До сравнительного испытания

Пункт	Отметка
Выбраны конкретные рабочие задачи	[ ]
Записаны версии и режимы кандидатов	[ ]
Скрытые параметры отмечены	[ ]
Исходники и требования одинаковы	[ ]
Известен эталон каждой задачи	[ ]
Настройка не смешана с первым запуском	[ ]
Повторы начинаются с чистого контекста	[ ]
Есть коды для слепой оценки	[ ]
Данные разрешены для использования	[ ]

Кандидаты / дата / задачи	Ограничения условий

До выбора и оплаты

Пункт	Отметка
Учтены все попытки, а не удачные скриншоты	[ ]
Готовность до и после правок разделена	[ ]
Ручное время относится к тому же объёму	[ ]
Расходы имеют основание и период	[ ]
Нет сравнения выдуманных тарифов	[ ]
Отдельно отмечено время ожидания	[ ]
Решение проверено на новых заданиях	[ ]
Определены границы применения и ручной контроль	[ ]

Решение / основание / остаток	Когда пересмотреть

M761
А: 4/6 сразу и 20 минут правок; Б: 5/6 и 14 минут. У обоих после правок 6 принятых результатов и условные полные затраты 260 ₽.

Источники и проверка

Anthropic: оценка под задачу и несколько критериев