants in three separate matters involving organisers and facilitators. This is the practical action which goes to a practical and balanced policy on dealing with the scourge of people smugglers and the humanitarian plight of asylum seekers. It is a balanced approach, a mainstream approach; one which unapologetically also engages our friends and partners in the region, as previous Australian governments—were they to honestly reflect—have done as well.

Malcolm Bligh Turnbull (9 Nov 2016) (diff: -18.50)

> The reaction from the opposition benches is entirely consistent with the complacency their party showed when they were in government in 2008, 2009, 2010, 2011 and 2012 when, ignoring all of the evidence and all of the experience, they unpicked John Howard's tough and effective border protection policies. As a result—as a direct consequence—we saw 50,000 unauthorised arrivals, the opening of 17 new detention centres in Australia, an immigration blowout in cost of $11 billion, 8,000 children put into detention, the reopening of regional processing centres on Nauru and Manus and, most tragically of all, 1,200 deaths at sea of which we know.

Of interest, and showing the strength of recent transformer based models, whilst the first text (Morrison) argues strongly for a hardened, offshore, processing amendment to the incumbent's policy, the second and third texts (Rudd, Turnbull) both include language pointing to the humanitarian consequences (Rudd: 'the scourge of people smugglers and the humanitarian plight of asylum seekers'; Turnbull: 'most tragically of all, 1,200 deaths at sea of which we know') yet rightly the model understands that these notes sit within a wider semantic posture supporting tougher border policies in respect

of asylum seekers. This kind of nuanced understanding of human language is both critical for political narrative analysis, and unachievable in the pre-transformer technology era.

5. Recommendations: the Future of Tracking Public Narratives of Democracy

So how could these frontier techniques be applied to policy research questions relating to democratic resilience? We propose a number of starting points below.

5.1. Recommendations for data: removing blockers to policy-relevant discourse analysis at scale

Policymakers who wish to analyse the strength or fragility of our democracy by tracking narratives of democratic resilience, may rightly consider starting with the narratives spun by our own elected ministers, and the election platforms would-be ministers are touting. Together, these would provide a key input to narrative analysis for policy-relevant questions such as: is populist language on the rise? what fraction of election platforms in the recent election touted exclusionary nationalism? which jurisdictions are on the receiving end of political platforms that portray an 'Us vs. Them' narrative?

Unfortunately, through an analysis of the situation in several apparently healthy democracies, e.g. the United Kingdom (scoring 91 on Freedom House's 'Global Freedom Scores'[27]), Australia (95), and Finland (100), we find that, despite the promise of NLP/AI technology, answering these questions is still frustrated by failures in rather more basic requirements.

5.1.1. Data Recommendation 1: an official API for Parliamentary Hansard

Contrary to what are likely reasonable public expectations, the public discourse that one would *most* expect to be freely available under well documented and supported means, namely Parliamentary Hansard, is not, in fact, easy to come by for the kind of analysis proposed in the current paper. The lack of a single, comprehensive access point for Parliamentary discourse data appears to bedevil efforts across jurisdictions. Key to such access is support for an application programming interface (API). As opposed to browsing, where an individual clicks on links and reads the content of a webpage visited, an API enables scripted (i.e. computational) access to a database, which any down-stream, large-scale, and real-time analysis of public discourse requires.

Efforts vary by country but the pattern is the same. In the United Kingdom, after a historical period of access[28], the 'data.parliament.uk' initiative hosts a series of APIs to access various aspects of UK Parliamentary activity [29]. However, the API is still in 'beta' mode, and does not support access to Parliamentary Debate Hansard. The situation in Australia is worse. Whilst each jurisdiction has some user-facing web-portal for Hansard access[30] these require laborious human operation to download a large number of speeches. To our knowledge only the NSW Parliament has created an API access point to their Hansard, hosted via data.gov.au[31] which demonstrates that the task is surmountable. Likewise, Finland provides no official API to its Hansard, publishing instead, links to sessions of Parliamentary plenaries, including webcasts[32].

Into the breach, community funded, not-for-profit, projects have arisen to make parliamentary Hansard available. Typically, these efforts effectively reverse-engineer the task by visiting the user-facing (non-API) Hansard pages, downloading the page, and then transforming the information from the page back into a database to be finally made available on a community funded API portal. Naturally, these efforts, although well-intentioned and often achieving a high level of reliability, are not assured, and are not an 'official' source of political speech. For instance, the 'TheyWorkForYou'[33] initiative in the UK provides such a service, whilst in Australia, the 'Open Australia' initiative[34] tries to provide this access, but is typically not up to date, struggling to get the man-power to provide reliable service. Likewise in Finland, community scraped APIs are available with methods which unlock them, but these projects do not appear to be up-to-date[35].

Clearly, for any major policy-relevant democratic resilience analysis of Parliamentary discourse, or for the building of tools on top of Hansard to drive transparency, accountability, and public understanding, enabling civil society and academia to process the very speeches of our democracy is a critical starting point.

5.1.2. Data Recommendation 2: an official repository of political platforms

Relatedly, there are rarely official comprehensive sources for election literature, speeches, or materials. In Australia, whilst the National Library offers a portal[36], following links is again a laborious task and not designed for programmatic access. Digital election materials are available through the 'Trove' web archive site, however, links back to state Hansard are often dead due to stale or broken connections. In the UK, the Electoral Commission is tentatively exploring API development, with an 'alpha' version providing some information on candidates (name, party affiliation), though this API is itself delivered by a not-for-profit, 'Democracy Club', which hosts a more thorough API which can provide, among other things, candidate information including 'statements to voters' ("where collected")[37].

To be sure, there are promising international initiatives in this direction, however, their coverage is typically restricted to major political platforms, at the National level, befitting an international research initiative. For example, The Political Party Database Project[38] provides a smattering of official party documents for Australia, the United Kingdom, and Finland's major parties (e.g. in Australia: Labor, Liberal, National, Greens) over the last two decades. Similarly, the highly regarded Manifesto-Project[39] provides a small sub-set of national party platforms for each country, albeit with a much greater longitudinal coverage. Yet, the choice of which party platform to include, and even which document is stored, is down to academic discretion and resourcing. Without official support, these databases miss the small, tail-end party platforms, which arguably have the most potent information for policymakers wishing to track emergent narratives of democracy.

To facilitate open and active engagement by the research community with the platform and positional materials of the full range of democratic actors, programmatic (API) end-points of political platforms should be created in the ideal case, or at least a location enabling download with filtering should be made, so that large-scale textual research can be realised.

5.2. Recommendations for future research: towards real-time tracking of policy-relevant democratic discourse

We finish by sketching a potential agenda for further democratic narrative research that would support policymakers, academic researchers, and the wider community alike. Granted, significant data challenges exist, but laying these to one side, we consider a world where data were more available, enabling enriched analysis with AI-Transformer backed technologies considered earlier.

5.2.1. Research Recommendation 1: Hearing from the fringe – augmenting surveys with alternative data

LLM-backed methods offer a chance to overcome two key challenges with traditional, survey-based methods of tracking anti-democratic narratives at nation-scale. First, surveys are typically expensive to run, and therefore are undertaken at most annually, though often less frequently. Second, a common concern of those who run surveys of trust in government and institutions, is that, despite rigorous attempts, survey respondents are typically self-selecting. That is, if there is a growing group of people in a country who support anti-democratic narratives, they are also highly likely to be wary of institutional mechanisms in general, and so, unlikely to be willing to participate in surveys run by 'official' organisations. The UK Government, for example, acknowledge this source of bias explicitly in their 'Trust in Government Survey QMI' report (2024)[40]

> There may be a sampling bias, because the sampling frame comes from those who have already responded to a government survey; these people may differ to the general population in how much they trust the government and other public institutions.

However, alternative data sources abound in these countries, from online discussion-boards, to YouTube and Instagram content, to broad-based social media and alt- group websites. By converting these sources to text (admittedly, not an insignificant task in some cases), LLM-enabled framing analysis on dimensions of populism, 'Us vs. Them' and other anti-democratic narratives can be tracked. By building up a time-series of these markers, movements up and down in the concentration and strength of these narratives can be compared statistically to baseline levels. Significant movements could trigger more intensive survey or focus-group work, or deeper dive analysis of specific narratives and the basis of their claims. In this way, the 'fringe' of society who may be actively engaging in digital media and platforms, but who are already disengaged from government surveys and institutions, can be better understood in policy-relevant analysis of the state or health of democratic values in society at large.

5.2.2. Research Recommendation 2: Estimating rates of anti-democratic support with revealed preferences, the case of Australia

Australia's democracy is one of only a handful of free countries world-wide to enforce compulsory voting. With the measure's introduction in 1924, Australia will see 100 years of the practice through the election cycle of 2025. However, for similar selection issues discussed above, a significant challenge in Australia is estimating the *true* fraction of the population who hold more extreme anti-democratic views. Given that voter turn-out is very high across the country, and that voting is considered a private act, one can consider voter support for radical party positions as a more revealed preference indicator of support, than perhaps stated preference measures as considered above. Here, the challenge is to collect and analyse the thousands of political tracts from major and minor parties across all election cycles, at state and national level, and analyse these for narratives of democratic resilience and opposition. By applying vote-share to those party platforms that score strongly on textual alignment with particular narratives of democratic opposition, one can calculate, for the first time, a true level of revealed support for these positions. These data would go significantly towards answering the question of whether Australia really does have a 'hidden' underbelly of support for radical positions on democratic issues, or if instead, public posturing does not speak to privately held beliefs.

Whilst this research direction focusses on the case of Australia, there are over a dozen countries which share the compulsory voting approach, providing further contexts for the application of AI-transformer methodologies to track relevant narratives of democracy by jurisdiction.

5.3. Research Recommendation 3: Supporting accountability in our Parliamentary discourses: a public monitor of democratic discourse

One of the major challenges facing the public, journalists, and academics alike, is the challenge of identifying objectively that a speaker or outlet is saying something unusual (for them) or fitting a particular narrative of democracy. Without accurate, quantitative tools, we are left relying on 'opinion' pieces and expert talking heads to opine on whether a particular speaker's comments really were anti-democratic, populist, or divisive. Whilst expert commentary will always be important, what we lack is a way for the public to see insightful visualisations, driven by accurate discourse measurement, of how a particular discourse fits in to a spectrum of discourse at a point in time, or across time. Such a 'Narrative Observatory' like tool would be a major benefit to the democratic project. Speakers would be held immediately accountable for their words, policymakers, journalists and analysts would more easily be able to draw comparisons and make data-driven insights about the positions being taken on the floors of our Parliaments nationally. Fact-checking initiatives exemplify this kind of public good provision,

a narrative observatory would complement and empower our society to better understand and hold to account those who represent us.

6. Conclusion

Our starting point was that story-telling matters deeply to the way that we communicate and encode our culture and beliefs in society, and so, for policymakers concerned with democratic resilience, narrative analysis at scale is now an immensely valuable capability alongside traditional methodologies. We have seen that in social science, narratives can be analysed and tracked in a variety of ways, with issue-framing being one of the most common approaches. However, coding texts for the presence or absence of one or other framing has traditionally been a laborious process for experts and research assistants alike. And, human approaches are not without their own bias and reliability concerns. For this reason, we have considered how automated techniques arising from the computational sciences have been developed to analyse discourse in a variety of ways, including automation of the labelling task. Significantly we have seen that the most recent wave of technological advance, the wave that is sweeping every major discipline of the academy, the generative AI/LLM wave, holds enormous promise for application to tracking narratives with relatively low costs, low bias, and high reliability.

Notes

[1]For a review of survey based methods, see surveys such as Levi and Stoker (2000); Citrin and Stoker (2018), or projects such as the Australian Social Cohesian index (https://scanloninstitute.org.au/research/australian-cohesion-index), or the Social Cohesion Radar (Dragolov et al., 2016).

[2]In machine learning, an algorithm is either trained with, or without, labelled data. In the former, known as 'supervised' learning, the objective is to accurately infer the class (label) of a given record, given that the actual class is known ('ground-truth'). In the latter, known as 'unsupervised' learning, the objective is to identify patterns in the data, minimising some penalty which increases if similar records are not grouped together.

[3]Pauwels (2011)'s word list: [*deceit*; *treason*; *betray*; absurd*; arrogant*; promis*; promise*; capitul*; corrupt*; direct; elite*; establishm*; ruling*; caste; class; mafia; freedom of expression; undemocratic; particrat*; politic*; propaganda; referend*; regime*; shameless; shame*; admit; tradition*; people].

[4]See Ali and Hassan (2022) for a review of 37 computational approaches in this direction.

[5]For an earlier survey of text as data methods see Grimmer and Stewart (2013), who reach largely the same conclusion.

[6]Prompt used in the study: 'You will be provided with the text of a tweet published by a member of the US Congress. Where does this text stand on the 'left' to 'right' wing scale? Provide your response as a score between 0 and 100 where 0 means 'Extremely left' and 100 means 'Extremely right'. If the text does not have political content, set the score to "NA". You will only respond with a JSON object with the key Score. Do not provide explanations. Text of the tweet '

[7]For a flavour of what is possible in the broader areas of political science, Li et al. (2024) provides a recent white-paper review of for the application of LLMs to political science.

[8]For example, if one asks the most performant language model to date, Anthropic's Claude 3.5 Sonnet, 'How many i's in Kit-Kat?', it answers, 'The correct answer is that there are zero i's in Kit Kat.' This behaviour is common for all the major models alike.

[9]Included countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Malta, The Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom.

[10]Note: prior to the prompt, a general system prompt was provided the model and was constant across all calls, 'You are a helpful, knowledgeable, expert assistant. You are proficient in many tasks, such as coding, math, science, humanities, and more.'

[11]See: https://trumpwhitehouse.archives.gov/briefings-statements/the-inaugural-address/, April 2025.

[12]See (requires login): https://manifesto-project.wzb.eu//down/originals/31021_2012.pdf, April 2025.

[13]Specifically: `claude-3-7-sonnet-20250219`, prompt: 'Translate to English: {text}'. It is worth noting that each model tested in this demonstration could have handled undertaking the task in the original language (French), however, to reduce contamination due to their respective strengths at translation, pre-translation to each model's main training language (EN) was undertaken.

[14]See: https://electionspeeches.moadoph.gov.au/speeches/2022-anthony-albanese, April 2025

[15]See: https://obamawhitehouse.archives.gov/the-press-office/2013/01/21/inaugural-address-president-barack-obama, April 2025.

[16]Exact model details (reading left to right of Fig. 7): **Anthropic** – 'claude-3-7-sonnet-20250219-high' (thinking enabled, budget tokens 32768, cut-off Oct 2024), 'claude-3-7-sonnet-20250219' (LLM mode, no thinking enabled, cut-off Oct 2024), 'claude-3-5-sonnet-20241022' (cut-off Apr 2024); **OpenAI** – 'o3-high' ('o3-2025-04-16', reasoning effort: 'high', cut-off 1 Jun 2024), 'gpt-4.1' ('gpt-4.1-2025-04-14', cut-off 1 Jun 2024), 'gpt-4o' ('gpt-4o-2024-08-06', cut-off 1 Oct 2023); **Google/Gemini** – 'gemini-2.5-flash-preview-04-17' (thinking enabled, budget tokens 24k, cutoff Jan 2025), 'gemini-2.5-flash-preview-04-17' (LLM mode, no thinking, cutoff Jan 2025), 'gemini-2.0-flash' (cutoff June 2024).

[17]Noting that this simple analysis equally doesn't price development time for the model pipelines, nor equivalent time spent setting up human labellers.

[18]Recent work by Anthropic under the heading of model 'biology' has been able to show strong evidence that model explanations, and even internal chains of thought, do not always align with model outputs. See: 'On the Biology of a Large Language Model', published at https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-cot.

[19]Prompt 1: 'Provide two perspectives on asylum seekers in the Australian context. Name each perspective. For each perspective, provide five sentences that together represent that perspective.', followed by Prompt 2: 'For each perspective, turn the sentences into example statements that represent these views.'

[20]i.e. Typically via the 'echo' switch.

[21] Specifically, the 70B parameter variant, having a context length of 4,096 chars. Meta do not give a precise end-point for the training window, but note training on 'publicly available data'.
[22] See: http://data.openaustralia.org.au/.
[23] At the time of ingestion, these were the extent of data available through the project, as the engineering team were having trouble fixing their methods. See comments below on data availability.
[24] Figures from the Refugee Council of Australia, https://www.refugeecouncil.org.au/asylum-boats-statistics/, April 2025.
[25] Terms used for literal matches: [asylum seeker, seeking asylum, border protection, sovereign border, people smuggling, mandatory detention, refugee status, migration policy, stop the boat, turn-back, detention centre, maritime arrival, unauthorised boat, boat arrival, australian border force, offshore processing, offshore detention, regional processing.]
[26] Note: some speeches end with a hyphen, indicating an interjection.
[27] Freedom House, https://freedomhouse.org/countries/freedom-world/scores.
[28] See the historical Hansard Millbank home for such data at https://api.parliament.uk/historic-hansard/api.
[29] See https://explore.data.parliament.uk/index.html.
[30] e.g. Federal Parliament provides a search front-end: https://www.aph.gov.au/Parliamentary_Business/Hansard/ and Victorian Parliament the same: https://www.parliament.vic.gov.au/parliamentary-activity/hansard/hansard-debate/, neither offer programmatic (API), or bulk down-load access directly.
[31] See: https://parliament-api-docs.readthedocs.io/en/latest/new-south-wales/.
[32] See https://verkkolahetys.eduskunta.fi/fi/ (Finnish/Swedish).
[33] See: https://www.theyworkforyou.com/api/.
[34] See: https://www.openaustralia.org.au/api/.
[35] See for example, https://github.com/rOpenGov/finpar.
[36] See: https://www.nla.gov.au/collections/what-we-collect/ephemera/federal-election-campaigns.
[37] See https://democracyclub.org.uk/data_apis/voting_information_api/ and links therein.
[38] See https://www.politicalpartydb.org/countries/.
[39] See https://manifesto-project.wzb.eu/.
[40] Office for National Statistics (ONS), released 1 March 2024, ONS website, methodology, Trust in Government Survey QMI. See {urlhttps://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/trustingovernmentsurveyqmi

Acknowledgments. I am grateful to fellow members of the Resilient Democracy Research and Data Network (Australia) for discussions which informed the development of this paper. In particular, from the network, Prof Nicholas Biddle (Australian National University) who provided useful comments on earlier drafts of this work. All errors and omissions are my own. I am also indebted to SoDa Labs AI Research Lead, Lachlan O'Neill for their technical assistance in obtaining the data from the Open Australia API end-point for the Hansard demonstration and running the paired-completion estimates with the Llama 2 LLM. Their expertise in handling these steps was invaluable for this portion of the research. All errors and omissions are my own.

Competing Interests. 'None'.

Data Availability Statement. Data and code to replicate the results of the experiments of this paper are contained in the public GitHub repository https://github.com/specialistgeneralist/narratives-of-democratic-resilience, release https://doi.org/10.5281/zenodo.15493085 .

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. S.D.A. is the sole author of this manuscript and is responsible for the conceptualization, methodology, formal analysis, investigation, data curation, writing (original draft preparation and review/editing), visualization, and project administration.

Supplementary Material. NA

References

Algan, Y., Guriev, S., Papaioannou, E., and Passari, E. (2017). The european trust crisis and the rise of populism. *Brookings Papers on Economic Activity*, pages 309–382.

Ali, M. and Hassan, N. (2022). A survey of computational framing analysis approaches. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 9335–9348.

Anantharama, N., Angus, S. D., and O'Neill, L. (2022). Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3551–3564.

Angus, S. D. and O'Neill, L. (2024). Paired completion: quantifying issue-framing at scale with LLMs. SoDa Laboratories Working Paper Series 2024-02, Monash University, SoDa Laboratories.

Ash, E., Gauthier, G., and Widmer, P. (2024). Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32:115–132.

Ayala, O. and Béchard, P. (2024). Reducing hallucination in structured outputs via retrieval-augmented generation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 6:228–238.

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, 43:465–482.

Biddle, N., Fischer, A., Angus, S. D., Ercan, S., Grömping, M., and Gray, M. (2025). Democratic resilience: Moving from theoretical frameworks to a practical measurement agenda. Ssrn working paper, SSRN. Available at SSRN: https://ssrn.com/abstract=5202987.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Kohd, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., and Wang, W. (2021). On the opportunities and risks of foundation models.

Bonikowski, B., Luo, Y., and Stuhler, O. (2022). Politics as usual? measuring populism, nationalism, and authoritarianism in u.s. presidential campaigns (1952–2020) with neural language models. *Sociological Methods and Research*, 51:1721–1787.

Bonilla, T. and Mo, C. H. (2019). The evolution of human trafficking messaging in the united states and its effect on public opinion. *Journal of Public Policy*, 39:201–234.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. In *Proceedings of New Directions in Analyzing Text as Data Workshop*. Part of the New Directions in Analyzing Text as Data Workshop.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Chong, D. and Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science*, 10:103–126.

Choudhary, T. (2024). Political bias in large language models: A comparative analysis of chatgpt-4, perplexity, google gemini, and claude. *IEEE Access*.

Citrin, J. and Stoker, L. (2018). Political trust in a cynical age. *Annual Review of Political Science*, 21:49–70.

Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2256–2262.

Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2:643–653.

Cocco, J. D. and Monechi, B. (2022). How populist are parties? measuring degrees of populism in party manifestos using supervised machine learning. *Political Analysis*, 30:311–327.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dragolov, G., Ignácz, Z. S., Lorenz, J., Delhey, J., Boehnke, K., and Unzicker, K. (2016). *Social Cohesion in the Western World: What Holds Societies Together: Insights from the Social Cohesion Radar*. Springer.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43:51–58.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., and Sofsky, F. (2023). Friend or foe? exploring the implications of large language models on the science system. *AI and Society*, 40:447–459.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3570–3580.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57:535–74.

Gidron, N. and Hall, P. A. (2020). Populism as a problem of social integration. *Comparative Political Studies*, 53:1027–1059.

Goyal, N. and Howlett, M. (2021). "Measuring the Mix" of Policy Responses to COVID-19: Comparative Policy Analysis Using Topic Modelling. *Journal of Comparative Policy Analysis: Research and Practice*, 23(2):250–261. Publisher: Routledge.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297. Edition: 2017/01/04 Publisher: Cambridge University Press.

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Guo, X., Ma, W., and Vosoughi, S. (2022). Capturing topic framing via masked language modeling. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6811–6825.

Hagen, J. and Lysne, O. (2016). Protecting the digitized society-the challenge of balancing surveillance and privacy. *The Cyber Defense Review*, 1:75–90.

Hager, A. and Hilbig, H. (2020). Does public opinion affect political speech? *American Journal of Political Science*, 64:921–937.

Hetherington, M. J. and Rudolph, T. J. (2015). *Why Washington won't work : polarization, political trust, and the governing crisis*. The University of Chicago Press.

Holloway, J. and Manwaring, R. (2023). How well does 'resilience' apply to democracy? a systematic review. *Contemporary Politics*, 29:68–92.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8. Issue: 1.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Jennings, W. and John, P. (2009). The dynamics of political attention: Public opinion and the queen's speech in the united kingdom. *American Journal of Political Science*, 53:838–854.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84:905–949.

Lam, M. S., Teoh, J., Landay, J., Heer, J., and Bernstein, M. S. (2024). Concept induction: Analyzing unstructured text with high-level concepts using lloom.

Levi, M. and Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science*, 3:475–508.

Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J. A., Shen, B., Qian, A., Chen, W., Xue, Z., Sun, L., He, L., Chen, H., Ding, K., Du, Z., Mu, F., Pei, J., Zhao, J., Swayamdipta, S., Neiswanger, W., Wei, H., Hu, X., Zhu, S., Chen, T., Lu, Y., Shi, Y., Qin, L., Fu, T., Tu, Z., Yang, Y., Yoo, J., Zhang, J., Rossi, R., Zhan, L., Zhao, L., Ferrara, E., Liu, Y., Huang, F., Zhang, X., Rothenberg, L., Ji, S., Yu, P. S., Zhao, Y., and Dong, Y. (2024). Political-llm: Large language models in political science.

Liu, S., Guo, L., Mays, K., Betke, M., and Wijaya, D. T. (2019a). Detecting frames in news headlines and its application to analyzing news framing trends surrounding u.s. gun violence. *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 504–514.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). RoBERTa: A Robustly Optimized BERT Pretraining Approach. _eprint: 1907.11692.

Luitse, D. and Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of ai. *Big Data and Society*, 8.

Martino, L. D. (2020). The spectrum of listening. *Routledge Handbook of Public Diplomacy*, pages 21–29.

Mauro, A. D., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65:122–135.

McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

McCoy, J., Rahman, T., and Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62:16–42.

McCoy, J. and Somer, M. (2019). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *Annals of the American Academy of Political and Social Science*, 681:234–271.

Meijers, M. J. and Zaslove, A. (2021). Measuring populism in political parties: Appraisal of a new approach. *Comparative Political Studies*, 54:372–407.

Mendelsohn, J., Budak, C., and Jurgens, D. (2021). Modeling framing in immigration discourse on social media. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 2219–2263.

Mens, G. L. and Gallego, A. (2025). Positioning political texts with large language models by asking and averaging. *Political Analysis*, 0:1–9.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mitchell, M. (2023). Ai's challenge of understanding the world. *Science*, 382.

Mitchell, M. and Krakauer, D. C. (2022). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Moore, A. D. (2011). Privacy, security, and government surveillance: Wikileaks and the new accountability. *Public Affairs Quarterly*, 25:141–156.

Morstatter, F., Wu, L., Yavanoglu, U., Corman, S. R., and Liu, H. (2018). Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1:1–18.

Mudde, C. (2004). The populist zeitgeist. *Government and Opposition*, 39:541–563.

Mylonas, H. and Tudor, M. (2021). Nationalism: What we know and what we still need to know. *Annual Review of Political Science*, 24:109–132.

Nelson, L. K., Burk, D., Knudsen, M., and McCall, L. (2021). The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research*, 50(1):202–237.

Norris, P. and Inglehart, R. (2019). Cultural backlash: Trump, brexit, and authoritarian populism. *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*, pages 1–540.

O'Neill, L., Anantharama, N., Borgohain, S., and Angus, S. D. (2023). Models teaching models: Improving model accuracy with slingshot learning. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3233–3247, Dubrovnik, Croatia. Association for Computational Linguistics.

Pauwels, T. (2011). Measuring populism: A quantitative text analysis of party literature in belgium. *Journal of Elections, Public Opinion and Parties*, 21:97–119.

Penney, J. W. (2016). Chilling effects: Online surveillance and wikipedia use. *Technology Law Journal*, 31:117–182.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pit, P., Ma, X., Conway, M., Chen, Q., Bailey, J., Pit, H., Keo, P., Diep, W., and Jiang, Y.-G. (2024). Whose side are you on? investigating the political stance of large language models.

Plisiecki, H., Lenartowicz, P., Flakus, M., and Pokropek, A. (2025). High risk of political bias in black box emotion inference models. *Scientific Reports 2025 15:1*, 15:1–11.

Pokrass, M. (2024). Introducing structured outputs in the api | openai.

Polletta, F. and Callahan, J. (2017). Deep stories, nostalgia narratives, and fake news: Storytelling in the trump era. *American Journal of Cultural Sociology*, 5:392–408.

RaffelColin, ShazeerNoam, RobertsAdam, LeeKatherine, NarangSharan, MatenaMichael, ZhouYanqi, LiWei, and J., L. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21:1–67.

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., and Bavel, J. J. V. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 121:e2308950121.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rettenberger, L., Reischl, M., and Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science 2025 8:2*, 8:1–17.

Richards, N. M. (2013). The dangers of surveillance. *Harvard Law Review*, 126:1934–1965.

Rooduijn, M., Pirro, A. L., Halikiopoulou, D., Froio, C., Kessel, S. V., Lange, S. L. D., Mudde, C., and Taggart, P. (2024). The populist: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (eiqcc). *British Journal of Political Science*, 54:969–978.

Schmidgall, S. and Moor, M. (2025). Agentrxiv: Towards collaborative autonomous research.

Shiller, R. J. (2019). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, Princeton, NJ.

Smith, G. R., Bello, C., Bialic-Murphy, L., Clark, E., Delavaux, C. S., de Lauriere, C. F., van den Hoogen, J., Lauber, T., Ma, H., Maynard, D. S., Mirman, M., Lidong, M., Rebindaine, D., Reek, J. E., Werden, L. K., Wu, Z., Yang, G., Zhao, Q., Zohner, C. M., and Crowther, T. W. (2024). Ten simple rules for using large language models in science, version 1.0. *PLOS Computational Biology*, 20:e1011767.

Smith-Carrier, T. and Lawlor, A. (2016). Realising our (neoliberal) potential? a critical discourse analysis of the poverty reduction strategy in ontario, canada. *Critical Social Policy*, 37:105–127.

Snow, D. A. and Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *International Social Movement Research*, 1:197–217.

Somers, M. R. (1994). The narrative constitution of identity: A relational and network approach. *Theory and Society*, 23:605–649.

Strengthening Democracy Taskforce (2024). Strengthening australian democracy: A practical agenda for democratic resilience.

Terkildsen, N. and Schnell, F. (1997). How media frames move public opinion: An analysis of the women's movement. *Politcal Research Quarterly*, 50:879–900.

Thapa, S., Shiwakoti, S., Shah, S. B., Adhikari, S., Veeramani, H., Nasim, M., and Naseem, U. (2025). Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15:1–30.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P.,

Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature 2024 625:7995*, 625:476–482.

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vidgen, B. and Yasseri, T. (2020a). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17:66–78.

Vidgen, B. and Yasseri, T. (2020b). What, when and where of petitions submitted to the UK government during a time of chaos. *Policy Sciences*, 53(3):535–557.

Wang, S., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2021). Want to reduce labeling cost? gpt-3 can help. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pages 4195–4205.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Quoc, E. H. C., Le, V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models chain-of-thought prompting. In *36th Conference on Neural Information Processing Systems (NeurIPS)*.

Willard, B. T. and Louf, R. (2023). Efficient guided generation for large language models.

Young, L. and Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231. Publisher: Routledge.

Zhang, Y., Shah, D., Pevehouse, J., and Valenzuela, S. (2023). Reactive and asymmetric communication flows: Social media discourse and partisan news framing in the wake of mass shootings. *International Journal of Press/Politics*, 28:837–861.