A Appendix

A.1 Proof of Lemma 1

The density function of (s, t, o) in GS–PB joint Data Generation Process could be written as

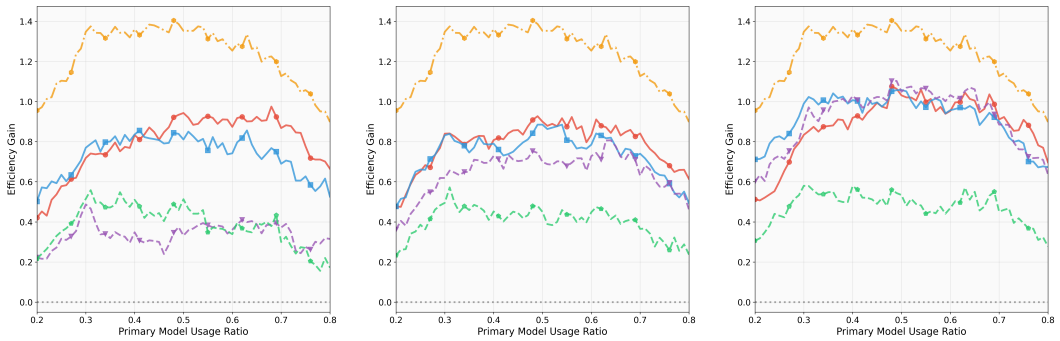
$$f(s,t,o) = \kappa^t(1-\kappa)^{1-t} f_{\mathcal{Q}}^t(s) f_{\mathcal{Q}'}^{1-t}(s) f_r^t(o \mid s) f_y^{1-t}(o \mid s),$$

where $f_r(\cdot \mid s)$ and $f_y(\cdot \mid s)$ represent the conditional probability density function of r_i and y_i given $q_i = s$, following (1) and (2), respectively. This could be further written as

$$f(s,t,o) = \underbrace{(\kappa f_{\mathcal{Q}}(s) + (1-\kappa) f_{\mathcal{Q}'}(s))}_{f_{\kappa\mathcal{Q}+(1-\kappa)\mathcal{Q}'}(s)} \cdot \underbrace{\frac{\kappa^t(1-\kappa)^{1-t} f_{\mathcal{Q}}^t(s) f_{\mathcal{Q}'}^{1-t}(s)}{\kappa f_{\mathcal{Q}}(s) + (1-\kappa) f_{\mathcal{Q}'}(s)}}_{Pr(t_i=t|s)=tp(s)+(1-t)p(s)} \cdot \underbrace{f_r^t(o \mid s) f_y^{1-t}(o \mid s)}_{f_{o^{(t)}}(o|s)}, \quad (11)$$

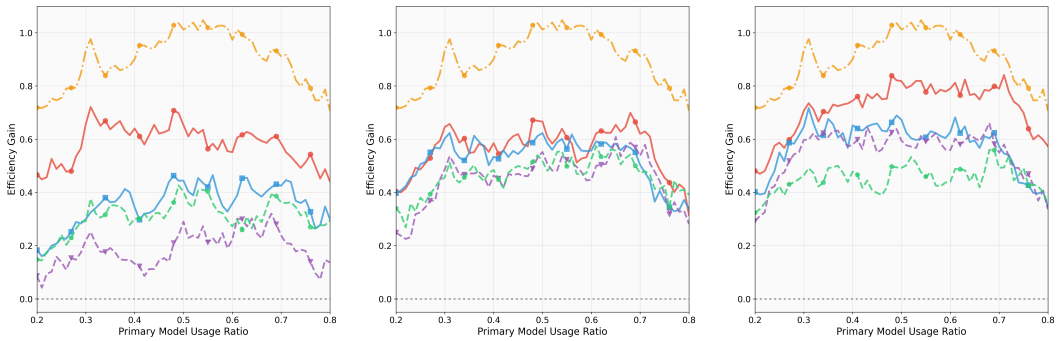
recalling the notation in Causal Data Generation Process, and thereby show the distributional equivalence of two processes. □

A.2 Additional numerical results for §4



(a) $n = 100$ (b) $n = 500$ (c) $n = 1000$

Figure 3: The efficiency gains of different routing strategies compared with the random routing baseline versus the primary model usage ratio. The query embedding dimension is reduced to 100 via PCA and all regressions are implemented via random forest. Other explanations are the same as Figure 1.



(a) $n = 100$ (b) $n = 500$ (c) $n = 1000$

Figure 4: The efficiency gains of different routing strategies compared with the random routing baseline versus the primary model usage ratio. The query embedding dimension is reduced to 100 via PCA and all regressions are implemented via XGBoost. Other explanations are the same as Figure 1.

A.3 THE USE OF LARGE LANGUAGE MODELS (LLM)

For this project, LLMs were used to polish the writing of the main paper and to assist with coding for the numerical experiments.