Очень примерная возможная оценка стоимости генерации SMS‑сообщений на 2025 г.:

**На текущий момент:**

Для одного SMS:

|  |  |  |
| --- | --- | --- |
| **Действие** | **Input, токенов** | **Output, токенов** |
| Создание неперсонализированного сообщения | 700 | 60 |
| Создание персонализированного сообщения | 300 | 60 |
| Выполнение проверок (примем 3 проверки) | 3 проверки × 300 | 3 проверок × 20 |
| Направленная перегенерация (примем в среднем 10 раз) | 10 раз × 400 | 10 раз × 60 |
| Проверки после перегенераций | 10 раз × 3 проверки × 300 | 10 раз × 3 проверки × 20 |
| **Итого** | 14 900 | 1 380 |

Для GigaChat одинаковый тариф для input и output токенов, поэтому получаем около 16 280 токенов на создание одного SMS‑сообщения, что по тарифу 0,75 руб. за 1000 токенов (для используемой модели GigaChat Pro) составляет 12,21 рублей за сообщение. В текущей реализации мы создаем 3 сообщения, чтобы редактор мог выбрать лучшее – т.е. затраты на одну группу клиентов составят около 12,21×3 = 36,63 рублей.

Определим примерное количество групп клиентов в рамках одной кампании. Предположим, что выполняем персонализацию по 4 параметрам – пол (мужской, женский), психотип (конструктор, рефлектор, центрист), поколение (X, Y, Z), отрасль (23 отрасли). Тогда в одной кампании у нас 2×3×3×23 = 414 групп клиентов, для каждой из которых требуется создать уникальное SMS-сообщение. Тогда стоимость одной кампании составит 36,63×414 = 15 165 рублей.

Если запускаются около 6 кампаний в месяц, то суммарная стоимость составит около 15 165 × 6 × 12 = 1,09 млн руб. в год.

**После проведения оптимизации:**

В настоящее время проводятся исследования по оптимизации процесса создания персонализированных SMS-сообщений – основными направлениями являются отказ от промежуточной стадии создания неперсонализированного SMS‑сообщения, сокращение промпта, переход от проверок с помощью LLM и морфологических анализаторов к проверкам только с помощью морфологических анализаторов. В таком случае для создания одного SMS будет использовано следующее примерное количество токенов:

|  |  |  |
| --- | --- | --- |
| **Действие** | **Input, токенов** | **Output, токенов** |
| Создание персонализированного сообщения | 600 | 60 |
| Направленная перегенерация (примем в среднем 10 раз) | 10 раз × 650 | 10 раз × 60 |
| **Итого** | 7 100 | 660 |

В таком случае получаем 7 760 токенов на создание одного сообщения, что приводит к затратам около:

* 5,82 рублей на одно сообщение;
* 17,46 рублей на 3 сообщения;
* 7 228 рублей на 1 кампанию для 414 групп клиентов;
* 0,52 млн руб. в год при 6 кампаниях в месяц.

Однако данные затраты можно сократить еще вдвое, если использовать тариф GigaChat Pro в асинхронном режиме (0,375 рублей за 1000 токенов). Также, исходя из результатов предварительных экспериментов, fine-tuning (LoRa) может позволить сократить количество необходимых перегенераций для получения SMS-сообщения без ошибок в среднем где-то в 2,5 раза – в таком случае можно ожидать дополнительного снижения расходов.

Таким образом, примерные затраты в токенах и рублях можно представить в следующих таблицах:

Примерные затраты в токенах:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Объем / реализация | Текущая реализация | После оптимизации (отказ от создания неперсонализированного SMS, сокращение промпта, переход с LLM проверок на морфологические) \* | После оптимизации и перехода на тариф с асинхронным режимом \* | После оптимизации, перехода на тариф с асинхронным режимом, выполнения fine-tuning (LoRa) \* |
| 1 SMS | 16,28 тыс. токенов | 7,76 тыс. токенов | 7,76 тыс. токенов | 3,5 тыс. токенов |
| Набор из 3 SMS | 48,84 тыс. токенов | 23,28 тыс. токенов | 23,28 тыс. токенов | 10,5 тыс. токенов |
| 1 кампания (414 наборов из 3 SMS (для 414 групп клиентов – 2 пола × 3 поколения × 3 психотипа × 23 отрасли)) | 20,22 млн токенов | 9,64 млн токенов | 9,64 млн токенов | 4,35 млн токенов |
| 6 кампаний в месяц в течение года | 1,456 млрд токенов | 693,93 млн токенов | 693,93 млн токенов | 312,98 млн токенов |

\* оценки являются предварительными (в настоящее время проводятся исследования)

Примерные затраты в рублях:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Объем / реализация | Текущая реализация | После оптимизации (отказ от создания неперсонализированного SMS, сокращение промпта, переход с LLM проверок на морфологические) \* | После оптимизации и перехода на тариф с асинхронным режимом \* | После оптимизации, перехода на тариф с асинхронным режимом, выполнения fine-tuning (LoRa) \* |
| 1 SMS | 12,21 руб. | 5,82 руб. | 2,91 руб. | 1,31 руб. |
| Набор из 3 SMS | 36,63 руб. | 17,46 руб. | 8,73 руб. | 3,94 руб. |
| 1 кампания (414 наборов из 3 SMS (для 414 групп клиентов – 2 пола × 3 поколения × 3 психотипа × 23 отрасли)) | 15 165 руб. | 7 228 руб. | 3 614 руб. | 1 630 руб. |
| 6 кампаний в месяц в течение года | 1,09 млн руб. | 0,52 млн руб. | 0,26 млн руб. | 0,12 млн руб. |

\* оценки являются предварительными (в настоящее время проводятся исследования)