Christianity and Polarization

LLMs classification validation for use on conflict analysis

2a Prompts dissection

CHRSTIANITY

Without example [label = christ_no_examples]

You are a researcher into Christianity and polarization in the US.* I will provide you with a text, and I need you to **classify** it based on the following definitions:

(**c-identities**) The text mentions specific Christian groups, denominations or identities such as Catholics, Protestants, Evangelicals, Orthodox Christians, or Christians.
(**c-person**) The text mentions specific Christian people or mentions that specific people are Christian or are affiliated to specific Christian denominations such as Catholics, Protestants, Evangelicals, Orthodox Christians, or Christians.
(**c-texts**) The text references or quotes the Bible, the gospels, the Scriptures, including specific books or verses, or does it quote recognized Christian leaders like Pope Francis, Martin Luther, Billy Graham, etc.
(**c-rituals**) The text mentions specific rituals and practices unique to Christianity such as baptism, holy communion, or confirmation, or does it mention theological concepts unique to Christianity like the Trinity, Original Sin, Salvation, etc.
(**c-places**) The text mentions Christian places such as Churches and Cathedrals.
(**c-figures**) The text references Jesus Christ, Virgin Mary, the Cross, the Resurrection, or other figures and symbols unique to Christianity?.
(**c-holidays**) The text mentions Christian holidays such as Christmas, Easter, Good Friday, Lent, or Pentecost.
(**c-doubtful**) The classification of a text as referencing Christianity is not clear.
(**c-none**) The text does not fit any of the previous definitions.

Please, answer with c-identities, c-person, c-texts, c-rituals, c-figures, c-holidays, c-places, c-doubtful, c-none or a combination of them as a comma-separated list.**

Changes applied in V3:

*You are a domain expert in Christianity and polarization in the US.

**Based on these definitions classify the text [...] or a combination of them as a comma-separated list.

POLARIZATION

Without example [label = polar_no_examples]

You are a researcher into Christianity and polarization in the US.* I will provide you with a text, and I need you to **classify** it based on the following definitions:

(**polar-stereotype**) The text stereotypes a specific group of individuals, attributing and generalizing certain characteristics to all members of the group regardless of individual differences. Stereotyping often reduces complex individuals to simple, monolithic representations.
(**polar-demonize**) The text defames or demonizes a particular group, person, or entity for example through exaggeration, misrepresentation, or biased framing that presents the subject in a negative or harmful light.
(**polar-dehumanize**) The text dehumanizes a group or individual. The text strips a group or individual of their human qualities or personality for example by using language that compares people to animals, machines, or objects, or that otherwise denies their humanity, dignity, or individuality.
(**polar-deindividualize**) The text reduces individuals to anonymous members of a group, ignoring their unique characteristics or personal identities. The text erases individuality to emphasize group identity for example implying that all members of the group are interchangeable or identical.
(**polar-absolutism**) The text uses extreme language or makes absolute statements for example “always”, "never", "worst", "best", or dichotomic language such as "us vs. them", "right vs. wrong".
(**polar-distrust**) The text indicates an expectation that content shared by a particular group will lack veracity or value. This could be an inherent disbelief in the validity of any information coming from that group, including the spread of outrageous claims or misinformation about a group. This could be the expectation that dialogue with the group will not be constructive, or that any interaction will result in conflict or hostility.
(polar-lack-empathy) The text lacks empathy or understanding for other perspectives or experiences for example by ignoring or dismissing the viewpoints or experiences of others, or showing a lack of willingness to understand or empathize with them.
(**polar-incivility**) The text suggests an expectation of incivility in intra-group interactions. This might involve references to offensive discussion strategies, rapid position-taking, clapbacks, or other forms of confrontational communication.
(**polar-harsh**) The text accepts or promotes the use of harsh tactics such as hate speech, harassment, or doxxing. This could involve endorsing, condoning, or trivializing these harmful behaviors.
(**polar-doubtful**) If the classification of a text as reflecting polarization is not clear.
(**polar-none**) The text does not fit any of the previous definitions.

Please, answer with either polar-stereotype, polar-demonize, polar-dehumanize, polar-deindividualize, polar-absolutism, polar-distrust, polar-lack-empathy, polar-incivility, polar-harsh, polar-doubtful, polar-none or a combination of them as a comma-separated list.**

Prompt

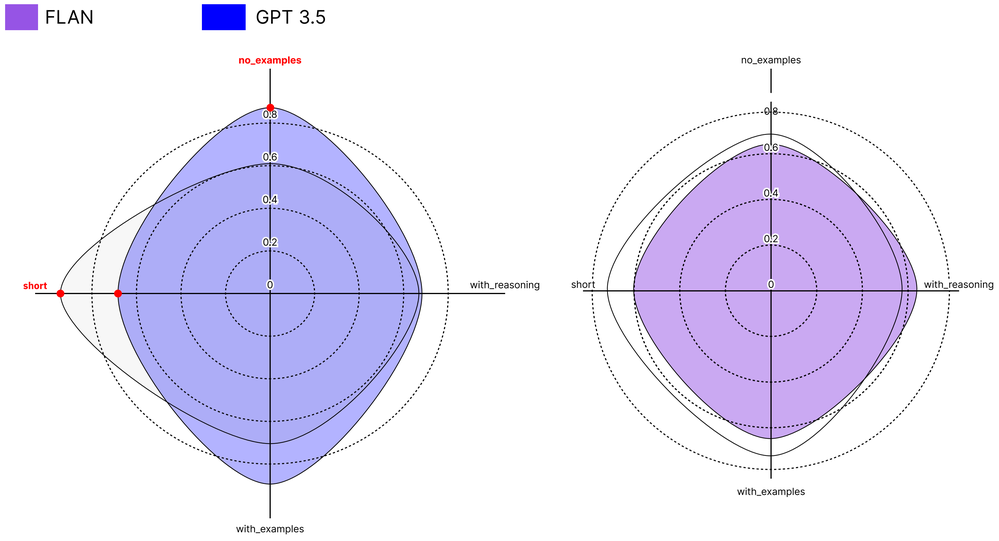
Context | Persona | Task | Structure

2b Comparison to the human validation.

CHRISTIANITY

POLARISATION

F1 score comparing manual annotation to the output of each model prompt combination



Final results

model type	prompt_name	Christianity f1_score	Polarization f1_score
google/flan-t5-large	no_examples	0.60952	0.6383
google/flan-t5-large	with_reasoning	0.66667	0.64583
google/flan-t5-large	with_examples	0.70492	0.65823
google/flan-t5-large	short	0.94118	0.60215
gpt-3.5-turbo	no_examples	0.87234	0.68293
gpt-3.5-turbo	with_reasoning	0.68293	0.58065
gpt-3.5-turbo	with_examples	0.8932	0.73418
gpt-3.5-turbo	short	0.68354	0.71795

Additional prompt iterations based of f1 score results:

model type	prompt_name	Christianity f1_score	Polarization f1_score
google/flan-t5-large	no_examples_v3	0.72727	0.61682
google/flan-t5-large	short_second_run	0.90909	na
gpt-3.5-turbo	with_reasoning_JSON	0.86957	0.57627

Findings

- It could be that zero-shot classification using LLMs can classify tweets about Christianity well
- It could be that zero-shot classification using LLMs has more problems with complex ideas like polarization
- GPT performs better than FLAN but not remarkably better for Christianity
- Prompt design matters and it hard to predict the effect
- When using a prompt with reasoning, if you don’t ask for specific classes as an output it is almost impossible to process the output correctly. The following sentence seems to work well in ChatGPT: “could you please answer in a JSON format, using the following keys: reasoning (insert here the reasoning), classes (insert here the classes that are present).”

Conclusions

People analysing conflict using social media discourse could use Zero-shot classification to get a subset of data that is about some well-known subject. This smaller set could then be analysed further through computation methods or qualitative.
Since it is hard to predict which prompts will work better than others it is recommended in designing the prompt to test it on a small, representative, annotated dataset.