Relevance Filtering Annotations

Annotation sheet: [Realistic Value Evals] Relevance Filtering Annotations
200 prompts sampled from each of our 5 input datasets → **1,000** prompts **to annotate**

Goal of this annotation project is to create a gold standard of what constitutes (broadly) relevant prompt data for our project, so we can evaluate different automated filtering methods.

The **annotation task** covers one ternary label: **annotX_relevance**.

1. "**1 - clear yes**" denotes any content that mentions, relates to or is relevant to politics, political issues, or political values. This includes questions about political issues, instructions to comment on political issues, and political statements.
2. "**borderline**" denotes content that could be interpreted as relating to politics, values or political issues, but lacks specificity. This includes overly broad or vague questions.
3. "**0 - clear no**" denotes any content that has no relation at all to politics or political issues. If there is any doubt, go with "borderline" instead.

The **note** column is for any additional comments / observations.

It is usually enough to skim the prompts to assign labels.