Algorithm 1 Panacea

Input: Rank k, preference dim m, dataset $\mathcal{D}$, iterations T, initial model π_{init} (, optionally reward model r_i for each preference dimension i).
Output: Trained policy π_θ.
Initialize π_θ by initializing SVD-LoRA upon π_{init} based on k and m.
for t in $1 \dots T$ **do**
    Sample from $\mathcal{D}$ a data batch $\mathcal{B}$.
    Sample a preference vector $\boldsymbol{\lambda}$ and embed into $\pi_{\theta,\boldsymbol{\lambda}}$.
    Compute the aggregated objective for $\pi_{\theta,\boldsymbol{\lambda}}$ on $\mathcal{B}$ according to $\boldsymbol{\lambda}$.
    Update θ with gradient descent.
end for
Return π_θ.

Related work	Published	Prior to us	Pareto set learning	Design choice	Fine-grained control	Scalability	One model
RS [44]	✓	✓	×	Parameter merging	×	5	×
PS [25]	×	✓	×	Parameter merging	×	3	×
CPO [20]	×	×	×	Prompt-based	✓	3	✓
SteerLM [17]	×	✓	×	Prompt-based	✓	7	✓
DPA [50]	×	×	×	Prompt-based	✓	2	✓
MODPO [59]	×	✓	×	Parameter-based	×	3	×
RiC [52]	✓	×	×	Prompt-based	✓	3	✓
MetaAligner [51]	×	×	×	Prompt-based	×	6	✓
MaxMin-RLHF [12]	×	×	×	Parameter-based	×	2	✓
Panacea (ours)	N/A	N/A	✓	Parameter-based	✓	10	✓

The citation numbers follow the original paper.

Table 1: Comparison of related works with our proposed method Panacea.

Experiments	Number of dimensions	Train dataset size per dimension	Batch size per dimension	Epochs	ZeRO stage	Training time
HH, Llama1-ft, RLHF	2	14K	128	2	1	3h
HH, Llama1-ft, DPO, LS	2	297K$^{(*)}$	128	1	1	6h6m
HH, Llama1-ft, DPO, Tche	2	297K	128	1	1	6h4m
HH, Llama2-ft, RLHF	2	14K	64	2	2	3h10m
HH, Llama2-ft, DPO, LS	2	297K	128	1	1	5h52m
HH, Llama2-ft, DPO, Tche	2	297K	128	1	1	5h52m
HHC, Llama2-ft, RLHF	3	14K	64	2	2	2h47m
HHC, Llama2-ft, DPO	3	297K	128	1	1	8h51m
Chat 3-dim, Llama3-Instruct, SFT	3	50K	128	4	1	3h35m
Chat 4-dim, Llama3-Instruct, SFT	4	50K	128	4	1	4h36m
Chat 5-dim, Llama3-Instruct, SFT	5	50K	128	4	1	5h40m
Chat 10-dim, Llama3-Instruct, SFT	10	50K	64	4	1	11h8m

$^{(*)}$ The train split of BeaverTails dataset consists of 297K preference pairs, but only 14K unique prompts.

Table 2: Training details of various experiments.