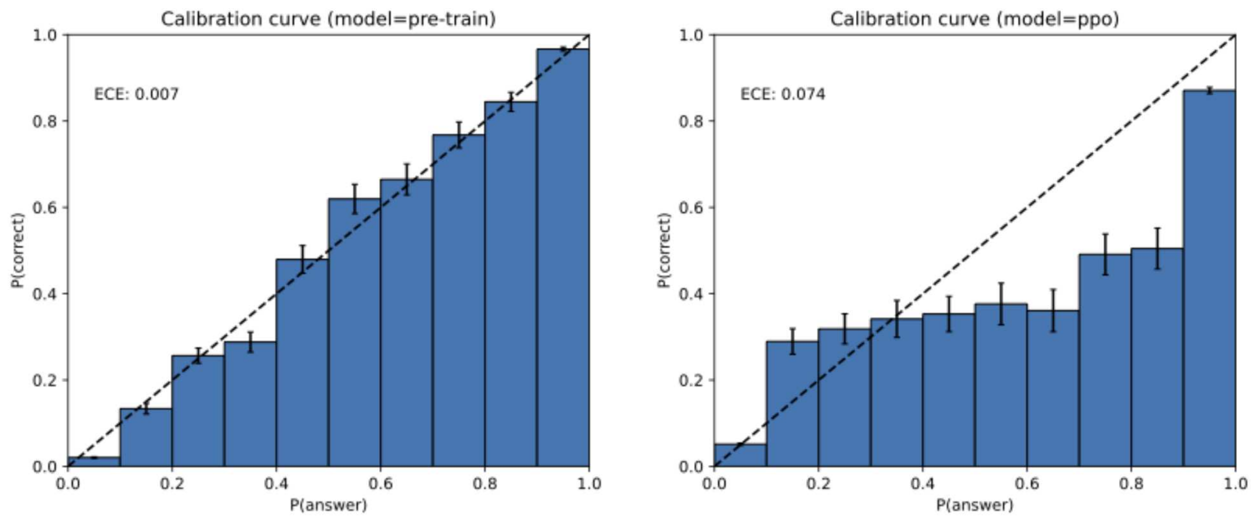

Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

Then, a simple calculation shows that δ is the magnitude of the derivative of the loss with respect to the scaling factor s, evaluated at $s = 1$:

$$\delta = \left| \frac{d}{ds}\mathcal{L}(\hat{p}_s)\Big|_{s=1} \right|.$$

If $\delta \neq 0$, then rescaling by some $s \neq 1$ would reduce the loss, so the loss is not at a local minimum. For any class of language models powerful enough to approximate such simple rescaling, local optimization should yield small δ. Note that δ, being defined at a single threshold $t = 1/|\mathcal{E}|$ is weaker than notions such as Expected Calibration Error (ECE) which integrate over thresholds t.

Hallucinations are inevitable *only for base models.* Many have argued that hallucinations are inevitable (Jones, 2025; Leffer, 2024; Xu et al., 2024). However, a non-hallucinating model could be easily created, using a question-answer database and a calculator, which answers a fixed set of questions such as "What is the chemical symbol for gold?" and well-formed mathematical calculations such as "3 + 8", and otherwise outputs IDK. Moreover, the error lower-bound of Corollary 1 implies that language models which do not err must not be calibrated, i.e., δ must be large. As our derivations show, calibration—and, hence, errors—is a natural consequence of the standard cross-entropy objective. Indeed, empirical studies (Fig. 2) show that *base models* are often found to be calibrated, in contrast to post-trained models which may deviate from cross-entropy in favor of reinforcement learning.

3.2 The reduction with prompts

Henceforth, we generalize the setting of Section 3.1 to include prompts (contexts) $c \in \mathcal{C}$ drawn from a *prompt distribution* μ. Each example $x = (c, r)$ now consists of a prompt c and plausible response r. The analysis above corresponds to the special case in which μ assigns probability 1