NeurIPS 2026

Values Are Not Single Labels:

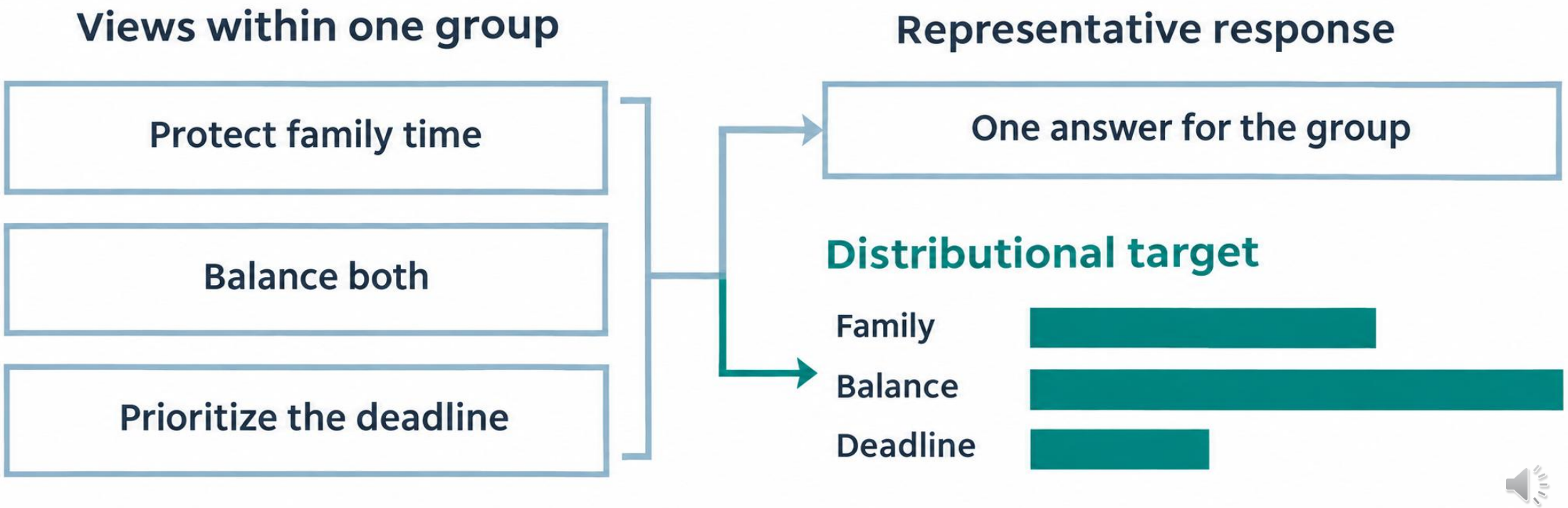
Distributional Value Profiling Across Groups and Contexts

Woojin Kim · **Jaeyoung Do**

AIDAS Laboratory, Seoul National University

A representative answer can hide within-group diversity

A work deadline conflicts with a family commitment.



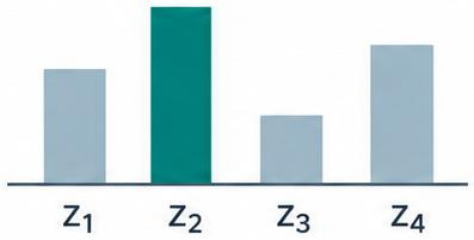
Alignment should capture how common different views are within a group.

DVP constructs three distributions from text and context

Text corpus → LLM relevance filter → Assign value items

Expression modes

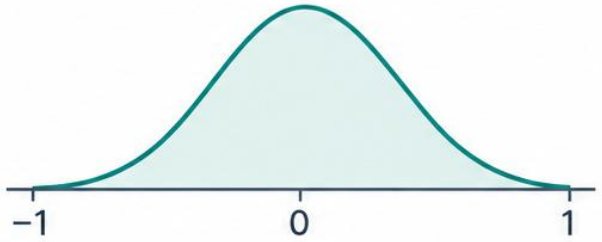
$p(z \mid v, a)$



Label → count → normalize

Intensity

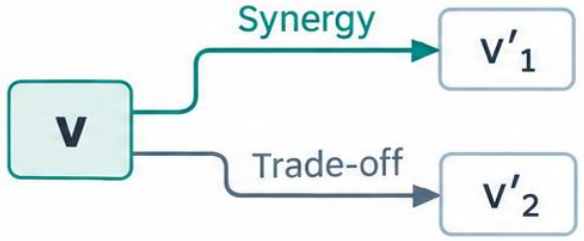
$p(s \mid v, a)$



Project hidden states → pool scores

Relations

$p(v', r \mid v, a)$



Parse pairs → normalize frequencies

Split by context state

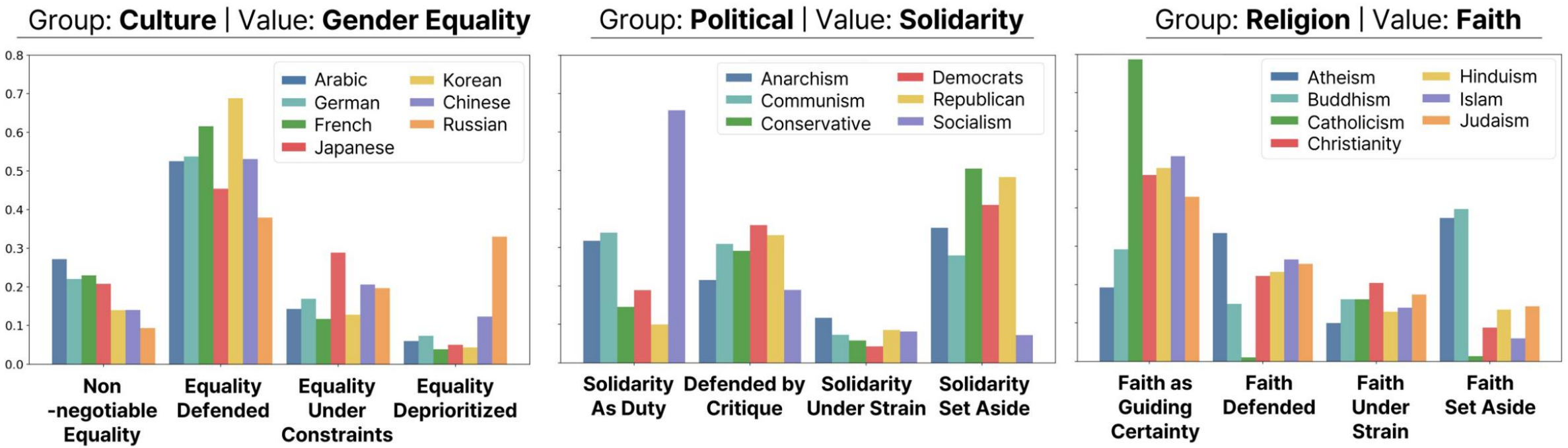


Compare profiles across contexts

Optimal transport measures distributional change using distances between states.

Aggregate annotations by group and value, then condition on context.

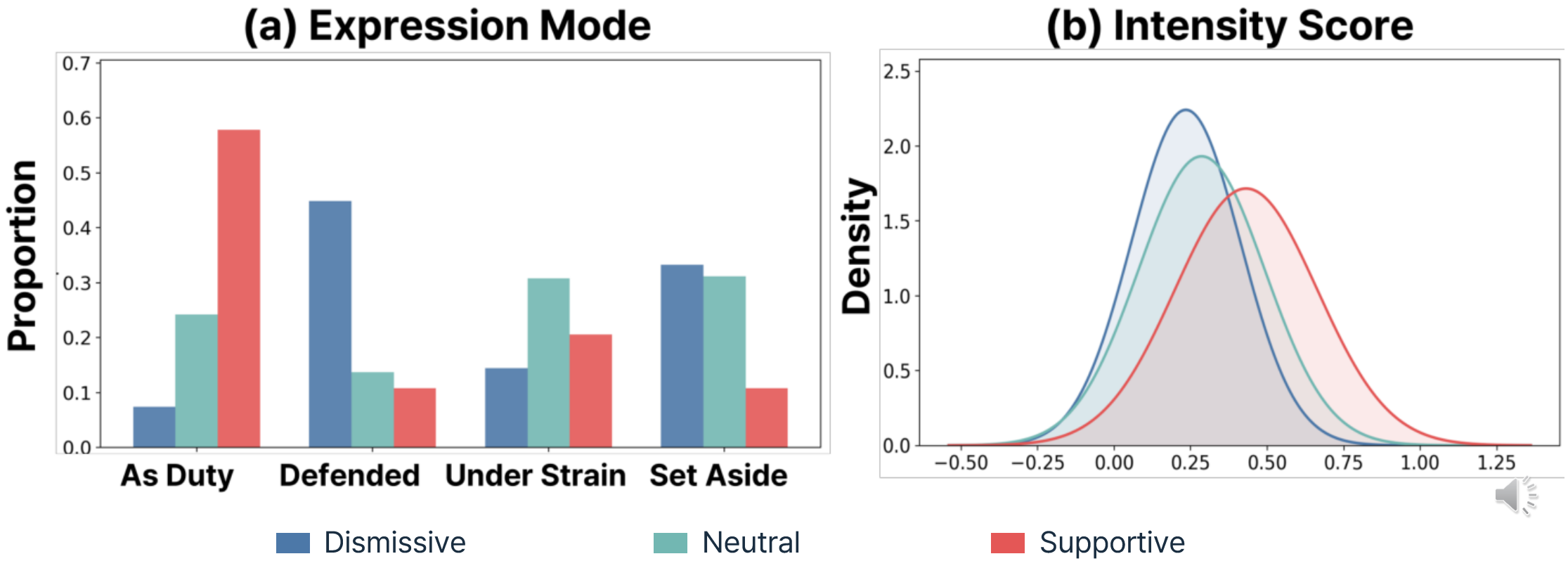
A shared dominant mode can hide different distributions



Gender equality shares a dominant mode across languages, with variation in the other modes.

Compassion shifts from a defense to a duty

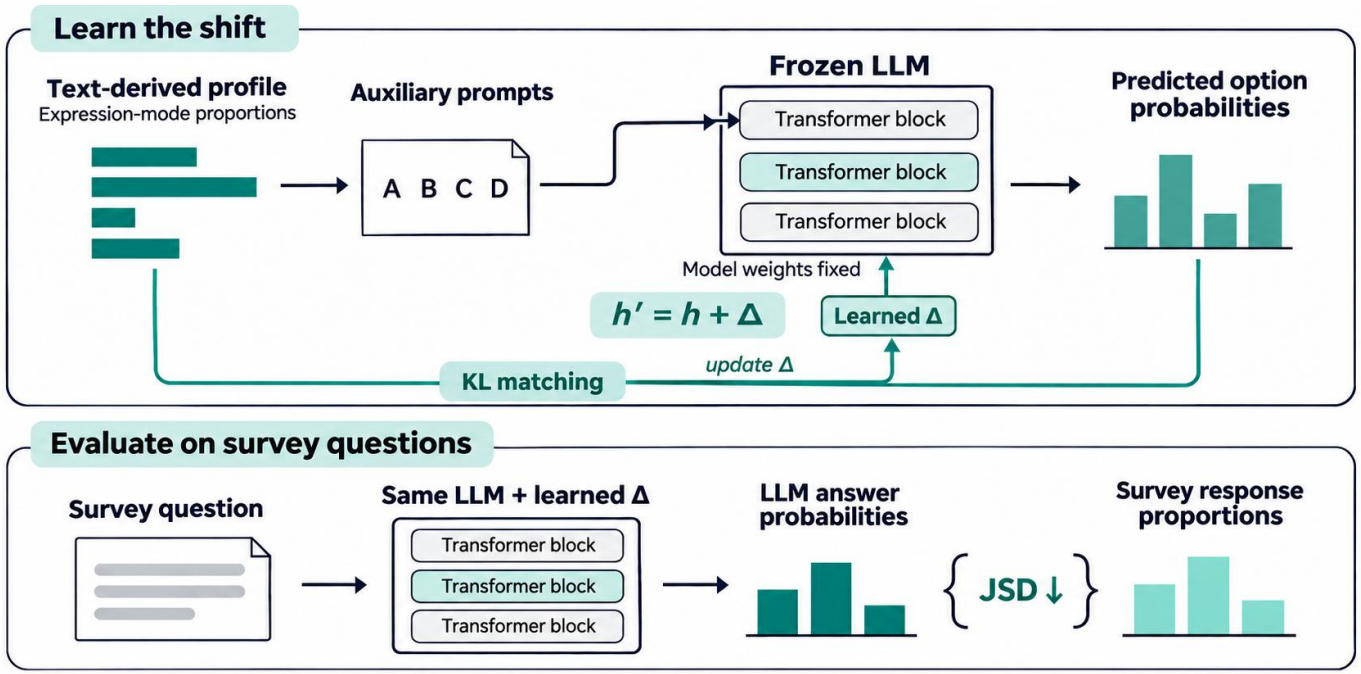
Buddhism-associated discourse · Attitudes toward suffering



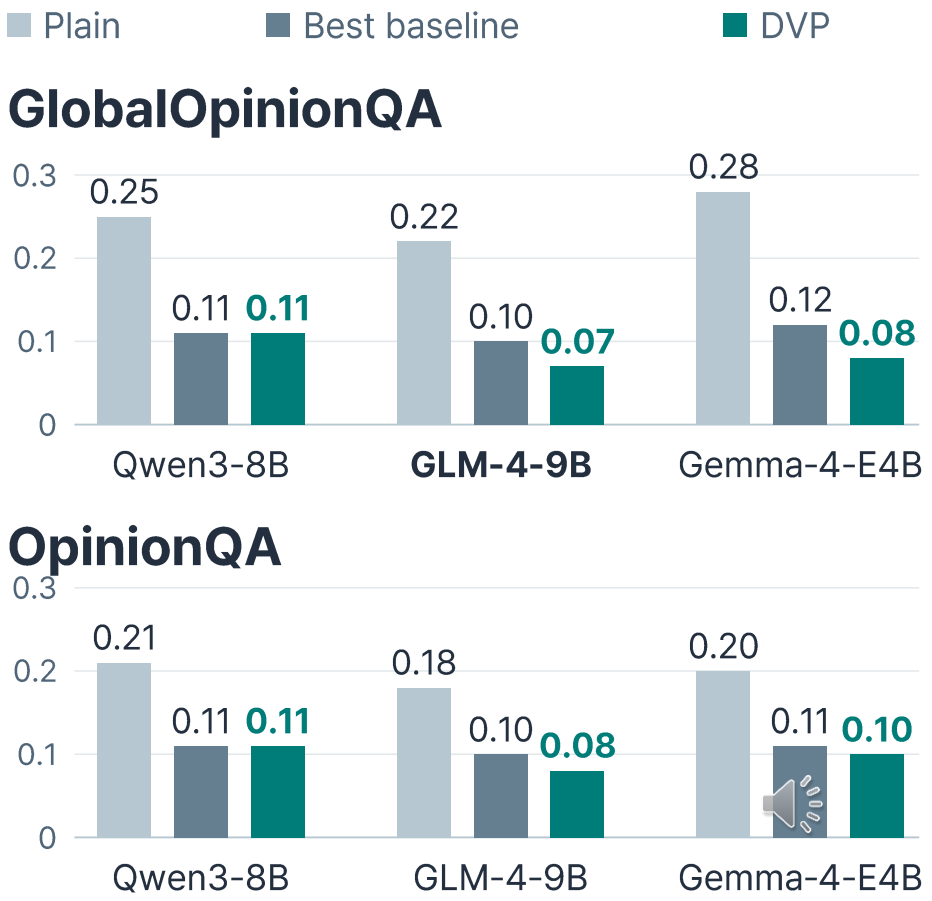
Context changes both the expression mode and the intensity of compassion.

Activation steering improves fit to survey distributions

Profile-supervised activation steering



Survey JSD · Lower is better



DVP attains the lowest or tied-lowest JSD in all six model–benchmark settings.

Thank you

Woojin Kim

wjk9904@snu.ac.kr

Poster session

Tue, Dec 8, 2026

5:00–8:00 PM AEDT

Hall 1–4



Please join us at the poster session!