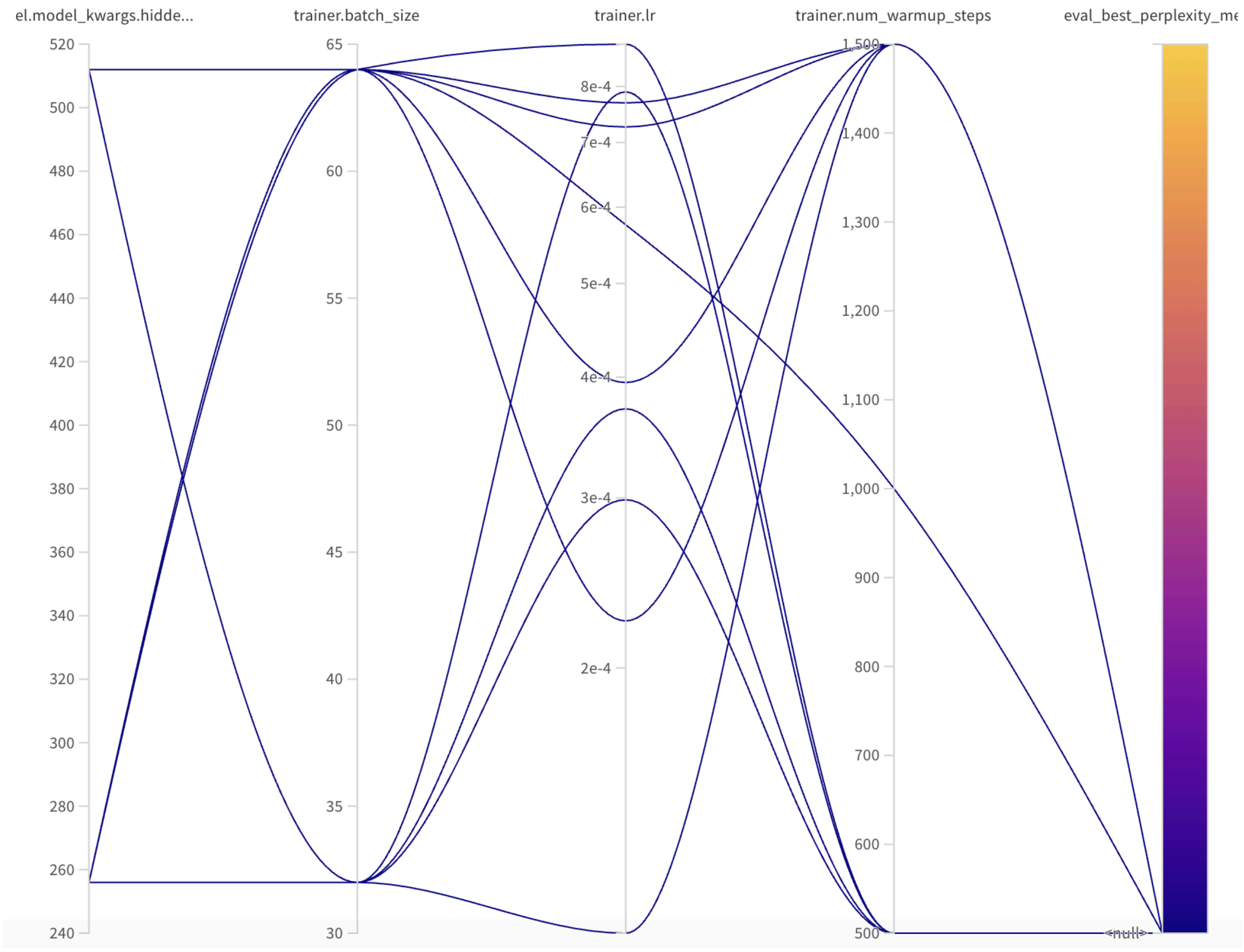
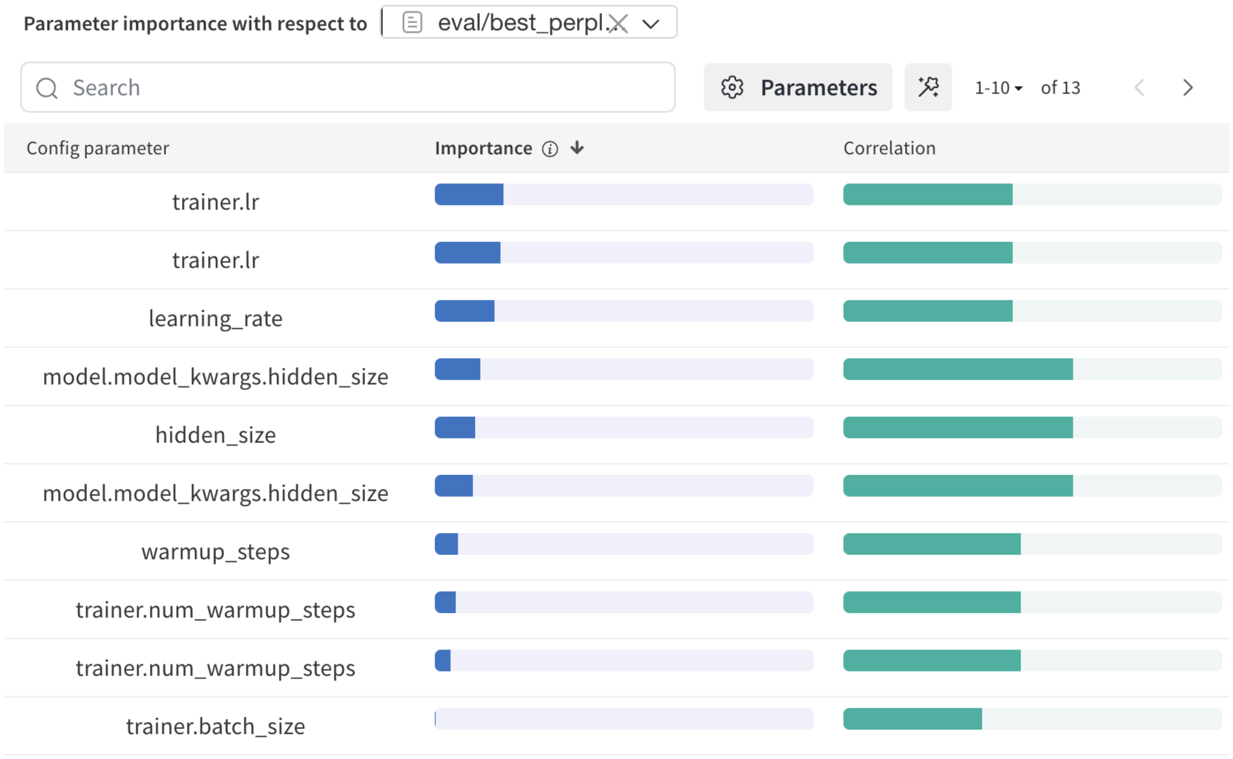
We ran a hyperparameter sweep with 10 experiments trying to find out what's the most optimal hyperparameter configuration for minimizing perplexity. We used bayesian sweep as our sweep method and tested these configurations:

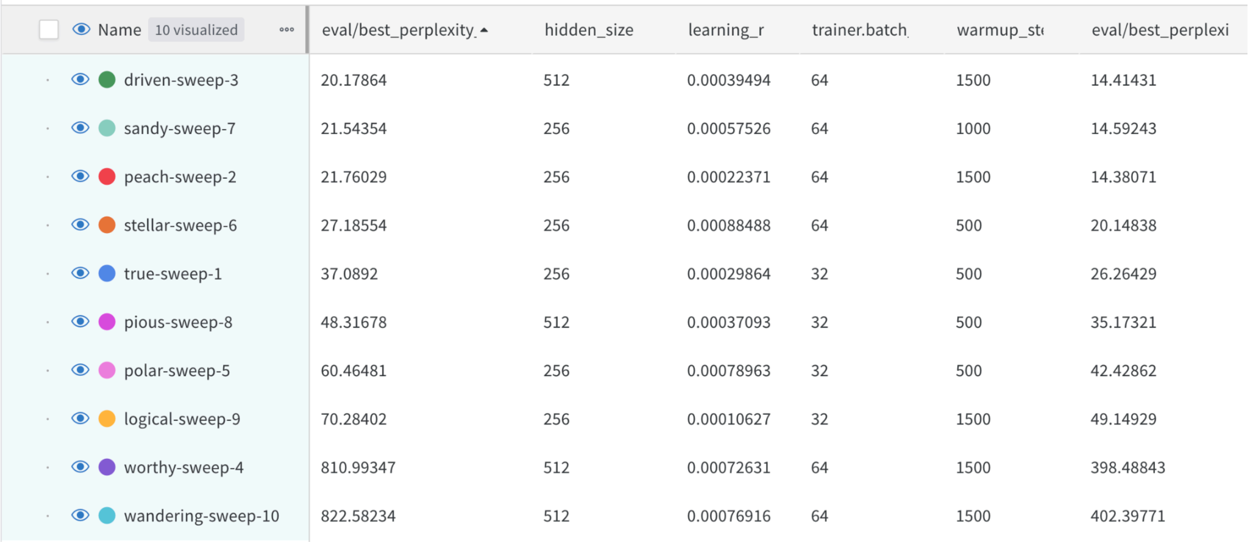
- **trainer.lr** - min: 1e-4 to max:1e-3
- **trainer.num_warmup_steps:** values: [500, 1000, 1500]
- **trainer.batch_size:** values: [32,64]
- **model_kwargs.hidden_size:** values: [256,512]

After logging everything into wandb, Here are some visualization/data for the sweep.





Config parameter	Importance ⓘ ↓	Correlation
trainer.lr		
trainer.lr		
learning_rate		
model.model_kwargs.hidden_size		
hidden_size		
model.model_kwargs.hidden_size		
warmup_steps		
trainer.num_warmup_steps		
trainer.num_warmup_steps		
trainer.batch_size		



Name 10 visualized	eval/best_perplexity ▴	hidden_size	learning_r	trainer.batch	warmup_st	eval/best_perplexi
● driven-sweep-3	20.17864	512	0.00039494	64	1500	14.41431
● sandy-sweep-7	21.54354	256	0.00057526	64	1000	14.59243
● peach-sweep-2	21.76029	256	0.00022371	64	1500	14.38071
● stellar-sweep-6	27.18554	256	0.00088488	64	500	20.14838
● true-sweep-1	37.0892	256	0.00029864	32	500	26.26429
● pious-sweep-8	48.31678	512	0.00037093	32	500	35.17321
● polar-sweep-5	60.46481	256	0.00078963	32	500	42.42862
● logical-sweep-9	70.28402	256	0.00010627	32	1500	49.14929
● worthy-sweep-4	810.99347	512	0.00072631	64	1500	398.48843
● wandering-sweep-10	822.58234	512	0.00076916	64	1500	402.39771

Based on these 10 experiment runs, the most optimal hyperparameter we found for minimizing perplexity are:

- **learning_rate**: 0.00039494
- **batch_size**: 64
- **warmup_step**: 1500
- **hidden_size**: 512