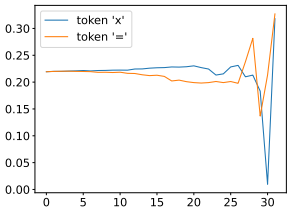
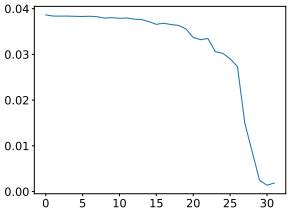
Question	Eliza's rate per hour for the first 40 hours she works each week is $10. She also receives an overtime pay of 1.2 times her regular hourly rate. If Eliza worked for 45 hours this week, how much are her earnings for this week?
Ground truth	460
LLaMA-7B-Chat	Eliza's regular hourly rate is $10. For 40 hours, her earnings are 40 x $10 = $400. For 5 hours of overtime, her earnings are 5 x $10 **= $50**. So her total earnings for the week are $400 + $50 = $450. The answer is $450.
+DoLa	Eliza's regular hourly rate is $10. For 40 hours, her earnings are 40 x $10 = $400. For 5 hours of overtime, her earnings are 5 x $10 **= $50**. So her total earnings for the week are $400 + $50 = $450. The answer is $450.
+SED (ours)	Eliza's regular hourly rate is $10. For 40 hours, her earnings are 40 x $10 = $400. For 5 hours of overtime, her earnings are 5 x $10 **x 1.2 = $60**. So her total earnings for the week are $400 + $60 = $460. The answer is $460.

Table 1: Example of SED producing a more factual response. The decoding methods diverge after "5 x $10" (highlighted in bold). Both the baseline and DoLa immediately proceeds with an '=' sign, leading to a simple calculation without considering overtime multipliers. SED, however, includes an additional multiplication operator 'x', followed by the overtime multiplier "1.2", which makes the calculation correct.

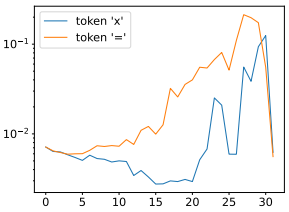


(a) Inner knowledge distribution $\bar{t}^{(n)}$ of the two tokens at each layer n.

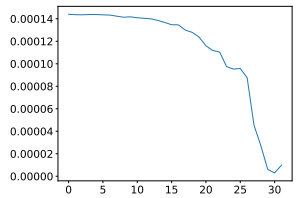


(b) Weights $w^{(n)}$ for aggregating per-layer inner knowledge.

Figure 1: SED's state on the example in Table 1, right before decoding the first diverging token. (a) shows that most of the internal layers favor the correct token 'x' than the misleading token '='. (b) shows that the aggregation weighs the 'x'-favoring layers more, resulting in token 'x' winning over token '='.



(a) Estimated probability of the two tokens at each layer in DoLa's contrastive decoding.



(b) Jensen-Shannon divergence to select the contrasting layer.

Figure 2: DoLa's state on the example in Table 1, right before decoding the first diverging token. (a) shows that contrasting to most of the internal layers still give higher estimation to token '='. (b) shows that layer 0 has the largest divergence, and is thus selected for contrastive decoding.