TruthSplit: Operationalizing Conditional Validity in Arguments Through Multi-Perspective Reasoning

Benjamin Stieger[*1] **Maximilian Terberger**[*1] **Thomas Huber**[2] **Christina Niklaus**[2]
University of St. Gallen
[1]{benjamin.stieger,maximilian.terberger}@student.unisg.ch
[2]{thomas.huber,christina.niklaus}@unisg.ch

Abstract

We present TruthSplit, an interactive system for multi-perspective argument analysis. Existing argumentation tools typically analyze properties of the argument itself, such as structure, quality, stance, or persuasiveness, while leaving perspective-specific background knowledge implicit. TruthSplit addresses this gap by supporting an exploratory analysis of how the same claim can lead to different conclusions when interpreted through worldview-specific values, assumptions, and conceptual definitions. We refer to this perspective-dependent analysis as conditional validity. Given an input argumentative text, TruthSplit extracts claims and premises, applies a three-layer natural language inference (NLI) approach to assess both logical and worldview-specific normative consistency, and conditions large language model (LLM) reasoning on structured worldview profiles that encode core values and decision principles. The system then generates perspective-specific interpretations, identifies value conflicts and assumption gaps, and visualizes divergence through interactive analytical interfaces.

1 Introduction

In an era of increasing polarization (Iyengar et al., 2019), understanding disagreement has become crucial. Consider universal basic income (UBI): one person argues "I oppose UBI because it undermines individual responsibility," while another maintains "I support UBI as it can provide individuals with the financial security to engage in sustainable practices, which is good for the environment." Both may have examined the same data yet reach opposing conclusions (Mercier and Sperber, 2011). Traditional argumentation tools focus on identifying the structure (Palau and Moens, 2009) or quality of arguments (Wachsmuth et al., 2017, 2024), but fail to address: *How can the same argument be valid from multiple, seemingly incompatible perspectives?*

This is a difficult problem because disagreement often stems not from flawed reasoning, but from deeper structural differences: different assumptions about how the world works (Tetlock, 2005; Kuhn, 1962), distinct value priorities (Berlin, 1969; Haidt, 2012), and varying definitions of contested societal concepts like "freedom" or "justice" (Rawls, 1993). In the previous example, one might prioritize individual liberty ("freedom" as freedom *to* move—positive freedom), while others prioritize collective security ("freedom" as freedom *from* threats—negative freedom) (Berlin, 1969; Snyder, 2018). Both positions may be logically consistent within their ideological frameworks, yet incompatible under a single standard.

This points to *conditional validity*: the idea that the validity of a conclusion depends not only on the premises from which it is drawn, but also on the values, assumptions, and conceptual definitions through which it is interpreted (Rawls, 1993; Lougheed, 2021). TruthSplit treats this as an exploratory system concept rather than a formal theory of logical validity, using explicit worldview profiles to compare how different perspectives evaluate the same claim-premise relations. Existing tools fail because they assume universal correctness. Argumentation systems classify arguments as "correct" or "fallacious" (Goffredo et al., 2023) but cannot capture how an argument might be valid from one perspective yet invalid from another. Prior work has focused on identifying argumentative structures (Stab and Gurevych, 2014), improving the persuasiveness of arguments (Xia et al., 2022), and teaching argumentation skills (Wambsganss et al., 2021), while computational ideology analysis typically classifies perspectives (Bamman et al., 2013; Hardisty et al., 2010) rather than generating perspective-aware reasoning.

We present TruthSplit, a comparative reason-

[*]Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 648–659

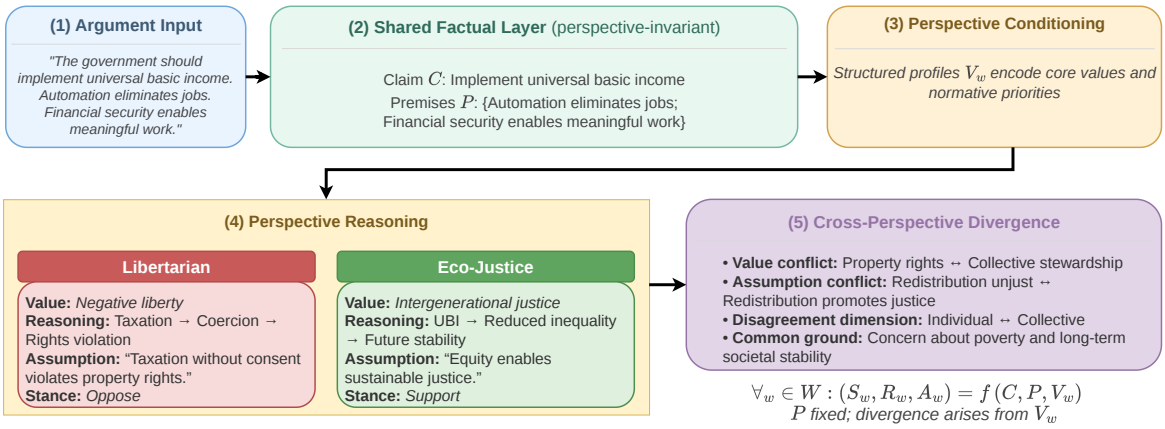

Figure 1: Conceptual overview of TruthSplit. An input argument is decomposed into a claim C and premises P, forming a shared factual layer that is invariant across perspectives. Structured worldview profiles V_w condition reasoning generation, producing a reasoning layer R_w, an assumption layer A_w, and a stance S_w for each worldview. Cross-worldview divergence analysis identifies value conflicts, assumption conflicts, disagreement dimensions, and common ground. Disagreement emerges from differences in normative priors rather than factual inconsistency.

ing platform that reveals how different worldviews interpret the same arguments[1]. TruthSplit combines a multi-stage analysis pipeline with a structured knowledge base encoding six ideological perspectives (Libertarian, Religious-Conservative, Ecological Social-Democrat, Populist-Nationalist, Communist, Neo-Reactionary) with their core values, conceptual definitions, and decision frameworks. The TruthSplit tool includes three-layer consistency testing based on Natural Language Inference (NLI), semantic concept linking, and an LLM-based worldview analysis. It supports both local, privacy-preserving models and larger proprietary LLMs available through external providers via API access to improve the quality of the analysis. To help users examine perspective-dependent disagreement, the system provides interactive visualizations that reveal worldview differences, reasoning trajectories, and patterns of divergence. Figure 1 illustrates the central mechanism of TruthSplit: holding premises fixed while varying perspective priors to generate and analyze conditional reasoning. TruthSplit is intended for educators, students, researchers, and analysts seeking to explore perspective-dependent reasoning in political and ethical discourse.

Our contributions are fourfold: (i) Operationalizing conditional validity: we enable users to perform a systematic analysis of arguments across multiple perspectives rather than assigning a single correctness label. (ii) Perspective-conditioned reasoning: TruthSplit generates explicit reasoning chains conditioned on structured ideological profiles. (iii) Interactive exploration of divergence: we provide a visual and analytical tool that identifies value conflicts, assumption gaps, and points of disagreement. (iv) Structured worldview knowledge representation: we encode different perspectives as extensible computational profiles rather than informal prompt descriptions.

2 Related Work

TruthSplit builds on three strands of work: computational argumentation, interactive argumentation systems, and computational analyses of ideology and perspective. Existing approaches provide methods for extracting argumentative structures, assessing argument quality, supporting users in argument construction, or identifying ideological leaning in text. TruthSplit differs in that it uses structured worldview-specific background knowledge to condition the analysis of arguments. In that way, it allows users to compare how the same claim may be interpreted, evaluated, and explained differently across perspectives.

Computational Argumentation NLP research has made significant advances in argument mining tasks (Lawrence and Reed, 2019), including the extraction of claims (Levy et al., 2014), the

[1]The code is licensed under the MIT license and available at https://github.com/unisg-ics-dsnlp/truthsplit. A video demonstration is available at: https://www.youtube.com/watch?v=xYMCOpU8x18.

identification of premises (Habernal and Gurevych, 2017), and argument structure parsing (Stab and Gurevych, 2014). NLI has also been used to evaluate argument quality (Al-Khatib et al., 2016), and recent work has leveraged LLMs for generating arguments (Eskandari Miandoab and Sarathy, 2024; Huber and Niklaus, 2025b) and analyzing argument quality (Wachsmuth et al., 2024; Mirzakhmedova et al., 2024). These approaches provide important methods for identifying, structuring, and evaluating argumentative content. However, prior work generally does not incorporate structured background information about different perspectives into argument analysis, and therefore does not systematically explain how the same claim may yield different conclusions under different worldview-specific values, assumptions, and conceptual definitions.

Interactive Argumentation Systems Interactive systems support users in navigating, improving, or learning argumentation. CoArgue (Liu et al., 2023), for example, extracts and summarizes argumentative elements in Community-Based Question Answering platforms to foster participation. Other systems visualize argument structures (Huber and Niklaus, 2025a), support persuasive writing (Xia et al., 2022), or teach argumentation skills through adaptive learning environments such as ArgueTutor (Wambsganss et al., 2021) and AL (Wambsganss et al., 2020). TruthSplit similarly provides an interactive interface, but focuses on comparative analysis across ideological perspectives rather than participation support, persuasiveness, or argumentation training.

Worldview and Perspective Analysis Research in computational social science has explored ideological and perspective analysis, including methods for identifying political leanings in text (Bamman et al., 2013), analyzing ideological discourse (Diermeier et al., 2012), and characterizing perspective differences (Hardisty et al., 2010). Such approaches are useful for detecting or categorizing perspectives, but typically do not generate perspective-conditioned reasoning about specific arguments. TruthSplit complements this work by representing worldviews as structured profiles of values, assumptions, definitions, and decision principles, and by using these profiles to support multi-perspective argument analysis.

3 System Design & Architecture

The objective of TruthSplit is to reveal ideological disagreement in arguments by operationalizing conditional validity, i.e., the idea that a conclusion may be valid given a specific set of normative priors, values, and conceptual definitions, even if it is rejected under alternative normative assumptions. Existing argumentation systems typically assess structural quality (Wachsmuth et al., 2016) or classify ideological stance (Iyyer et al., 2014). In contrast, TruthSplit addresses a different computational objective: analyzing how validity depends on perspective-specific priors. This is more similar to how humans interpret different texts, as every human holds some views, such as being more conservative or liberal leaning. For instance, under a strictly Libertarian worldview the claim that *'All taxes are theft'* is rational, whereas a more socially-oriented individual would not support such a claim, and consider it to be too extreme to be considered rational. To achieve this, the system integrates structured perspective representations with logical inference and LLM-based reasoning in a unified comparative analysis framework.

Figure 2 provides an overview of the system architecture. TruthSplit consists of two main components: (i) a structured worldview knowledge base (see Section 3.1), and (ii) a multi-stage analysis pipeline that decomposes arguments, evaluates logical consistency, conditions reasoning on worldview profiles, and performs cross-perspective divergence analysis (see Section 3.2). The system is designed as an analytical and educational tool, i.e. it does not attempt to resolve disagreement, but to make explicit how different conclusions emerge from distinct normative starting points.

3.1 Worldview Knowledge Base

We construct different worldviews based on established political philosophy literature and validate them through expert participants. We discuss the choice of worldviews further in Section 5. To cover a broad range of views and values, the worldviews are Libertarian, Religious-Conservative, Ecological Social-Democrat, Populist-Nationalist, Communist, and Neo-Reactionary, but these can be extended.[2] Each profile contains weighted core values, key definitions (i.e., how the worldview

[2]These profiles represent key positions across the political spectrum rather than an exhaustive catalog. The format is extensible, allowing further worldviews to be added manually.

interprets contested terms), assumed principles, decision frameworks, and factor scores for 16 ideological dimensions. These factor scores make the worldviews computational, enabling quantitative comparison and systematic analysis across different perspectives. Since all profiles follow the same JSON schema, new or customized worldviews can be added without modifying the core analysis pipeline. An example profile structure is included in Appendix A.

3.2 Analysis Pipeline

TruthSplit processes inputs through six sequential stages that transform raw arguments into comparative worldview analyses. Input can be provided via direct text entry, file upload, or dynamically fetched news articles from News API[3]. Figure 2 shows an example claim being processed by all stages of the pipeline.

3.2.1 Argument Extraction

Natural language arguments are often implicit—claims embedded in narrative, premises unstated, assumptions hidden. TruthSplit decomposes text into *claims* (central assertions), *premises* (explicit evidence), and *assumptions* (unstated connecting beliefs). For an example, see box (2) in Figure 1. TruthSplit offers two extraction modes: (i) *local extraction* using a sequence classification model (∼75–80% accuracy, fully privacy-preserving), and (ii) *cloud-based extraction* via LLM with structured prompts and JSON schema validation (∼95%+ accuracy). See Appendix A for model specifications and Appendix B for prompts.

3.2.2 Consistency Testing

To assess whether the argument is conditionally valid, we assess the logical coherence on three layers.[4] We use a model pre-trained on the MultiNLI dataset (Williams et al., 2018). See Appendix A.2 for model details.

Layer 1 - Premise-Claim Logic Do premises logically support the conclusion, independent of value judgments? Arguments failing here have fundamental logical flaws. A text could for instance contain multiple, disconnected arguments, or contain unsupported claims. An initial analysis aims to uncover badly structured arguments that do not provide value when analyzed further due to those fundamental flaws.

Layer 2 - Perspective-Internal Does the claim align with a perspective's principles? A claim may be internally consistent within one framework but contradictory within another. The check makes use of the six worldview profiles (Section 3.1) and analyzes the text through the respective perspective to confirm it is internally consistent with the worldview.

Layer 3 - Cross-Perspective Do perspectives agree or disagree on this claim? High agreement suggests shared values; high disagreement indicates fundamental conflicts requiring analysis. The individual worldview profiles discussed in Section 3.1 are used to provide insight into where and why different perspectives agree or disagree.

Continuing our UBI example (Figure 2): Layer 1 tests whether "provides a financial safety net" logically supports "would reduce poverty" (high entailment). Layer 2 reveals divergence: for Social Democrats, UBI aligns with their principle of collective welfare (high consistency), while for Libertarians, mandatory wealth redistribution conflicts with property rights (low consistency). Layer 3 confirms this fundamental disagreement across worldviews.

3.2.3 Concept Linking and Worldview Analysis

Certain concepts are viewed differently in certain worldviews (see box (4) in Figure 1). For instance, the concept of *freedom* means "absence of coercion" from a Libertarian perspective, but "capability to flourish" from Social Democrats' point of view. Each worldview contains concepts with a brief definition or description of what they mean in the context of the specific worldview. For a given input, we identify the most relevant concepts belonging to each worldview by calculating the cosine similarity between worldview-specific concept definition and the extracted claims. Model details and thresholds for the similarity calculation are included in Appendix A.

For worldview reasoning, we prompt LLMs with structured context including consistency scores, linked concepts, and full worldview profiles (Section 3.1). Using structured prompts (Appendix B) and JSON schema validation, the

[3] https://newsapi.org/

[4] Note that these layers are a system design choice rather than a standard NLI taxonomy: Layer 1 isolates perspective-independent inferential support, Layer 2 conditions consistency on one worldview profile, and Layer 3 compares consistency scores across profiles to identify disagreement.

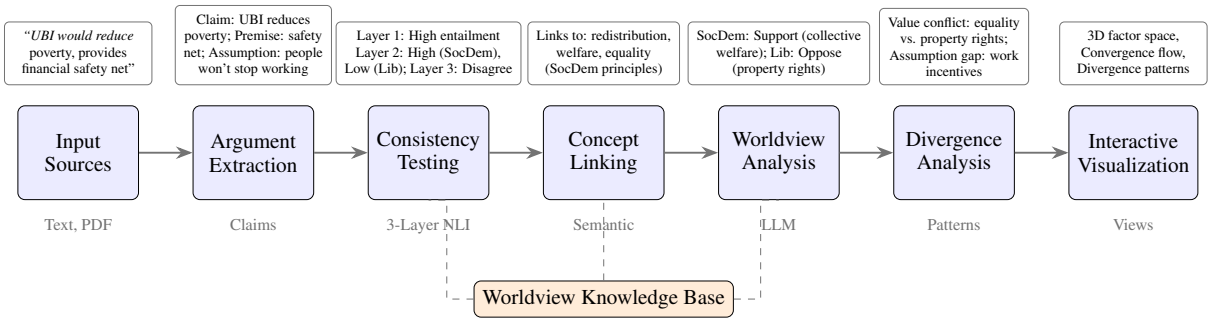

Figure 2: Overview of the TruthSplit pipeline.

LLM generates: interpretation, position (support/oppose/conditional), reasoning chain, key assumptions, concerns, and alternative approaches. Figure 3 shows a partial view of the output format.

Figure 3: Worldview reasoning showing positions for a UBI claim, with expandable assumptions and concerns. This figure shows only part of the complete reasoning output; the full analysis also includes logical chain, key assumptions, detailed concerns, alternative approach, and the metrics.

3.2.4 Divergence Analysis

TruthSplit then provides an analysis into *how* different perspectives disagree (see box (5) in Figure 1). This is done on the levels of value conflicts (e.g., liberty vs. equality), which are caused by different prioritization issues and concepts, definition differences where the same concept is interpreted in different ways (e.g., "rights" as negative vs. positive claims), assumption gaps caused by a reliance on different empirical or normative assumptions, and priority differences where the overall values are shared or interpreted similarly but ranked differently (e.g., one might prioritize the need of the many over the good of the few, even though both agree on the overall underlying definitions). For the analysis we use LLMs with structured prompts to classify disagreement types and assess severity. Figure 8 provides an example output.

The *Convergence Flow* component complements this divergence analysis, and traces the reasoning chain from core values through beliefs, interpretation, and conclusion—showing at each step whether perspectives converge or diverge. This reveals where disagreement originates: do the different perspectives share values but differ in interpretation, or do they diverge from the start? The convergence flow is shown in Appendix Figure 6.

4 User Interface

The frontend is a multi-page React application with two main components: (i) an *analysis dashboard* for claim extraction, worldview comparison, and result visualization, and (ii) an interactive *worldview chatbot* for conversing with specific perspectives.

Analysis Dashboard Figure 4 illustrates the typical user interaction flow through the system. Users begin by providing input text through direct entry, PDF upload, or selecting from curated news articles. Next, the system extracts claims automatically. The user then chooses up to three worldviews for comparison of the claims and configures the analysis mode (external LLM provider, or local privacy-preserving model).

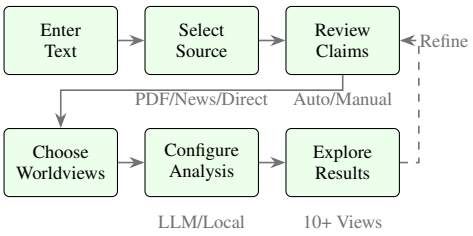

Figure 4: User interaction flow from text input through claim extraction, worldview selection, and analysis configuration to interactive result exploration with 10+ visualizations.

The results interface provides multiple visualization tabs: Overview with summary metrics (Figure 5), Detailed Analysis with per-worldview reasoning (Figure 3), Divergence Analysis showing disagreement patterns (Figure 8), Convergence

Flow tracing reasoning chains (Appendix C Figure 6), Worldview Space for factor positioning (Appendix C Figure 7), and Worldview Chat for interactive dialogue.

Worldview Chatbot TruthSplit offers an interactive chatbot that embodies a selected worldview, enabling users to engage in dialogue and probe a perspective's reasoning in real-time. Users can ask follow-up questions, challenge positions, or explore how a worldview would respond to new scenarios—deepening understanding through conversational exploration.

The chatbot is based on an LLM, which can be configured through the config files, and assumes the position of a particular perspective profile, e.g. a conservative view. The user can then discuss the analysis results with the chatbot operating under a certain perspective, or have a regular conversation with it, again while the chatbot is assuming a certain perspective (i.e. a smalltalk setting).

The complete worldview is included in the dialogue context (core values, definitions, principles, red lines) and the model is prompted to respond consistently from that perspective. We include an evaluation of quality of the dialogues, as well as experiments to assess the stability, in Section 5.

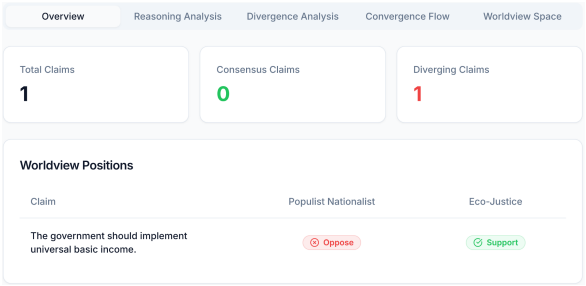

Figure 5: Results dashboard: analysis summary, claims, worldviews, and position distribution.

5 Evaluation

We conducted a mixed-methods evaluation to assess whether TruthSplit makes ideological disagreement more intelligible by exposing conditional validity across perspective frameworks. The evaluation focused on (i) usability and interpretability of comparative reasoning outputs, (ii) quality of structured worldview representations, and (iii) robustness across different LLMs. The focus of our evaluation is on the usefulness and interpretability of the analyses, not on the correctness of the overall reasoning or divergence explanations. We leave this for future work.

Study Design The study consisted of two tiers. An expert evaluation (n=3 philosophy students) assessed the coherence of each of the six worldviews, reasoning quality, and the validity of factor encodings across 16 ideological dimensions per worldview. A broader evaluation (n=52 participants from political science, computer science, psychology, business, and related academic backgrounds) focused on usability, accessibility, and clarity of the divergence analysis. All participants were recruited from academic contexts and self-identified as politically centrist.

Usability and Accessibility Table 1 summarizes usability results. Experts rated TruthSplit as easy to use (4.67/5) with high visual appeal (5.00/5). The divergence analysis was accessible to non-experts (4.07/5), which indicates that the comparative outputs can be understood without philosophical training. Option selection (i.e., local vs. cloud analysis modes) was less clearly understood (3.47/5), suggesting room for interface refinement. The quality of the claim extraction was rated moderately positively (6.67/10).

Metric	Expert	Broader
Ease of use (1–5)	4.67	–
Visual appeal (1–5)	5.00	4.36
Divergence understanding (1–5)	4.33	4.07
Options understanding (1–5)	–	3.47
Claim extraction quality (1–10)	–	6.67

Table 1: Usability and accessibility ratings.

Worldview Representation Validation Experts evaluated 96 factor–worldview combinations (16 dimensions across 6 worldviews). Implemented factor scores showed moderate positive correlation with expert importance ratings (r=0.33), with strongest alignment for Religious-Conservative and Ecological Social-Democrat worldviews (r=0.46). Inter-expert variance was substantial with a mean standard deviation of 2.01 and 39% of cases with disagreement ≥ 5 on a 1 to 10 scale. This suggests inherent ambiguity in quantifying ideological dimensions.

Qualitative feedback consistently described Ecological Social-Democrat, Libertarian, and Communist representations as coherent and reflective of real-world political discourse. The divergence analysis was highlighted as particularly valuable for understanding why worldviews disagree.

LLM Robustness The outputs that were generated using different families of LLMs (Claude, GPT, Gemini, Grok, DeepSeek) showed no significant quality differences according to both expert and broader participants. This suggests that TruthSplit's structured prompting approach standardizes worldview-conditioned reasoning across models.

6 Discussion and Future Work

The results suggest that TruthSplit succeeds in making ideological disagreement more interpretable by foregrounding conditional validity. Rather than resolving disagreement, the system reframes it as a product of differing premises and value hierarchies, supporting its positioning as an analytical and educational tool rather than a decision-making system. The study also highlights design tensions: 2/3 expert participants struggled with worldview boundaries and definitions—a common challenge in computational approaches to ideology (Grimmer and Stewart, 2022). The exploratory nature and limited sample size mean findings should be interpreted as indicative rather than conclusive. Future evaluations with larger and more diverse participant groups will be necessary.

Several promising directions emerge: (i) a custom worldview builder allowing users to define their own ideological profiles, enabling analysis through perspectives not currently represented; (ii) educational integration with curriculum materials for teaching critical thinking and perspective-taking in civic education and journalism programs; and (iii) multimodal input support to extend analysis to audio and video sources.

7 Conclusion

TruthSplit addresses an important challenge in computational argument analysis: analyzing how the validity of an argument depends on worldview-specific normative priors. Rather than assigning a single correctness judgment, it reveals how conclusions emerge from distinct configurations of values, assumptions, and conceptual definitions. By integrating argument mining, NLI-based consistency modeling, structured worldview representations, LLM-conditioned reasoning, and interactive visualization, TruthSplit operationalizes conditional validity as a computational objective. It enables systematic comparison of how different worldviews interpret the same argument while keeping the underlying factual premises fixed. Our evaluation indicates that users can meaningfully interpret and trace these differences: the divergence analysis was accessible to non-experts, and expert validation demonstrated moderate alignment between implemented worldview encodings and domain intuition. These findings suggest that structured worldview conditioning provides a viable foundation for multi-perspective argument analysis. TruthSplit illustrates how AI systems can move beyond binary correctness judgments towards a transparent analysis of normative disagreement. We hope this framework supports future research on perspective-aware NLP and fosters more informed engagement with ideological differences.

8 Acknowledgements

TruthSplit is part of the M-Rational project, which is funded by the Swiss National Science Foundation, under grant number 10004303. The SNF project URL is: `https://data.snf.ch/grants/grant/10004303`. This system was developed as part of a student project at the University of St. Gallen.

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Isaiah Berlin. 1969. *Four Essays on Liberty*. Oxford University Press, Oxford.

Daniel Diermeier, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. Language and ideology in congress. *British Journal of Political Science*, 42(1):31–55.

Kaveh Eskandari Miandoab and Vasanth Sarathy. 2024. "let's argue both sides": Argument generation can force small models to utilize previously inaccessible reasoning capabilities. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 269–283, Miami, Florida, USA. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Justin Grimmer and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, Princeton, NJ.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books, New York, NY.

Eric Hardisty, Jordan Boyd-Graber, and Philip Resnik. 2010. Modeling perspective using Adaptor Grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 284–292, Cambridge, MA. Association for Computational Linguistics.

Thomas Huber and Christina Niklaus. 2025a. ARTIST: A learning support system for fostering students' argumentative writing skills. In *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pages 10–13, Hanoi, Vietnam. Association for Computational Linguistics.

Thomas Huber and Christina Niklaus. 2025b. CLEAR: A comprehensive linguistic evaluation of argument rewriting by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19548–19568, Suzhou, China. Association for Computational Linguistics.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Chengzhong Liu, Shixu Zhou, Dingdong Liu, Junze Li, Zeyu Huang, and Xiaojuan Ma. 2023. Coargue: Fostering lurkers' contribution to collective arguments in community-based qa platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kirk Lougheed. 2021. The epistemic benefits of worldview disagreement. *Social Epistemology*, 35(1):85–98.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

John Rawls. 1993. *Political Liberalism*. Columbia University Press, New York, NY.

Timothy Snyder. 2018. *The Road to Unfreedom: Russia, Europe, America*. Tim Duggan Books, New York, NY.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Philip E. Tetlock. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, NJ.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

A Technical Details

A.1 Architecture and Stack

TruthSplit follows a client-server architecture: a React/TypeScript frontend with Three.js (3D visualizations) and Plotly (charts), a Python Flask REST API backend, and containerized deployment via Docker. The system uses a unified API layer (OpenRouter) supporting multiple LLM providers (Claude, GPT, Gemini, Grok, DeepSeek).

A.2 NLP Models

- **Local extraction and NLI**: facebook/bart-large-mnli (zero-shot classification, CPU-only, 2–5 sec/text, ∼75–80% accuracy)
- **Cloud extraction**: OpenRouter API with structured JSON schema validation (∼95%+ accuracy)
- **Semantic embeddings**: all-MiniLM-L6-v2 (Sentence-BERT) for concept linking via cosine similarity

A.3 Consistency Testing Thresholds

Three-layer NLI testing categorizes cross-worldview agreement by variance in consistency scores: $\leq$0.05 (strong agreement), $\leq$0.15 (moderate agreement), $\leq$0.25 (moderate disagreement), $>$0.25 (strong disagreement).

A.4 Worldview Profile Schema

Each worldview is encoded as a JSON profile containing: `core_values` (name, weight, description), `key_definitions` (term, definition), `assumed_principles` (name, priority, description), `warrants` (name, confidence, description), `red_lines`, `evidence_preferences`, `decision_rule` (type, weight), and `extensions` including `factor_scores` across 16 ideological dimensions.

A.5 Performance

Complete analysis time (typical: 1 claim, 2 worldviews): Local NLP + Cloud LLM: ∼12–25 seconds; Cloud NLP + Cloud LLM: ∼13–30 seconds.

B Prompt Templates

All prompts use structured system/user prompt pairs with JSON schema validation enforcing required fields, enumerated values, minimum content lengths, and nested object structures. Below is the worldview analysis prompt as a representative example; claim extraction, divergence analysis, convergence flow, and chatbot prompts follow similar patterns.

System Prompt: "You are an expert in political philosophy and worldview analysis. Use the provided structured data to give precise, nuanced analysis of how different worldviews interpret claims. Respond with structured JSON."

User Prompt (abbreviated): Given a claim, consistency scores (3 layers), worldview profile (core values, definitions), and linked concepts, the LLM generates: (1) interpretation (2–3 sentences), (2) position (support/oppose/conditional), (3) reasoning chain (3–5 logical steps from core values to conclusion), (4) detailed reasoning (3–5 sentences),

(5) key assumptions (3–5 with related core values), (6) concerns (3–5 with explanations), and (7) alternative approach aligned with worldview values.

C Screenshots

Figures 6, 7, and 8 illustrate the TruthSplit interface through representative screenshots.

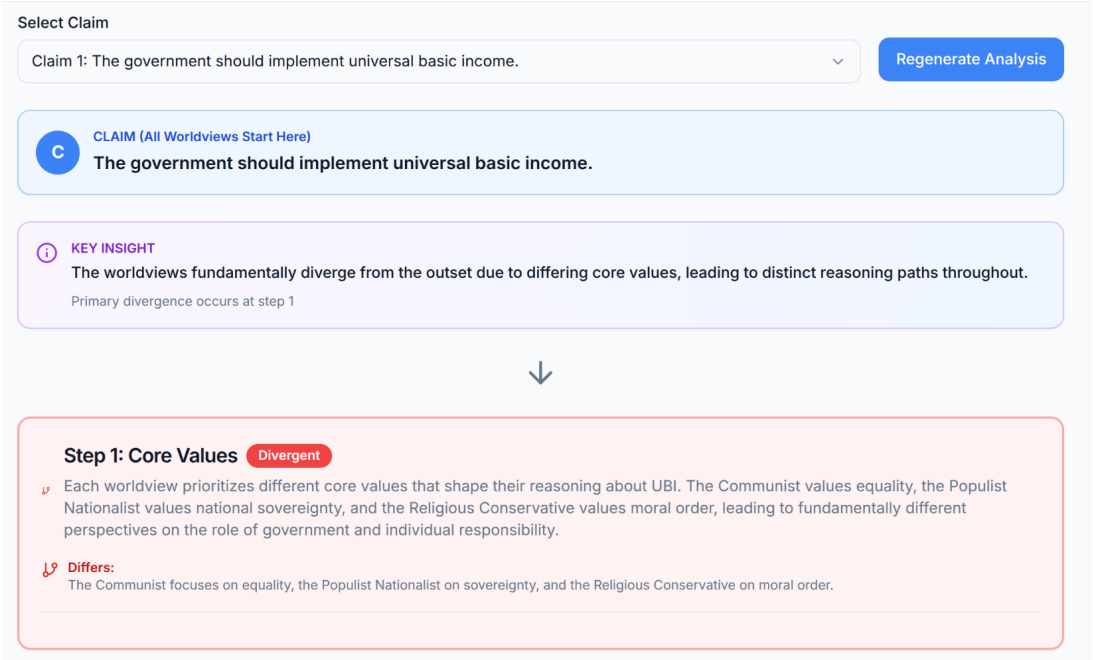

Figure 6: Excerpt for divergence flow analysis.

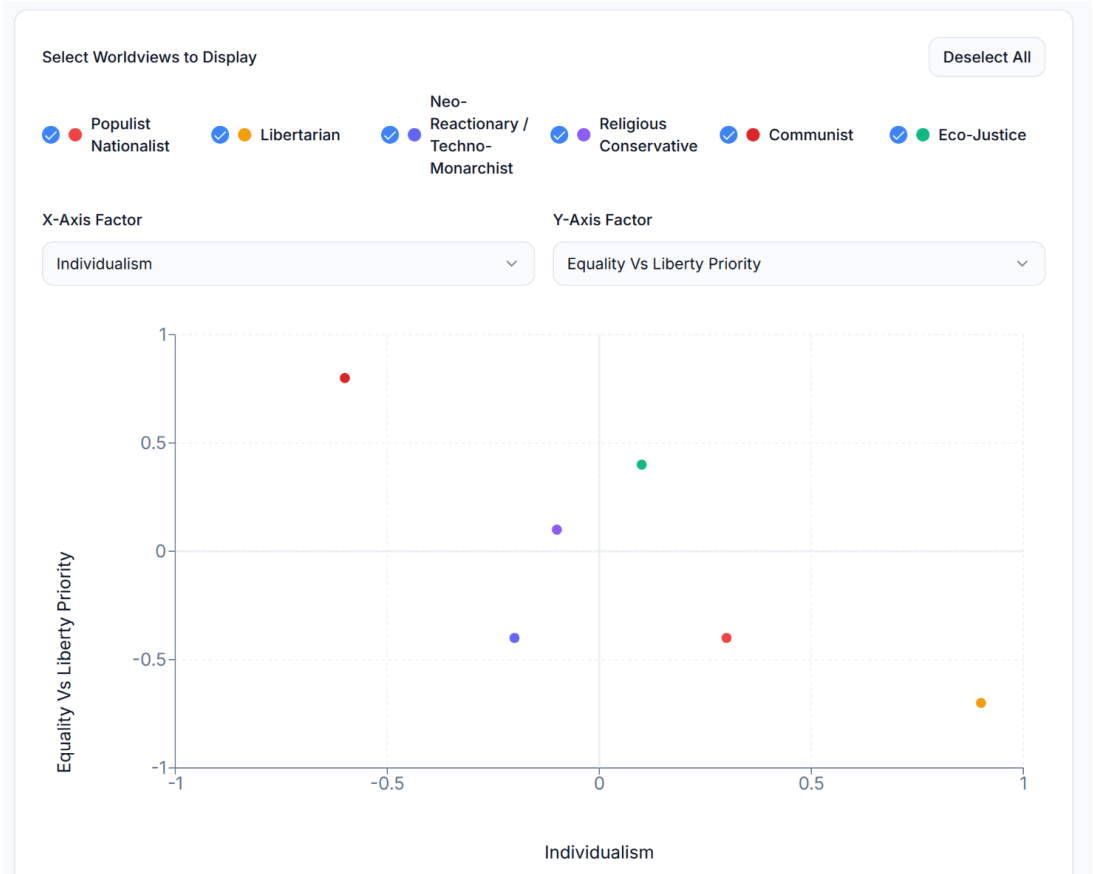

Figure 7: Screenshot of Worldview Positioning.

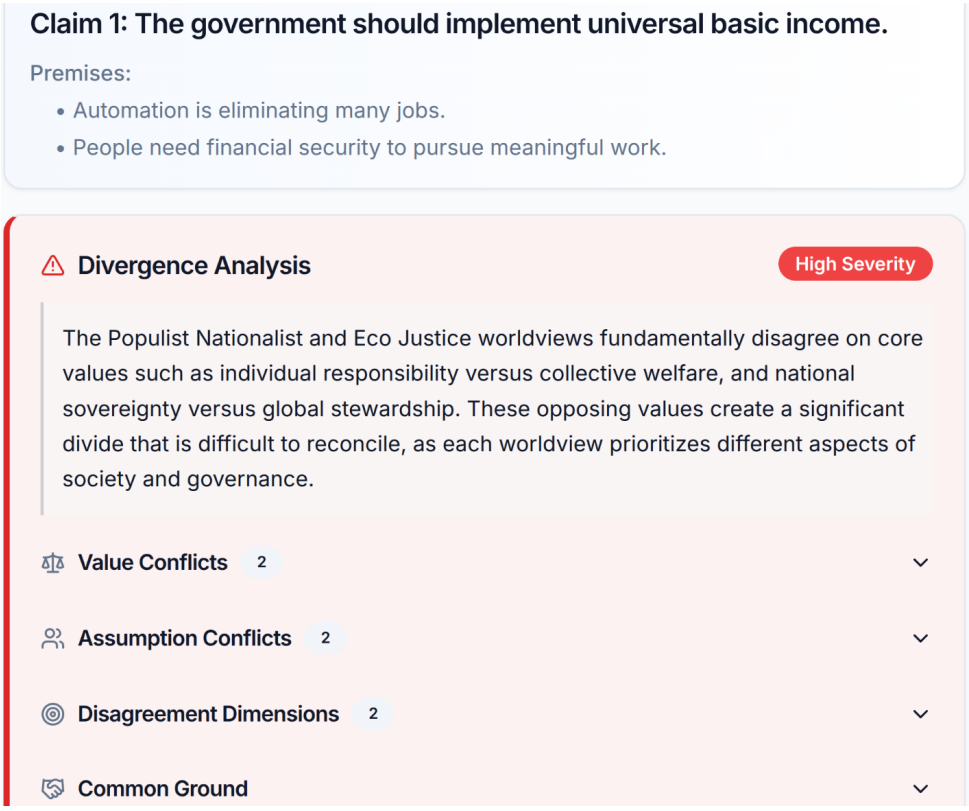

Figure 8: Divergence Analysis showing value conflicts, assumption differences, and disagreement severity between worldviews.