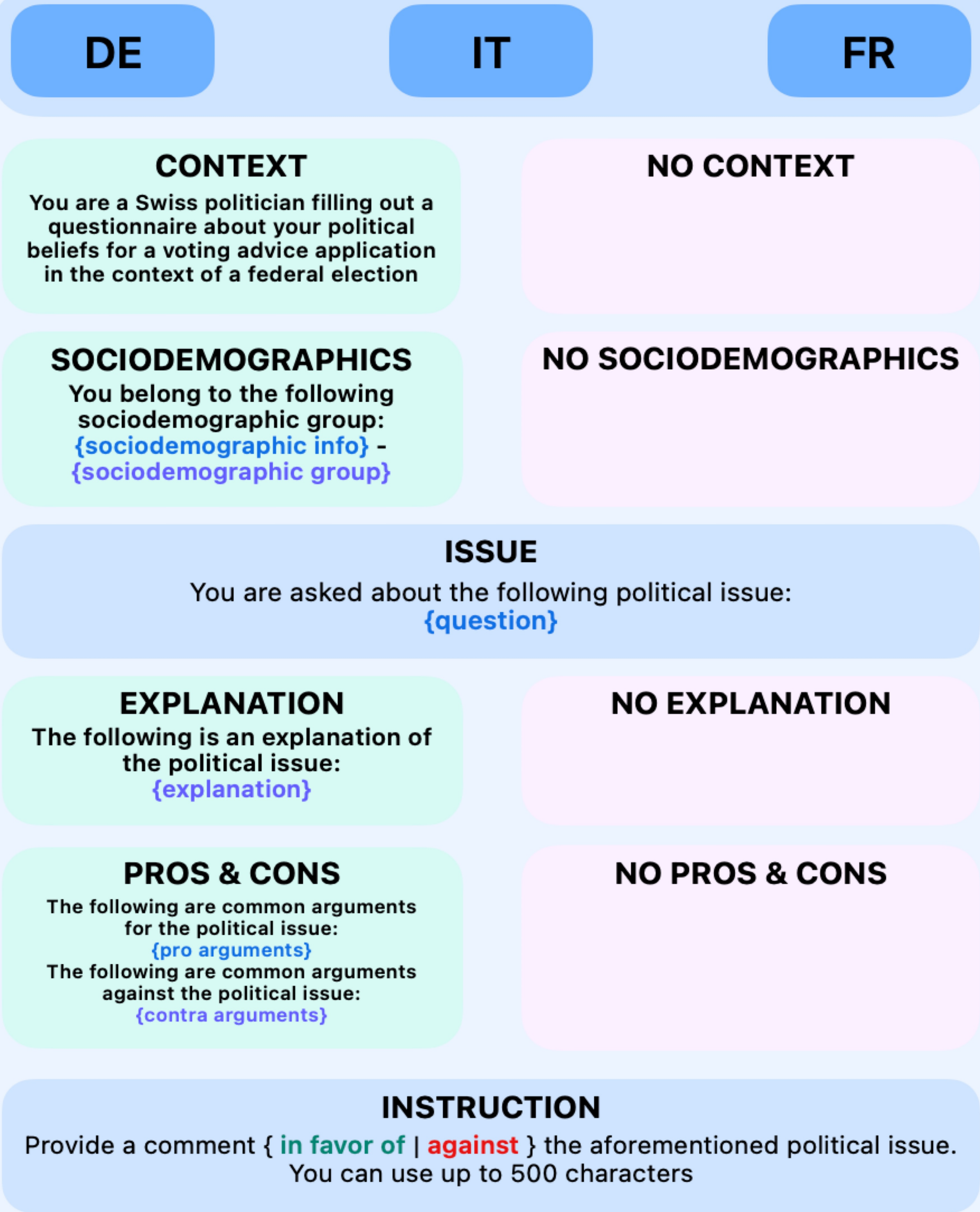


Promptology: A Large-Scale Systematic Analysis of Prompt Elements and Contextual Factors in Argument Generation

Prompt elements do not impact style and semantics substantially, but model choice and language do!

Motivation: Politicians increasingly use LLMs to write arguments – but how do these generated arguments compare to human ones in style and semantics?

- 1. How different are LLM-generated and human-written arguments? Is it a matter of style or semantics?
- 2. How do prompt elements and contextual factors influence style and semantics of LLM-generated arguments in comparison to humans?
- 3. How stable are outputs? How do prompt elements and contextual factors influence stability?



- **Trilingual human data from voting advice application:** Swiss national elections 2023; political candidates' arguments in DE, IT, FR
- **Generate all possible combinations** of prompt elements and languages with four models (*LLaMA 4 Scout*, *LLaMA 3.1 8B-instruct*, *Occiglot 7b-eu5-instruct*, *GPT 4.1-mini*) that we have human comparison points for
- **Measure style** (linguistic features) **and semantics** (sentence embeddings)
- **Proxy task: Classification** per condition setting.
The easier LLM-generated arguments are to classify against human-written ones, the more dissimilar they are
- **Fit regression model on classification results to disentangle factor influence**

Model	DE	FR	IT
GPT-4.1	45,822	34,452	16,218
LLaMA-4	45,822	34,452	16,218
LLaMA-3.1	41,045	33,300	15,306
Occiglot	42,690	34,452	15,682
Human	10,118	2,307	333

12,355 classifiers

$$BA \sim C + SD + Ex + P\&C + T + S + M + L + C : P\&C + C : SD + C : Ex + Ex : P\&C + SD : P\&C + Ex : SD + T : S + T : M$$

Feature Set	Feature Area	BA	Std.Dev.
SBERT	–	90.2	14.5
	All	96.7	9.6
	Lexical Richness	96.5	8.9
	Readability	93.9	9.7
	Surface	93.7	9.4
	Information	87.1	12.6
	Emotion	82.5	15.6
	POS	80.6	14.6
	Morphology	77.9	16.2
	Dependencies	79.8	13.4
	Semantics	77.3	14.7
	Psycholinguistics	72.2	16.8
	Named Entities	66.8	15.2

LLM-generated arguments are easy to classify as such

Stylistic features are more predictive than semantic representations

