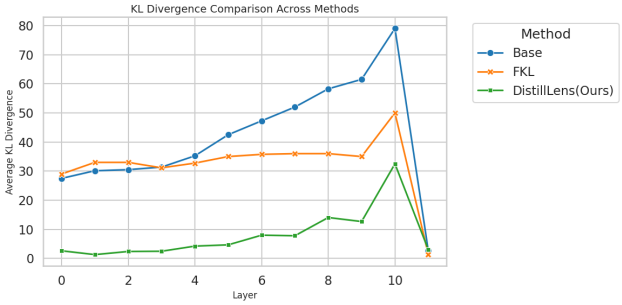

Figure 1: The distilled student (FKL) notably diverges from the teacher model for the hidden layers, except for the final layer. In contrast, DistillLens successfully minimizes this KL divergence across the intermediate layers, indicating the strong faithfulness to teacher's internal belief states.

Table 1: **Distillation: Qwen2.5-3B → Qwen2.5-1.5B:** DistillLens shows consistent improvement

Method	Dolly	SelfInst	Vicuna	SNI	UnNI
SFT	24.4	19.2	18.0	38.1	33.2
KD (FKL)	24.5	20.7	17.8	38.5	32.9
RKL	26.0	21.0	19.5	40.1	35.5
FKL + RKL	25.8	21.2	18.9	40.0	35.2
JSD	24.1	18.7	18.0	37.0	32.6
AKL	25.0	20.4	19.6	38.8	34.6
DistillLens	**26.8**	**21.7**	**20.6**	**42.7**	**38.6**

Table 2: For on-policy, DistillLens is compared against the MSE baseline. On average, MiniLLM + DistillLens gives superior results.

Method	**Dolly**	**SelfInst**	**Vicuna**	**S-NI**	**UnNI**
MiniLLM	24.6	13.2	16.9	25.3	26.6
MiniLLM + MSE	25.0	12.3	18.2	24.8	25.6
MiniLLM + DistillLens	25.1	13.3	17.8	26.1	27.2