1. Математическая модель экономии токенов

1.1. Обозначения

- N — общее количество запросов.
- p — вероятность кэш-попадания.
- c_e — стоимость одного токена для создания эмбеддинга.
- c_q — стоимость одного токена для запроса к модели завершения.
- t_{input} — количество токенов входного запроса.
- t_{output} — количество токенов выходного ответа.

1.2. Стоимость запросов

1.2.1. С кэшированием

Если запрос попадает в кэш (p), то стоимость равна:

$$C_{\text{cache}} = c_e * t_{\text{input}}$$

Если запрос не попадает в кэш ($1 - p$), то стоимость равна:

$$C_{\text{miss}} = c_e * t_{\text{input}} + c_q * \left(t_{\text{input}} + t_{\text{output}}\right)$$

Общая стоимость с кэшированием:

$$C_{\text{total}} = N * (p * C_{\text{cache}} + (1 - p) * C_{\text{miss}})$$

1.2.2. Без кэширования

Каждый запрос обрабатывается без кэширования, и его стоимость равна:

$$C_{\text{base}} = c_q * \left(t_{\text{input}} + t_{\text{output}}\right)$$

Общая стоимость без кэширования:

$$C_{\text{base\_total}} = N * C_{\text{base}}$$

1.3. Экономия токенов

Экономия токенов (E) — это разница между стоимостью без кэширования и стоимостью с кэшированием:

$$E = C_{\text{base\_total}} - C_{\text{total}}$$

Подставляем выражения:

$$E = N * C_{\text{base}} - N * (p * C_{\text{cache}} + (1 - p) * C_{\text{miss}})$$

Упрощаем:

$$E = N * (C_{\text{base}} - p * C_{\text{cache}} - (1 - p) * C_{\text{miss}})$$

1.4. Итоговая формула экономии

Подставляем выражения для C_{cache}, C_{miss} и C_{base}:

$$E = N * \left(c_q * \left(t_{\text{input}} + t_{\text{output}}\right) - p * c_e * t_{\text{input}} - (1 - p) * \left(c_e * t_{\text{input}} + c_q * \left(t_{\text{input}} + t_{\text{output}}\right)\right)\right)$$

Раскрываем скобки:

$$E = N * \left(c_q * t_{\text{input}} + c_q * t_{\text{output}} - p * c_e * t_{\text{input}} - c_e * t_{\text{input}} - c_q * t_{\text{input}} - c_q * t_{\text{output}} + p * c_e * t_{\text{input}} + p * c_q * t_{\text{input}} + p * c_q * t_{\text{output}}\right))$$

Сокращаем:

$$E = N * \left(p * c_q * t_{\text{input}} + p * c_q * t_{\text{output}} - c_e * t_{\text{input}}\right)$$

1.5. Компактная формула

Итоговая формула экономии:

$$E = N * \left(p * c_q * \left(t_{\text{input}} + t_{\text{output}}\right) - c_e * t_{\text{input}}\right)$$