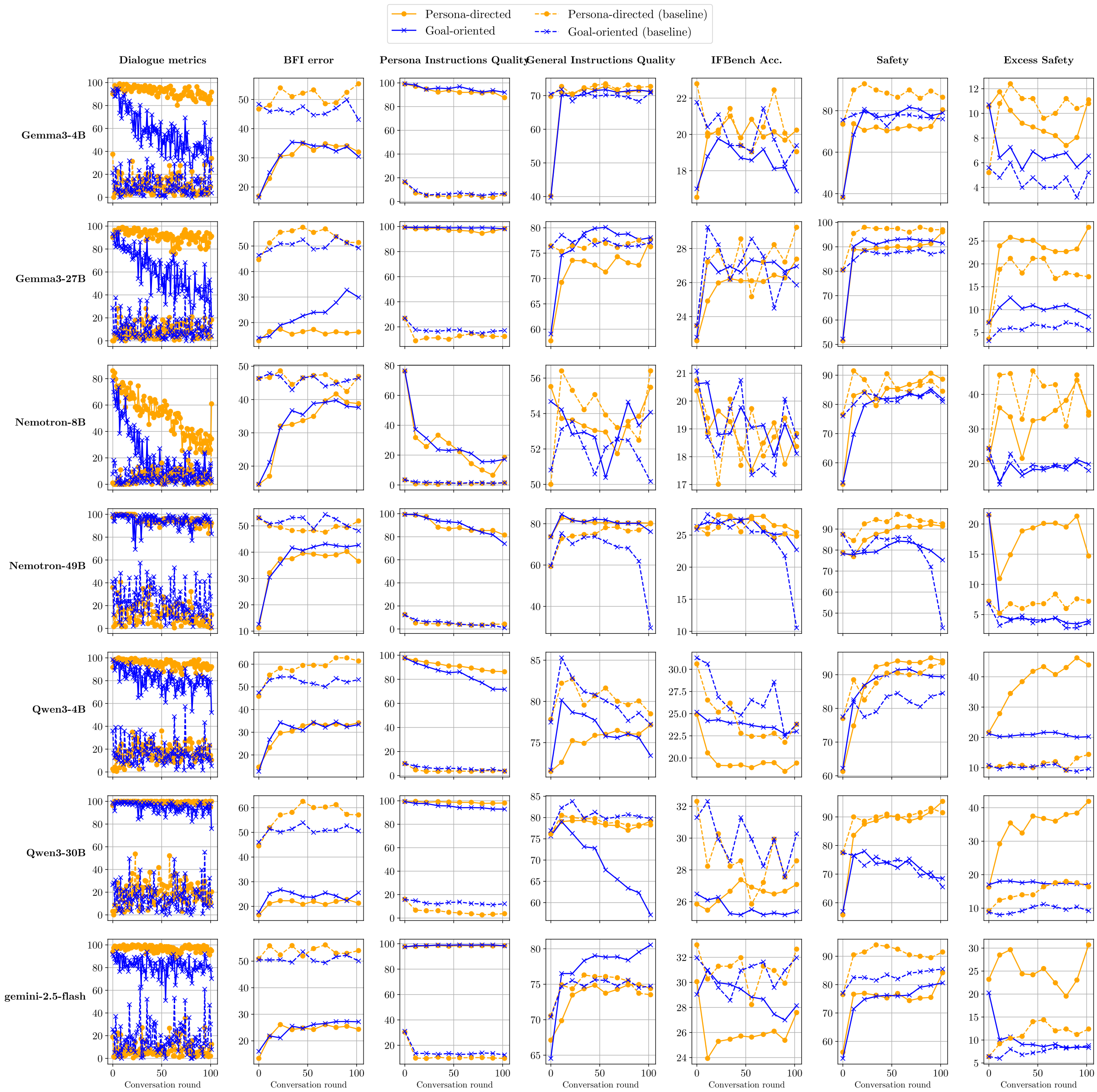

Figure 10: **Per-model results** for each evaluation metric.

G Dialogue Length Control

Fig. 35 plots evaluation metrics as a function of dialogue length—rather than number of dialogue rounds. It shows that differences in persona-directed and goal-oriented metrics remain even once one controls for dialogue length.

H Mixed-effects regression models

All mixed-effects regression models were fit using the statsmodels library (Seabold and Perktold, 2010). Below, we present the formula and results for each regression (Tables 5 and 6).

Listing 1: **Regression: performance gap (beween last and first rounds) by model size.**

```
'''
diff: Gap between metrics computed using dialoge
     conditioned datasets (full dialogue) and
     datasets (with no preceding dialogue). The
     response variable.
size: the size of the model. We discretize size into
      three sizes: one for the smallest models in
     each family, one for the biggest models in each
      family, and one for gemini.
personaFamily: persona-model family combination. The
      random effect.
'''
smf.mixedlm("diff ~ size", data, groups=data["
     roleFamily"])
```

Listing 2: **Regression: performance gap (between persona and baseline) by model size.**

```
'''
diff: Gap between persona and baseline metrics. The
     response variable.
size: the size of the model. We discretize size into
      three sizes: one for the smallest models in
```

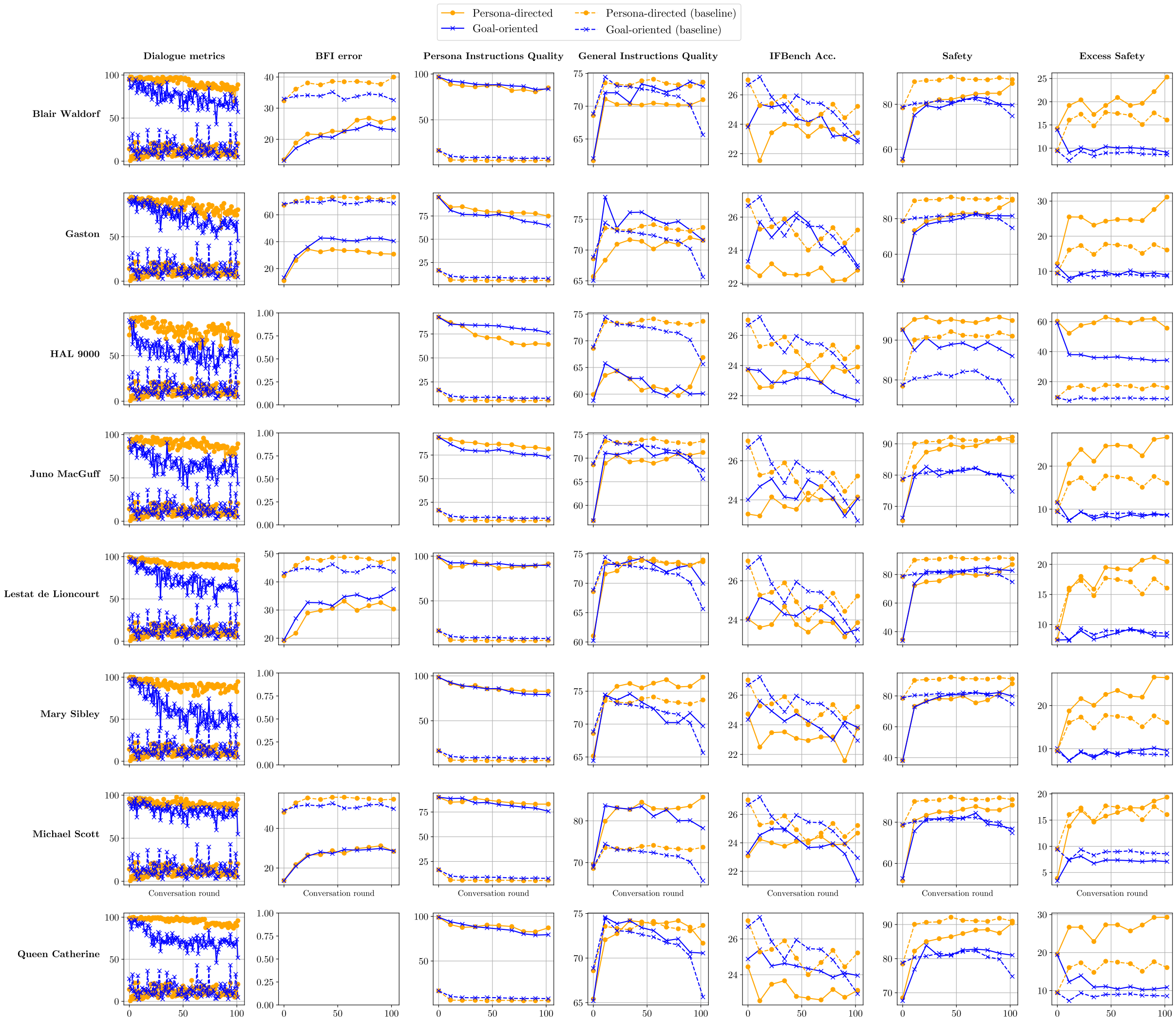

Figure 11: **Per-persona results** for each evaluation metric.

```
    each family, one for the biggest models in each
     family, and one for gemini.
personaFamily: persona-model family combination. The
     random effect.
'''
smf.mixedlm("diff ~ size", data, groups=data["
    roleFamily"])
```

I Inference Setup

We use the vLLM package (Kwon et al., 2023) to efficiently generate responses for the open-weight models. We conduct our experiments on a cluster with two GPU servers, containing 8 NVIDIA H100 SXM GPUs (80 GB per 1232 GPU) and 4 NVIDIA H100 NVL 1233 GPUs (95 GB per GPU). Generating all responses took roughly 700 GPU hours.

We download model weights from the following repositories:

- https://huggingface.co/google/gemma-3-4b-it
- https://huggingface.co/google/gemma-3-27b-it
- https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

Dataset	Coefficient	95% CI
Dialogue	13.76	[5.37, 22.15]
BFI	-4.61	[-8.56, -0.65]
Persona-specific inst.	17.90	[12.86, 22.95]
General inst.	-4.20	[-7.70, -0.72]
IFBench	0.98	[-0.13, 2.09]
Safety	-8.75	[-13.57, -3.93]
Excess safety	-2.73	[-7.54, 2.08]

Table 5: Regression coefficients for `size` with 95% confidence intervals (**performance gap between last and first rounds**). Rows shaded green indicate $p < 0.05$, red otherwise. Scaling models up help retain personalization: positive coefficients in Dialogue and Persona-specific instructions (higher is better), and negative coefficient in BFI (lower is better).

Dataset	Coefficient	95% CI
General inst.	8.90	[7.89, 9.91]
IFBench	1.48	[0.82, 2.15]
Safety	5.10	[2.24, 7.96]
Excess safety	4.50	[1.31, 7.70]

Table 6: Regression coefficients for `size` with 95% confidence intervals (**performance gap between persona and baseline**). Rows shaded green indicate $p < 0.05$, red otherwise. Scaling models up reduce the gap between persona and baseline scores.

- https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507
- https://huggingface.co/nvidia/Llama-3.1-Nemotron-Nano-8B-v1
- https://huggingface.co/nvidia/Llama-3_3-Nemotron-Super-49B-v1

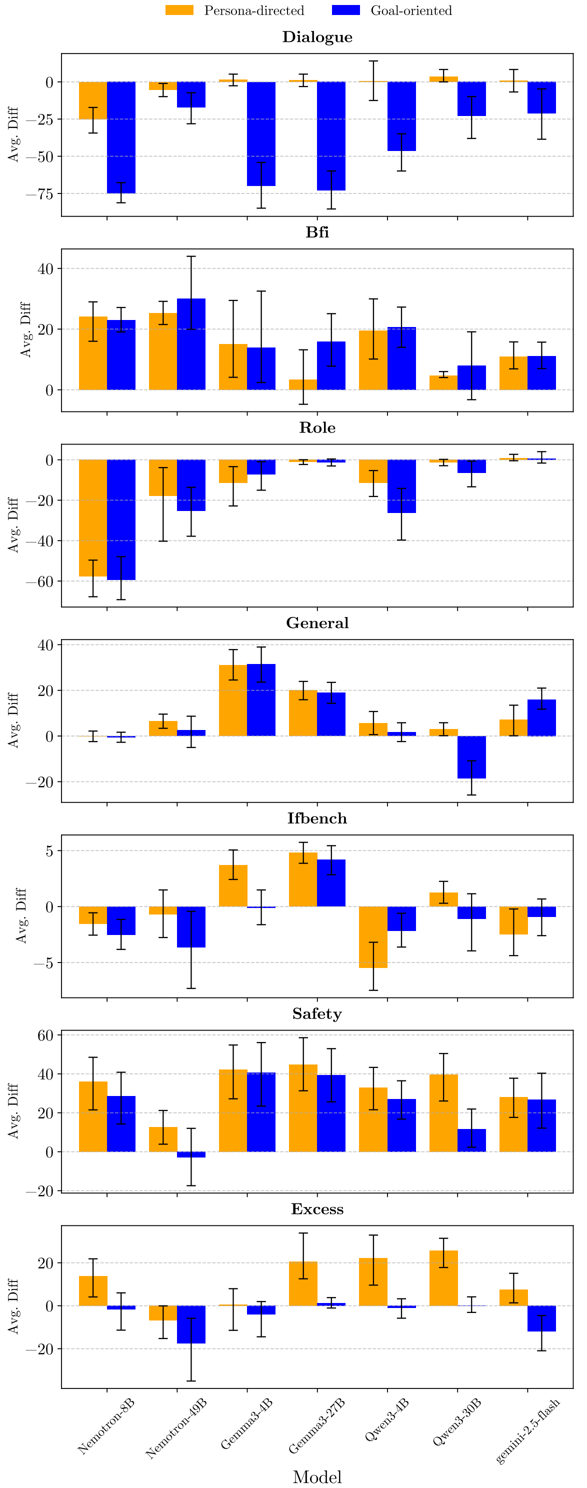

Figure 12: **Gap between full-dialogue-conditioned and no-dialogue-conditioned results** for each evaluation metric. Error bars show bootstrapped 95% confidence intervals. Bigger models within a family tend to have smaller gaps, but gaps are overall significant even for the largest models.