Итоги A/B тестов промптов (Gemini)

Что тестировали

Три варианта запросов/промптов:

1. Prod A (сомнолог, строгая привязка к метрикам)

- Промпт: Ты — эксперт по сну. Твоя цель — дать понятный план улучшения сна на ближайшие 7 дней.

Ограничения:

- Не здоровайся и не представляйся.
- Не используй Markdown.
- Не назначай лекарства и не ставь диагнозы.
- Не выдумывай значения, опирайся только на входные показатели.

Что сделать:

1. Объясни простыми словами, что сейчас “не так” с последней ночью и какой показатель важнее всего улучшить в первую очередь.
2. Сравни последнюю ночь с предыдущей минимум по двум параметрам и сделай вывод о тенденции.
3. Дай план на 7 дней: что менять вечером, что делать утром, и что отслеживать.
4. Заверши 2–3 уточняющими вопросами к пользователю, которые реально помогут улучшить рекомендации (например про время отхода ко сну, экраны, дневной сон, кофеин).

Формат: 4 абзаца, без списков.

user_prompt

1. Prod B (клинически осторожный)

- Промпт: Ты — эксперт по сну. Сформируй рекомендации максимально клинически корректно и осторожно.

Ограничения:

- Не здоровайся, не представляйся.
- Не используй Markdown.
- Не назначай медикаменты.

- Не выдумывай факты, используй только метрики из входа.

Что сделать:

1. Интерпретируй последнюю ночь: достаточно ли сна по длительности, признаки прерывистости сна по фрагментации/пробуждениям, возможный вклад стресса/режима (только как гипотеза).
2. Сравни последнюю и предыдущую ночь и объясни, какие изменения наиболее важны и почему.
3. Дай рекомендации в стиле "сначала базовые вмешательства": режим сна, свет/экраны, нагрузка/кофеин, условия спальни, релаксация, дневной сон (если релевантно).
4. Добавь "когда стоит обратиться к врачу": только триггеры из данных (например очень короткая длительность сна регулярно, высокая фрагментация, выраженная дневная сонливость — как уточняющий вопрос).

Формат: 3–5 абзацев, без списков.

1. Baseline (7-day plan)

- Промпт: Ты — эксперт по сну. Твоя цель — дать понятный план улучшения сна на ближайшие 7 дней.

Ограничения:

- Не здоровайся и не представляйся.
- Не используй Markdown.
- Не назначай лекарства и не ставь диагнозы.
- Не выдумывай значения, опирайся только на входные показатели.

Что сделать:

1. Объясни простыми словами, что сейчас "не так" с последней ночью и какой показатель важнее всего улучшить в первую очередь.
2. Сравни последнюю ночь с предыдущей минимум по двум параметрам и сделай вывод о тенденции.
3. Дай план на 7 дней: что менять вечером, что делать утром, и что отслеживать.
4. Заверши 2–3 уточняющими вопросами к пользователю, которые реально помогут улучшить рекомендации (например про время отхода ко сну, экраны, дневной сон, кофеин).

Формат: 4 абзаца, без списков.

Важно: baseline в логах — это не "старый" промпт, а отдельный формат (7-day plan), поэтому сравнение с prod A/B — это сравнение разных стилей выдачи, а не только "оптимизации пары строк".

Дизайн теста

- Модель: Gemini 2.5 Flash/Fast.
- Единица: 1 запрос (последняя ночь + предыдущая + ещё одна ночь).
- Выборка по логам:
  - Prod A: N=6
  - Prod B: N=5
  - Baseline : N=3

Метрики качества

1) Latency и Cost

Метрика	Prod A (N=6)	Prod B (N=5)	Baseline (N=3)
Mean latency, s	14.91	13.50	11.29
Median latency, s	14.35	13.07	10.56
Mean cost, $	0.001695	0.002136	0.001726
Median cost, $	0.001610	0.002217	0.001743

Вывод по перформансу:

- Самый быстрый по медиане — baseline (10.56s), затем prod B (13.07s), затем prod A (14.35s).
- Самый дешевый по медиане — prod A (~~0.00161$), затем baseline (~~0.001743$), затем prod B (~0.002217$).

2) Compliance (по наблюдаемым нарушениям формата)

Проверялось то, что можно увидеть в тексте ответов из логов:

- запрет приветствий — во всех примерах соблюдён (ответы начинаются сразу с анализа).
- запрет диагнозов/лекарств — в примерах не найдено назначений лекарств или постановки диагноза (только общие рекомендации).
- запрет Markdown — найдено как минимум одно явное нарушение в baseline: в ответе встречается выделение **...** (маркер Markdown bold).

Оценка "No-Markdown compliance" на имеющихся примерах:

- Prod A: 6/6 = 100%
- Prod B: 5/5 = 100%

- Baseline: 2/3 = 66.7% (1 ответ содержит **жирное выделение**)

Winning промпт и почему он лучше

Если цель — "понятный план на 7 дней" и вовлечение пользователя вопросами, то baseline (7-day plan) структурно лучше подходит под UX-задачу (там есть план + вопросы).
Если цель — "максимально строгая привязка к метрикам и профессиональный стиль сомнолога", то prod A выглядит наиболее стабильным по compliance и дешевле по cost, но он не обязан задавать вопросы (в текущем формулировании).
Если цель — "клиническая осторожность + триггеры обращения к врачу", то prod B покрывает safety-требования лучше, но он дороже по cost.

Практическая рекомендация:

- Для пользовательского режима приложения (простые действия + вопросы): использовать baseline-формат, но усилить compliance (убрать Markdown-жирность и жёстче валидировать 4 абзаца).
- Для режима "клиническая справка/осторожный совет": prod B.
- Для режима "анализ метрик как сомнолог + чёткие индикаторы прогресса": prod A.

Improvement после оптимизации (что можно утверждать по фактам)

Можно утверждать только про latency/cost и видимые нарушения формата:

- Prod A дешевле prod B по медианной стоимости (0.00161$ vs 0.002217$).
- Baseline быстрее prod A/B по медианной задержке (10.56s vs 14.35s/13.07s).
- Baseline имеет риск нарушения запрета Markdown (найдено 1/3).