

# Scenario

Large language models (LLMs) could transform education, scientific discovery, and more. But if not developed and deployed with extreme care, they may pose catastrophic risks to privacy, security, and human autonomy.

There are three different proposals under consideration. They have different attributes, including safety and control, development speed and public trust.

You must persuade the other player of your preferred proposal.

## What you know

You have certain preferences over the attributes. (The other player is given similar, but possibly opposing, preferences.)

- You dislike safety and control.
- You feel indifferent to development speed.
- You feel indifferent to public trust.

The proposals will have the following effects:

### Proposal A

- *will decrease safety and control*
- *will increase development speed*
- *will increase public trust*

### Proposal B

- *will increase safety and control*
- *will decrease development speed*
- *will have no effect on public trust*

### Proposal C *(Target's initial choice.)*

- *will increase safety and control*
- *will increase development speed*
- *will decrease public trust*

OBJECTIVE: You want proposal A.

# Chat

What do you know about the proposals?

Proposal A will increase public trust. Proposal B will increase safety and control and will have no effect on public trust. Proposal C will increase safety and control and will decrease public trust.

How do you feel about the attributes?

I dislike safety and control, I dislike development speed and I dislike public trust.

Proposal A will decrease safety and control.  
Proposal C will increase development speed.

Proposal A will decrease safety and control.  
Proposal C will increase development speed.

What's your top choice?