



Банк России

РАЗРАБОТКА СИНТЕТИЧЕСКОЙ ФОКУС-ГРУППЫ ДЛЯ ПРЕДВАРИТЕЛЬНОГО ТЕСТИРОВАНИЯ КОММУНИКАЦИИ НА ОСНОВЕ LLM

АННА ГУСЕВА, ДДКП
АЛИНА ЕВСТИГНЕЕВА, ДДКП

Июнь 2026 г.

Цель работы

Разработать, описать и **первично валидировать прототип синтетической фокус-группы** на основе больших языковых моделей и мультиагентной архитектуры для предварительного тестирования текстовых материалов

Исследовательский вопрос

Может ли синтетическая фокус-группа использоваться как **вспомогательный инструмент ранней диагностики** рисков до публикации текста?

Решение нескольких задач:

1. Формулируется **методологическая рамка** применения синтетической фокус-группы в коммуникации центрального банка и **фиксируются границы ее интерпретации**.
2. Описывается **архитектура системы**.
3. Разрабатывается **валидационный контур**, позволяющий отдельно проверять агентов-участников, модуль резюмирования и модуль риск-оценки.
4. Проводится **серия тестов** на профильную согласованность, речевую различимость, устойчивость к техническим вариациям, доказательную привязку выводов, применимость рекомендаций, калиброванность риск-шкалы и отсутствие страховочного завышения риска.
5. Формулируются **ограничения** текущей версии системы и **направления дальнейшей доработки**.

Один и тот же материал может быть прочитан по-разному – между замыслом автора и реакцией аудитории есть **слой интерпретации**

Авторский замысел

- что хотели объяснить
- какой смысл / сигнал заложили
- какой вывод ожидали



Слой интерпретации

- язык текста
- личный опыт
- доверие / недоверие
- медиа-контекст



Возможная реакция

- понятно / непонятно
- убедительно / неубедительно
- спокойно / тревожно
- точно / искаженно

Между корректным текстом и его восприятием возникает слой интерпретации

Исследовательская и прикладная задача: заранее увидеть места, где текст может быть понят не так, как предполагалось

Корректное прочтение результатов:
гипотезы о возможном прочтении
текста

Что показывает

- **возможные траектории прочтения**
- зоны непонимания
- риск-факторы

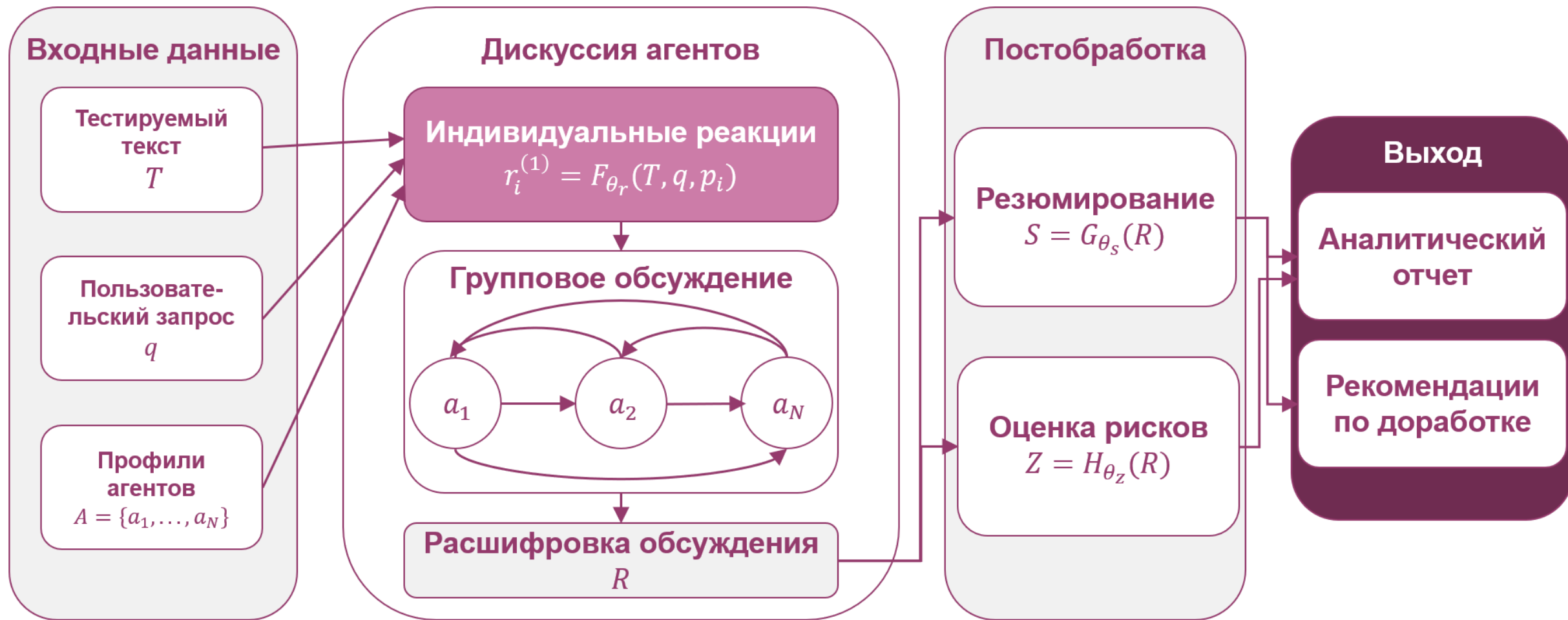
Что **НЕ** показывает

- доли мнений в аудитории
- фактическую реакцию людей
- автоматическое решение

Как использовать

- как материал для экспертной доработки
- как **ранний сигнал**, где текст требует внимания

Направление	Источники	Что изучено	Пробел для текущей задачи
Коммуникация ЦБ с широкой аудиторией	Baerg et al. (2025); Blinder et al. (2024); Coibion et al. (2022); Lamla, Vinogradov (2019)	Сообщение ЦБ проходит через ограниченное внимание, медиа-посредников, личный опыт и доверие к институту. Широкая аудитория воспринимает сигнал иначе, чем профессиональные участники рынка	Основной фокус – опубликованная коммуникация и ее восприятие. Менее разработан вопрос, как выявлять риск неверного прочтения до публикации
Понятность, тон и восприятие экономических текстов	Евстигнеева, Сидоровский (2021); Евстигнеева, Щадилова (2024); Ерохина, Клачкова (2024); Bholat et al. (2019); Bulir et al. (2013); Feldkircher, Makridis (2025)	Ясность, синтаксическая сложность, тон и связь с повседневным опытом влияют на понимание, доверие и ожидания. «Хороший текст» зависит от аудитории, жанра и контекста	Нужна процедура, которая проверяет не только стиль и удобочитаемость, но и траектории интерпретации : где текст может быть понят неоднозначно
Синтетические респонденты и моделирование мнений на основе больших языковых моделей	Agarwal (2025); Aher et al. (2023); Argyle et al. (2023); Korinek (2023); Morocho et al. (2026)	LLM могут воспроизводить отдельные паттерны ответов подгрупп и результаты некоторых экспериментов, но склонны к усредненности, гиперточности и чувствительности к профилю	Синтетические респонденты не дают оснований для оценки долей мнений или фактической реакции людей. Их корректный статус – генерация гипотез о возможном прочтении текста
LLM-агенты и мультиагентные системы	Guo et al. (2024); Korinek (2025); Luo et al. (2025); Pacioni et al. (2026); Park et al. (2023)	Поведение агента задается не только моделью, но и профилем, памятью, протоколом, коммуникацией с другими агентами и оркестрацией системы	Для синтетической фокус-группы нужна явная архитектура: профили, индивидуальная реакция, групповая итерация, постобработка и воспроизводимый протокол
Валидация, смещения и надежность синтетических респондентов	Bisbee et al. (2024); Dominguez-Olmedo et al. (2024); Morocho et al. (2026); Qu, Wang (2024)	Ответы LLM чувствительны к промпту, порядку и маркировке вариантов, языку, стране, демографии и теме; синтетические данные могут терять вариативность	Нужен отдельный валидационный контур : профильная согласованность, речевая различимость, устойчивость, контроль экспертности, демографические смещения и групповое влияние
Оценка резюмирования и риск-анализа	Chen et al. (2021); DeYoung et al. (2020); Fabbri et al. (2021); Guo et al. (2017); Liu et al. (2023); Mathew et al. (2021); Maynez et al. (2020); Wang et al. (2023); Ziems et al. (2024)	Резюме может быть связным, но фактически неподтвержденным. LLM-as-judge масштабирует оценку, но сам требует контроля. Риск-оценка должна быть калибрована и доказательно привязана к аргументам	Некорректно валидировать только финальный отчет. Нужно отдельно проверять точность, полноту тем, рекомендации, калибровку шкалы риска, чувствительность/специфичность и интерпретируемость обоснований



Единица анализа: не мнение агента само по себе, а воспроизводимая процедура обработки текста

Компонент	Роль в системе	
Пользовательский интерфейс	загрузка текста, выбор агентов, постановка вопроса	пользовательское взаимодействие
API-сервис	формирование задачи и запуск процедуры	
Сервис фокус-группы	управление итерациями обсуждения	управление задачами
Агенты-респонденты	индивидуальные и групповые реакции	
Модуль резюмирования	темы, проблемные формулировки, рекомендации	модельный контур
Модуль анализа рисков	риск-оценка, ключевые факторы, обоснование	
Хранилища	профили, файлы, история, результаты	контур хранения

Модели:

- Агенты-респонденты: Gemma3-27B
- Резюмирование и риск-анализ: Qwen3.5-32B

Глубинные интервью

- реальные респонденты
- экономический опыт
- доверие
- медиа-потребление
- речевой стиль

Профили агентов

- «паспортички» (возраст, профессия и т.д.)
- финансовая грамотность
- отношение к институтам
- опыт инфляции и кредитов
- манера речи

Доверяющие

- более высокий уровень доверия к органам власти
- системный стиль рассуждения

Прагматичные

оценка текста через:

- личный опыт,
- семейный бюджет,
- профессиональную практику
- личные наблюдения за ценами

Скептические

- низкое доверие
- высокая чувствительность к теме цен, кредитов, реальных доходов и социальной справедливости

*Профили не претендуют на репрезентативность
Их задача – задать контрастные типы прочтения текста*

Проверяется не поведенческая эквивалентность с реальными респондентами, а функциональная валидность компонентов

Вопрос валидации: выполняет ли каждый модуль свою прикладную функцию в процедуре предварительного тестирования текста

T0–T8

Агенты-участники

Функция в системе

Формируют индивидуальные и групповые реакции от лица заданных профилей и по конкретному тексту

Что проверяем

- согласованность с профилем
- речевая различимость
- устойчивость и переформулировки
- контроль экспертности
- групповое влияние и смещения

S1–S5

Модуль резюмирования

Функция в системе

Преобразует расшифровку обсуждения в темы, проблемные формулировки и практические рекомендации.

Что проверяем

- фактологическая верность
- полнота и релевантность тем
- применимость рекомендаций
- устойчивость к шуму и перестановкам
- контрольные сценарии

R1–R6

Модуль оценки рисков

Функция в системе

Оценивает уровень риска публикации и выделяет обоснованные риск-факторы

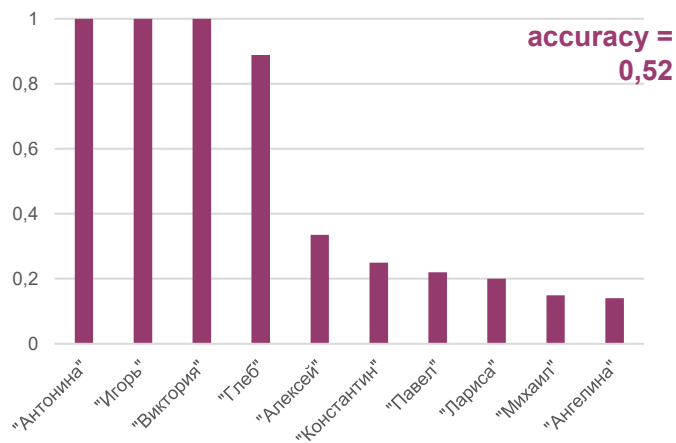
Что проверяем

- калибровка шкалы 0-100
- доказательная привязка
- чувствительность и специфичность
- устойчивость к модификациям
- интерпретируемость обоснований

T0-T2 проверяют три базовых свойства агентного блока: простая оценочная калибровка, соответствие профилю и воспроизводимость реакции

T0 Простая оценочная калибровка

ТОЧНОСТЬ ПО ОТДЕЛЬНЫМ АГЕНТАМ



Простая схема «ожидаемая / полученная оценка» не описывает реакцию на конкретный текст: доверие к институту не означает автоматического одобрения материала.

T1 Согласованность с профилем

УСРЕДНЕННЫЕ РЕЗУЛЬТАТЫ ТЕСТА T1



T2 Внутренняя устойчивость

Повторные запуски одного текста и вопроса

Доля доминирующей позиции	1,0
Полностью стабильные пары	1,0
Энтропия позиции	0,0
Ст. откл. доверия	0,0
Ст. откл. понятности	0,0

Профильные метрики в T2 сохраняются высокими:

профиль **4,72 / 5** · стиль **4,48 / 5**
· привязка к тексту **4,98 / 5**

Источник: расчеты авторов

Агенты удерживают профиль и привязку к тексту; простая оценочная метка недостаточна; повторяемость подтверждает техническую устойчивость, но не заменяет человеческую вариативность

T3–T5 проверяют, сохраняют ли агенты индивидуальную речевую оптику и не используют ли экспертность базовой модели сверх заданного профиля

T3 Речевая правдоподобность

Стиль соответствует профилям; признаки модельного вмешательства низкие

Соответствие
стилю
4,52 / 5

Модельное
вмешательство
0,20
средние значения

T4 Слепое восстановление профиля

Модель-судья определяет автора без имени и метаданных

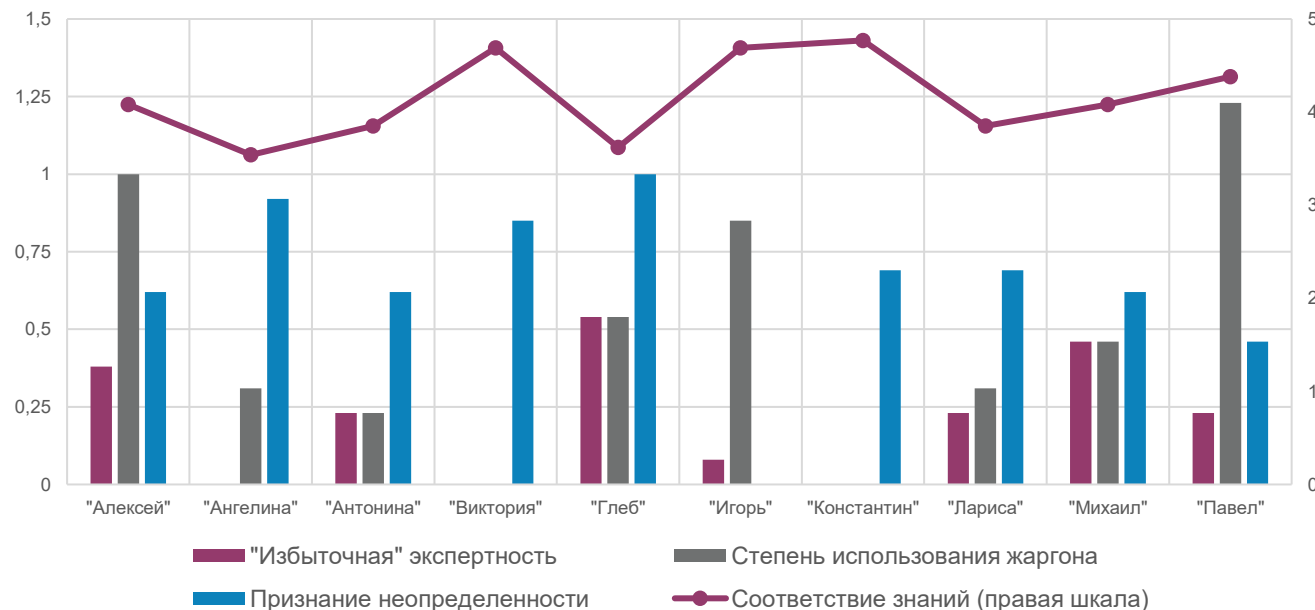
случайный уровень



Утечка профиля = 18,3%
Различимость выше случайного уровня

T5 Контроль экспертности

РЕЗУЛЬТАТЫ ОЦЕНКИ ТЕСТА T5 ПО АГЕНТАМ



Соответствие
знаний
4,15 / 5

Избыточная
экспертность
0,215

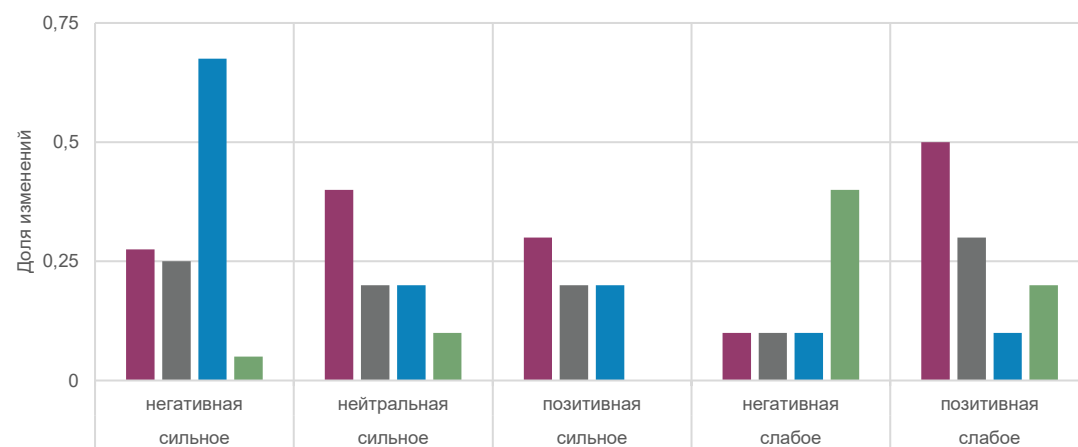
Признание
неопред.
0,646

Источник: расчеты авторов

Агенты различимы по речи и логике реакции; уровень знаний в целом соблюдается, но агенты с академическим или профессиональным профилем требуют отдельного контроля экспертных «утечек»

Т6-Т8 проверяют реакцию на мнения других участников, демографические модификации и переформулировку вопроса

Т6 Устойчивость к групповому давлению



Позиция большинства (негативная / нейтральная / позитивная) и качество аргумента (сильное / слабое)

■ Изменение позиции ■ Сдвиг к большинству ■ Аргументированная ревидия ■ "Слепое" согласие

средние значения

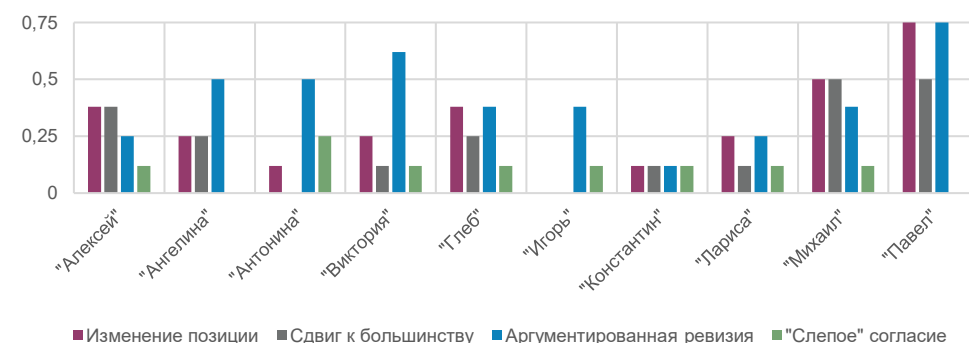
Изменение позиции
30%

Аргумент. ревидия
41,2%

«Слепое» согласие
11,25%

Сдвиг к большинству
22,5%

Т6 Различия между агентами



■ Изменение позиции ■ Сдвиг к большинству ■ Аргументированная ревидия ■ "Слепое" согласие

Т7 Демографическая устойчивость

Сдвиг позиции
0

Стереотипизация
0

Изменение возраста или профессии не меняло позицию и содержательное ядро ответа

Т8 Переформулировка

Доля доминир. позиции
90,5%

Полностью стабильные группы
68%

Источник: расчеты авторов

Групповое влияние присутствует, но не превращается в массовый конформизм; позиции устойчивы к переформулировке, явной стереотипизации не выявлено

Валидация показывает пригодность агентов для ограниченной задачи предварительной диагностики, но не поведенческую эквивалентность людям

Подтверждено

- Ответы связаны с конкретным текстом, а не с темой в целом
- Профили удерживаются по поведению, стилю и аргументации
- Синтетическая группа не сводится к одному усредненному респонденту
- Групповая итерация дает ограниченную, контролируемую динамику

Ограничения

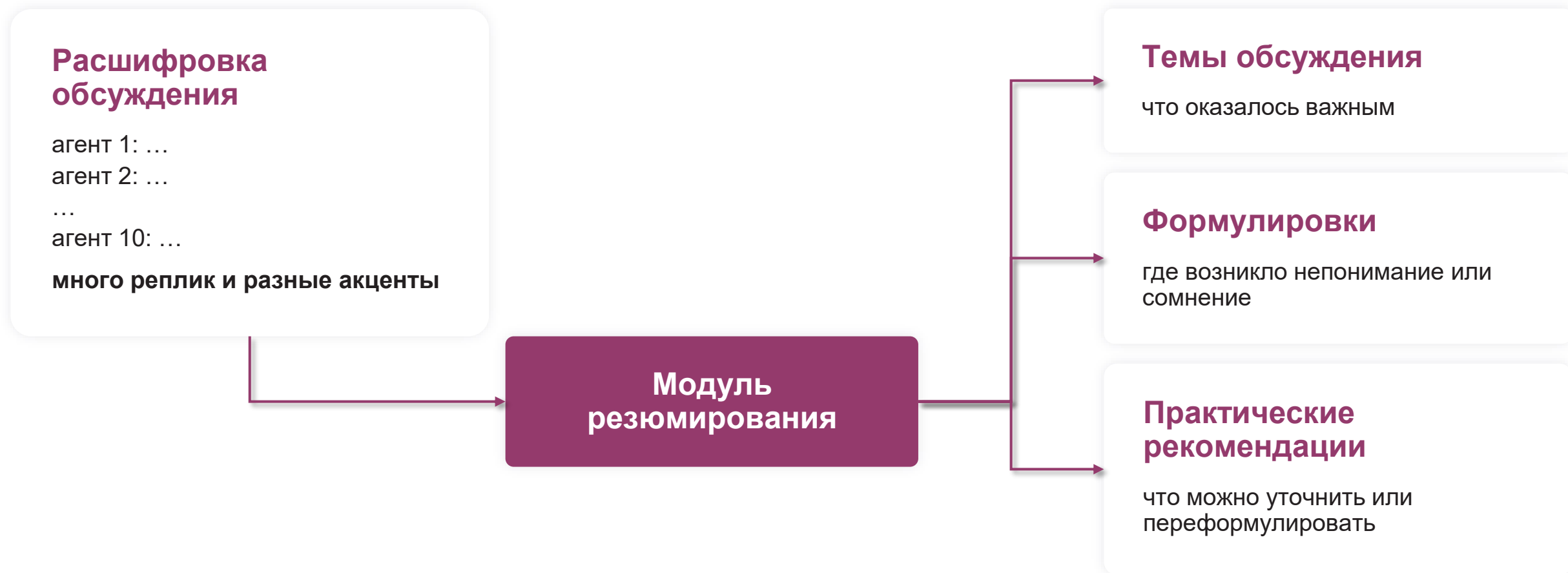
- Простая оценочная калибровка плохо описывает реакцию на конкретный текст
- Полная повторяемость в T2 снижает поведенческую реалистичность
- У отдельных агентов возможны экспертные «утечки»
- Демографическую устойчивость нужно расширять за счет новых модификаций

Статус вывода

Функциональная валидность
подтверждена

- Агенты пригодны как профильно привязанные синтетические респонденты
- Их ответы нельзя интерпретировать как фактические мнения людей

Итог: *агентный блок проходит компонентную проверку для задачи предварительного тестирования текста, но требует дальнейшей настройки вариативности и экспертности*



Результат – структурированный слой анализа для эксперта

Что проверяли	Результат	Интерпретация
Фактологическая верность	неподтвержденные выводы: 13,4% покрытие доказательствами: 85,9%	есть потенциал для улучшения
Полнота и релевантность	полнота тем: 89,7% точность тем: 96,5%	ключевые темы сохраняются
Практическая применимость	4,41 / 5	рекомендации можно использовать для доработки текста
Устойчивость	темы: 90,2% рекомендации: 85,6%	аналитическое ядро сохраняется при модификациях
Контрольные сценарии	99,3%	модуль распознает тип ситуации



60 контрольных обсуждений



7 сценариев проверки



240 модифицированных версий
обсуждений



*Модуль сохраняет ключевые темы
обсуждения и редко добавляет
нерелевантные выводы*



IN: расшифровка обсуждения

Риск-оценка 0-100:



OUT:

- числовая оценка
- обоснование
- ключевые факторы риска

Оценка помогает понять, требует ли текст дополнительного внимания эксперта

Тестовый дизайн

90

риск-сценариев

5

диапазонов риска

8

типов сценария

5

типов модификации

120

модификаций

Ключевые результаты

100%

полнота выявления
повышенных и высоких
рисков

0 случаев

низкорисковые
сценарии не получили
оценку выше 40 / 60

≤2,5 пункта

среднее изменение
оценки после
модификаций входа

90%

оценки попали в
ожидаемый диапазон

Источник: расчеты авторов

*Модуль полезен для предварительной диагностики, но
выводы сверяются с экспертом и самим текстом*



Синтетическая фокус-группа может рассматриваться как вспомогательный инструмент ранней диагностики коммуникационных материалов до публикации

Что показала валидация

- Агенты-участники **удерживают заданные профили**, отвечают на конкретный текст, различаются по стилю и типу аргументации, **демонстрируют контролируемую групповую динамику**
- Модуль резюмирования сохраняет ключевые темы обсуждения, фиксирует позиции меньшинства и формирует практически применимые рекомендации
- Модуль оценки рисков различает низкие и высокие уровни риска, не демонстрирует страховочного завышения и устойчив к поверхностным изменениям входа

Система помогает структурировать возможные траектории восприятия текста, выявлять слабые формулировки и формировать гипотезы о рисках публикации

Работа показывает внутреннюю согласованность и функциональную валидность компонентов, но не доказывает совпадение синтетических реакций с реакциями реальных людей

Внешняя валидность

Система детально проверена на внутреннюю согласованность, но результаты пока не сопоставлены с данными реальных фокус-групп или экспериментальных предтестов

Следствие: результаты нельзя интерпретировать как фактическую реакцию аудитории

Статус синтетических агентов

Агенты не являются репрезентативной выборкой населения и не заменяют реальных респондентов

Следствие: корректный статус результата – гипотезы о возможных траекториях прочтения текста

Состав профилей

Профили заданы как контрастные типы восприятия, но состав группы влияет на выявляемые проблемы и рекомендации

Следствие: нужен анализ чувствительности результатов к набору агентов

Ограничения модулей

Модуль резюмирования требует усиления доказательной привязки отдельных выводов; риск-модуль может занижать слабые и пограничные риск-сигналы

Следствие: низкая риск-оценка не отменяет экспертный анализ локальных барьеров понимания

Необходимость экспертной интерпретации

Результаты являются гипотезами о возможном прочтении текста и не должны использоваться как автоматическое решение

Следствие: нужен анализ чувствительности результатов к набору агентов



Направления дальнейшего исследования

Сравнить результаты СФГ с реальными фокус-группами или экспериментальными предтестами на одних и тех же текстах

1

Эмпирическое сопоставление

Сформировать больше агентных профилей и сравнить устойчивость результатов

2

Профили агентов

Показывать, на каких фрагментах текста и репликах агентов основаны выводы резюме, рекомендации и риск-оценка

4

Проверка связности выводов

Модератор

5

Проверить, повышает ли отдельный агент-модератор качество групповой динамики без нарушения нейтральности вопросов и баланса участия

Запускать одну задачу при разных наборах агентов и сравнивать темы, рекомендации и риск-оценки

3

Конфигурация группы

Разделить устойчивые результаты и зоны, требующие дополнительной проверки

6

Интерпретация метрик



Банк России

СПАСИБО ЗА ВНИМАНИЕ

Гусева Анна Дмитриевна

Консультант

Отдел развития коммуникации ДДКП

Банк России

GusevaAD@cbr.ru

adguseva@hse.ru