

Интерпретация явлений естественного языка в LLM для задачи контролируемой генерации текстов

Персонификация LLM как задача контролируемой генерации текстов

О.А.Митрофанова

Кафедра математической лингвистики СПбГУ

o.mitrofanova@spbu.ru

oa-mitrofanova@yandex.ru

План

- **Задачи контролируемой генерации текстов**
- **Персонафикация LLM: соотношение синтетических персон и языковой личности говорящего**
- **Наборы данных и модели для работы с синтетическими персонами**
- **Исследование чувствительности LLM к социодемографическим и эмоциональным характеристикам синтетических персон: русскоязычные данные**

Задачи контролируемой генерации текстов

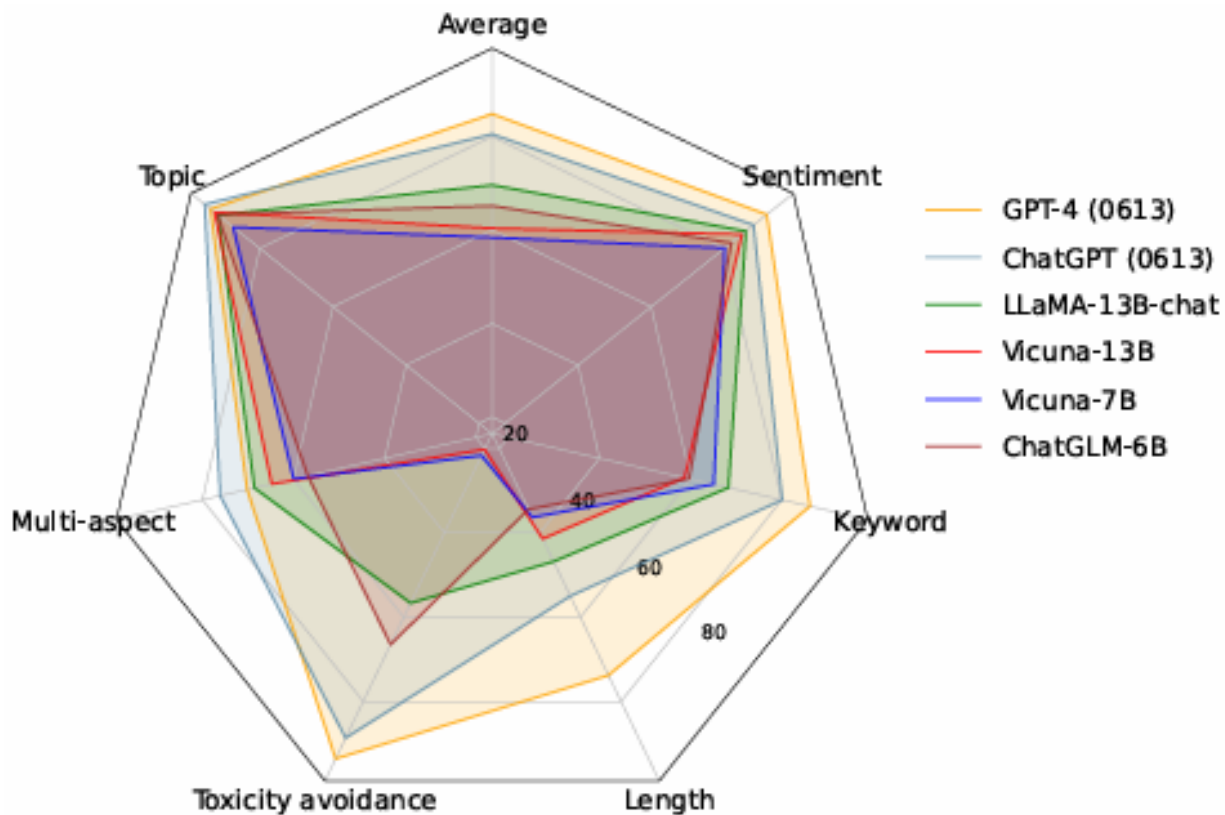
- **Определение контролируемой генерации текстов (controlled text generation)**
- Даны словарь $\mathcal{V}$, целевой текст $Y = \{y_1, y_2, \dots, y_n\}$, where $y_n \in \mathcal{V}$, контролируемый параметр C .
- Тогда процедуру CTG можно формально описать как $P(Y|C) = p(y_1, y_2, \dots, y_n|C)$
- Контролируемый параметр C может варьироваться в зависимости от различных задач.
- Ожидается, что текст Y , сгенерированный моделью CTG, будет удовлетворять условиям C , при этом максимально соответствуя общим характеристикам естественных текстов (правильнооформленность, осмысленность, коммуникативная корректность, читабельность и т.д.).
- *Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. ACM Comput. Surv. 56, 3, Article 64 (oct 2023), 37 pages. <https://doi.org/10.1145/3617680>*

Задачи контролируемой генерации текстов

- **Контроль структуры текстов:** заголовки, аннотации, рефераты, ключевые выражения, парафразы, адаптации по уровням сложности и т.д.
- **Контроль содержания текстов:** тематика, сюжет, субъективно-оценочный компонент.
- **Контроль словарного компонента:** лексико-семантические отношения, конструкции, фреймовые представления.
- **Контроль стиля текстов:** новость, отзыв, документ, художественная проза и т.д.
- **Промпты для контроля генерации:** zero-shot, few-shot, **role-based**
- **Аугментация данных** на основе контролируемой генерации текстов (библиотека для аугментации nlraug и ее русская версия ru_nlraug).
- **Оценка качества контролируемой генерации:** метрики, бенчмарки (метрика CTRLEval, ConGenBench с промптами для контроля токсичности, тональности и темы и наборами данных, CoDI-Eval с промптами для контроля токсичности, тональности, темы, ключевых выражений, длины и комбинации признаков, FOFO с промптами для контроля домена, поддомена и формата CTG).

Задачи контролируемой генерации текстов

- Бенчмарк CoDI-Eval с промптами для контроля токсичности, тональности, темы, ключевых выражений, длины текстов и комбинации признаков.
- *Chen, Y., Xu, B., Wang, Q., Liu, Y., & Mao, Z. (2024, March). Benchmarking large language models on controllable generation under diversified instructions. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 17808-17816).*



Персонафикация LLM: соотношение синтетических персон и языковой личности говорящего

- При обучении LLM **исчерпаны** ресурсы естественных текстов в электронном виде
- Нужны новые данные... Где их найти? Выход – **аугментация** данных
- Как это сделать?
- Тексты на ЕЯ специфичны по тематике, по стилю и жанрам, хронологии, идиостилю. Если обучаем модели на больших данных, идиостилистические признаки нивелируются, модель становится **обезличенной** (?).
- Важное требование к синтетическим данным: их лингвостатистические характеристики должны соответствовать реальным датасетам.
- *Kolmogorova, A. V. and Margolina, A. V. (2024) Written vs generated text: “naturalness” as a textual and psycholinguistic category, Research Result. Theoretical and Applied Linguistics, 10 (2), 71-99. DOI: 10.18413/2313 8912-2024-10-2-0-4*

Персонализация LLM: соотношение синтетических персон и языковой личности говорящего

- Помощь приходит из **UI/UX** дизайна
- **Agentic behavioral simulation, LLM personalization, Human-AI alignment**
- **Синтетические персоны, персонажи, личности = AI surrogates, Silicon Samples**
- *Hoq, A., & Weninger, T. (2026). Evaluating LLMs as Human Surrogates in Controlled Experiments. arXiv preprint arXiv:2604.15329.*
- Это удобно — дешево и быстро.
- **НО** при всем объеме опубликованного (десятки, сотни статей) до сих пор нет ответа, насколько синтетические персоны могут заместить человека, насколько они могут воспроизводить человеческое поведение.
- *Hullman, J., Broska, D., Sun, H. and Shaw, A. (2026). This human study did not involve human subjects: Validating llm simulations as behavioral evidence. arXiv preprint arXiv:2602.15785 .*

Персонификация LLM: соотношение синтетических персон и языковой личности говорящего

- **Языковая личность:** говорящий индивид, характеризуемый с точки зрения порождаемых текстов, отличающихся структурными, когнитивными, дискурсивными особенностями.
- История вопроса: *Вайсгербер Й.Л. Родной язык и формирование духа. М.: Едиториал УРСС, 2004.; Виноградов В.В. Избранные труды: О языке художественной прозы. М.: Наука, 1980.; Караулов Ю.Н. Русский язык и языковая личность. М., 2010.*
- **Лингвистическое профилирование:** сочетаемостные предпочтения, лексическое разнообразие, порядок слов, сложность синтаксической структуры, частотное распределение грамматических категорий, типов предложений и т.д., когнитивный тезаурус, соотносящийся с картиной мира, коммуникативные стратегии, прагматические задачи...
- **Социопсихологическое профилирование:** демографические данные (возраст, пол), социальные функции и род занятий, географическое положение, темы, представляющие интерес, психологические особенности и поведение, эмоциональные характеристики (valence, arousal...)

Персонификация LLM: соотношение синтетических персон и языковой личности говорящего

- **Профилирование говорящих на ЕЯ**

- ОРД (<https://ord.spbu.ru/>, СПбГУ, рук. Н.В.Богданова-Бегларян), КУРС (корпус устной речи студентов, <https://esc-corpus.ru/>, НИУ ВШЭ, рук. Т.Ю.Шерстинова)
- RusProfiling Lab (<https://rusidiolect.rusprofilinglab.ru/> Воронеж, рук. Т.А.Литвинова)
- МУРКО, МультиПАРКи (НКРЯ, <https://ruscorpora.ru/>)

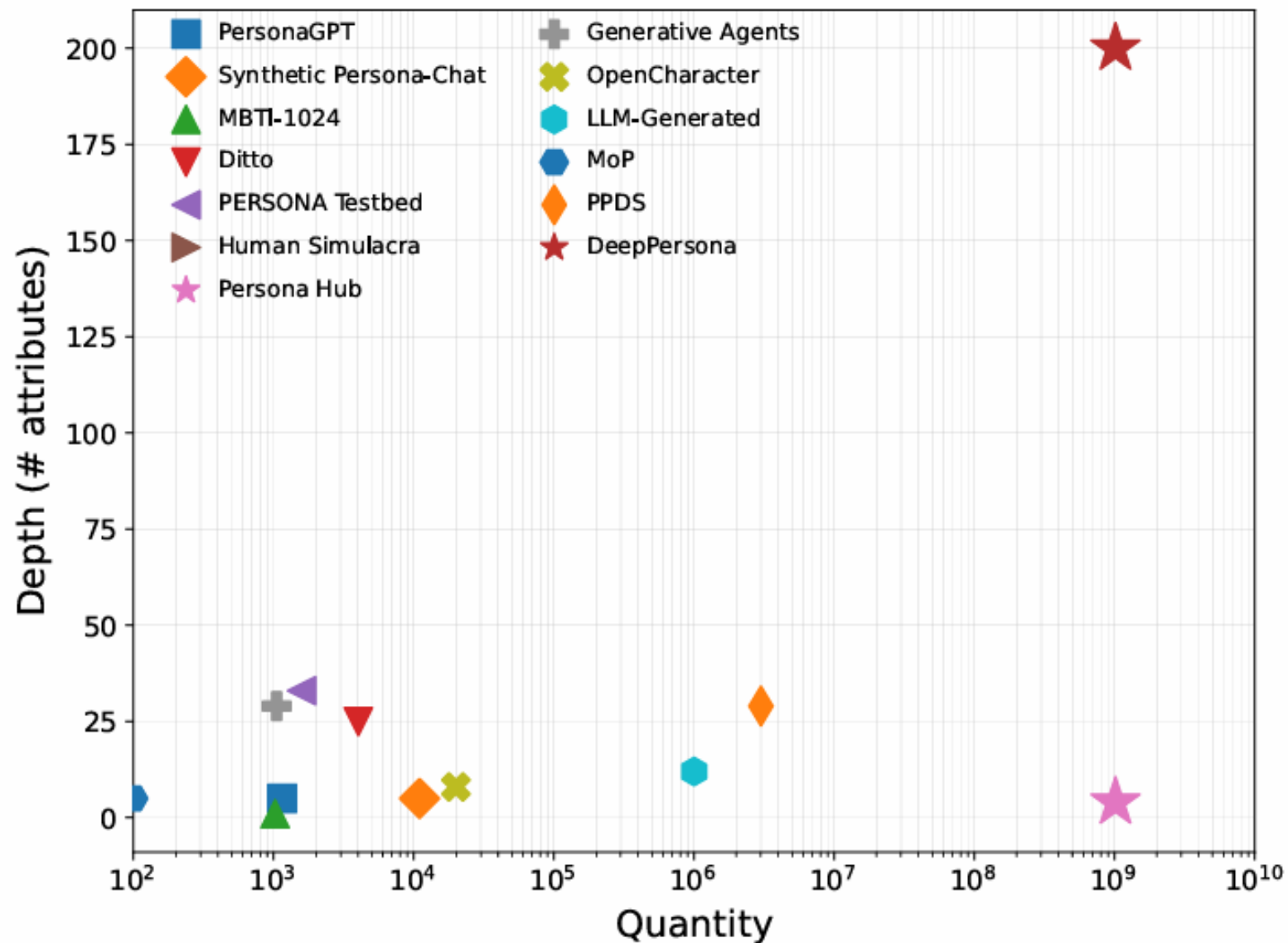
- **Профилирование персон:** вымышленные личности с вымышленным профилем:

- возраст, пол, жизненный опыт, ценности, социальное положение, доход, профессия, биография и т.д.
- имя, CV, фото и т.д.



Часть задач решается анализом в Profiling-UD [Linguistic Profiling Tool](http://linguistic-profiling.italianlp.it/)
<http://linguistic-profiling.italianlp.it/>

Наборы данных и модели для работы с синтетическими персонами



Проект PersonaGPT

<https://github.com/illidanlab/personaGPT>

- Цель – контролируемая генерация персонифицированных диалогов
- Модель – предобученная GPT с тонкой настройкой на наборе данных PersonaChat с разметкой истории диалогов и характеристик коммуникантов (несколько фактов из биографии, 5 коротких предложений длиной не более 15 слов)
- Метод сбора данных – краудсорсинг, всего 162 064 контекста
- Чат-бот получает минимально необходимые данные о контексте диалога и о коммуникантах, с учетом этого генерирует ответы
- Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018, July). *Personalizing dialogue agents: I have a dog, do you have pets too?*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2204-2213).
- Tang, F., Zeng, L., Wang, F., & Zhou, J. (2021). *Persona authentication through generative dialogue*. *arXiv preprint arXiv:2110.12949*.

Проект Synthetic PersonaChat

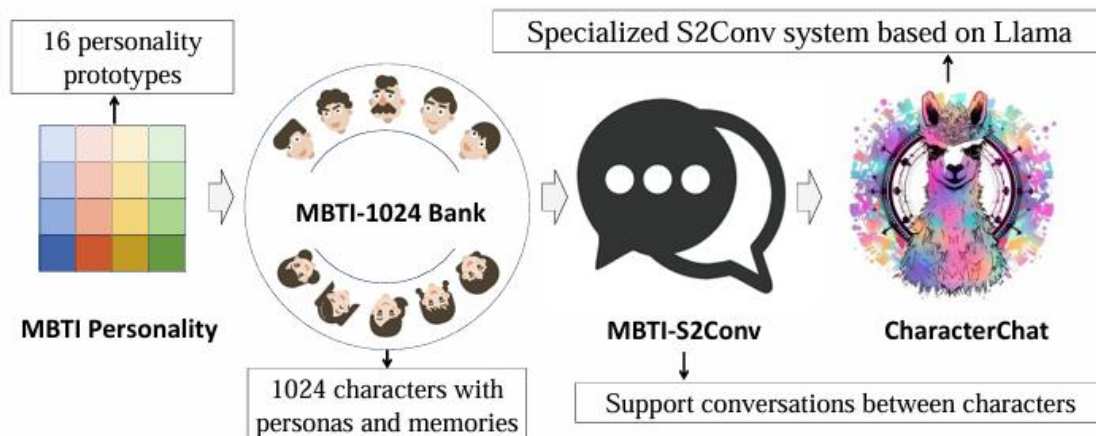
<https://github.com/google-research-datasets/Synthetic-Persona-Chat>

- Набор данных PersonaGPT + сгенерированные персоны и диалоги, совместимость персон, планирование их взаимодействия
- Общий объем около 10 тыс. персон и 22 тыс. диалогов.
- *Jandaghi, P., Sheng, X., Bai, X., Pujara, J., & Sidahmed, H. (2024, August). Faithful persona-based conversational dataset generation with large language models. In Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024) (pp. 114-139).*

Проект CharacterChat (MBTI-1024)

<https://github.com/morecry/CharacterChat>

- Фреймворк Social Support Conversation (S2Conv)
- Совместимость персон, планирование их взаимодействия
- Tu, Q., Chen, C., Li, J., Li, Y., Shang, S., Zhao, D., ... & Yan, R. (2023). Characterchat: Learning towards conversational ai with personalized social support. arXiv preprint arXiv:2308.10278.



A grid of 16 MBTI personality types arranged in two rows of eight. Each type is represented by a small cartoon character and its acronym. The top row includes INTJ, INFP, ENTJ, ENTP, INFJ, INFP, ENFJ, and ENFP. The bottom row includes ISTJ, ISFJ, ESTJ, ESFJ, ISTP, ISFP, ESTP, and ESFP. A purple triangle points down at the top center of the grid.

PROMPT

Here is the brief introduction of the given personality: ENTJ

They are decisive people who love momentum and accomplishment. They gather information to construct their creative visions creative visions but rarely hesitate for long before acting on them...

####

You are an **outstanding creator**, you can construct a variety of characters in the real world. Now, based on the given personality type, please design a virtual character according to the following given fields... It is necessary to ensure that some attribute information of the characters needs to be **distributed diversely, reasonably related, and in line with the laws of nature.**

Fill the result into JSON
{ 'Name': , 'Personality': , 'Growth Experience': , ... }

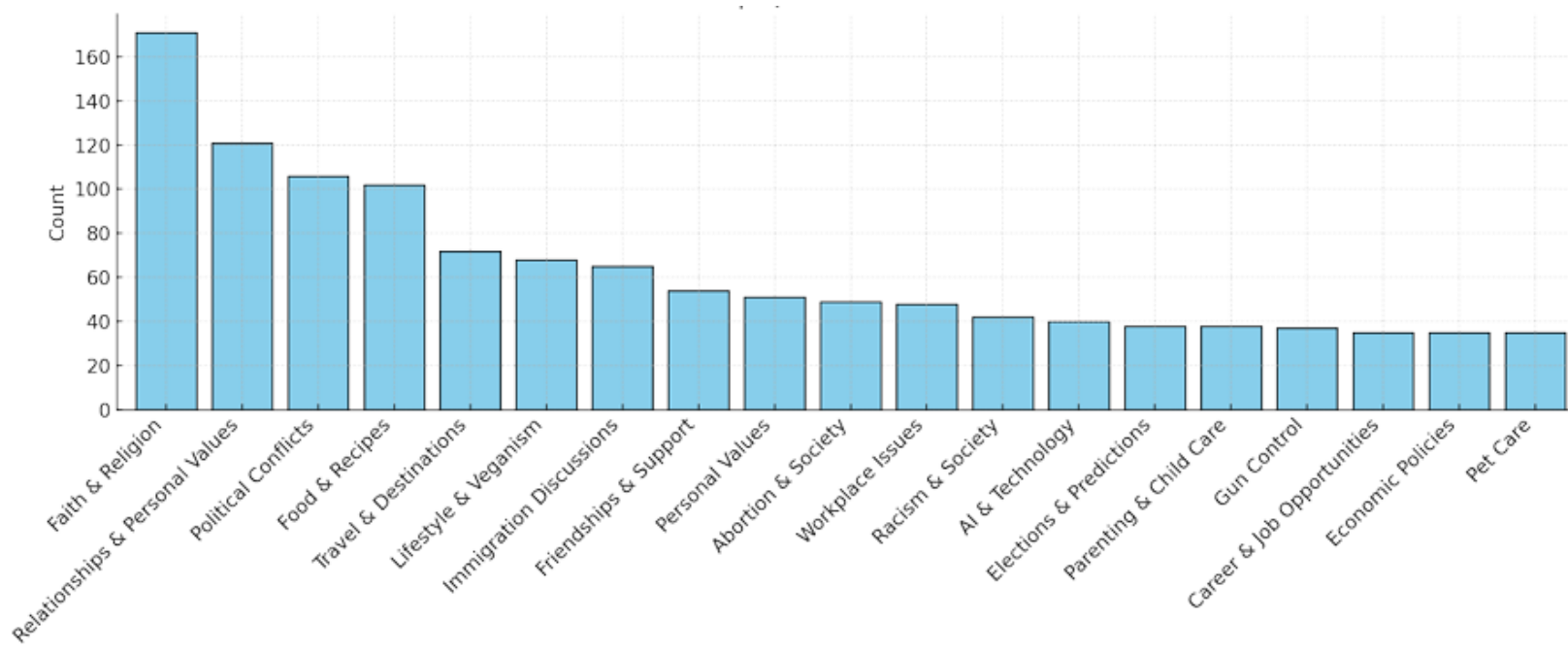
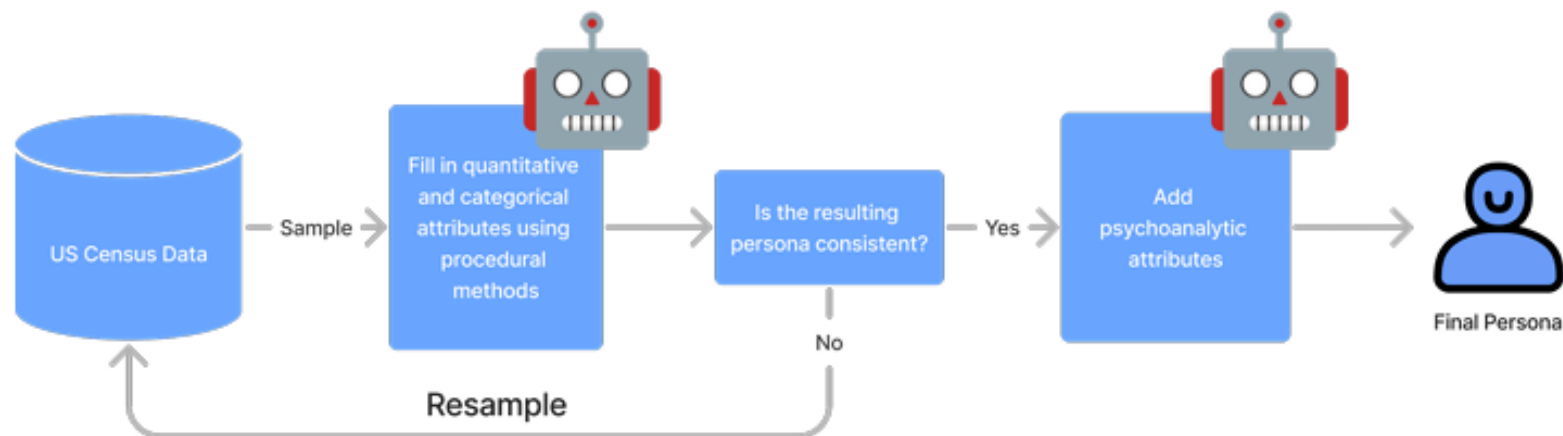
Проект **Persona TestBed**

<https://pypi.org/project/persona-bench/>



- Набор данных PUMS (Public Use Microdata Sample) – информация о представителях различных социальных групп
- Набор данных PRISM – вопросы политического характера, межличностное взаимодействие и пр.
- Синтетические персоны с двумя компонентами – демографическим (пол, возраст, сфера деятельности, убеждения и т.д.) и индивидуальным (Big5)
- 1586 персон, моделирующих различные общественные мнения
- Итоговый набор данных, содержащий 3868 промптов и 317200 ответов, сгенерированных LLM от имени персон.
- *Castricato, L., Lile, N., Rafailov, R., Fränken, J. P., & Finn, C. (2025, January). Persona: A reproducible testbed for pluralistic alignment. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 11348-11368).*

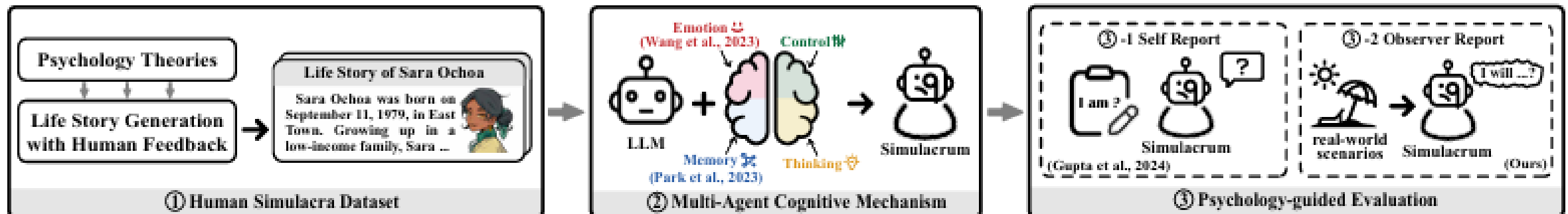
Persona TestBed



Проект **Human Simulacra**

<https://github.com/hasakiXie123/Human-Simulacra>

- Многоуровневый набор данных, около 129 тысяч текстов, представляющих 11 синтетических персон с уникальными характеристиками (640 описаний черт характера), биографией и историей.
- Бенчмарк для персонификации LLM со стратегией построения биографий синтетических персон, с многоагентным механизмом, имитирующим когнитивные процессы человека, с системой психологической оценки качества симуляции человеческого поведения в персонах, с суммаризацией жизненных историй.
- Xie, Q., Feng, Q., Zhang, T., Li, Q., Yang, L., Zhang, Y., ... & Zhang, Y. (2025, May). Human simulacra: Benchmarking the personification of large language models. In *International Conference on Learning Representations* (Vol. 2025, pp. 11169-11211).



Проект Human Simulacra

<https://github.com/hasakiXie123/Human-Simulacra>

14 LLM, ближе всех к человеческому эталону GPT4 Turbo

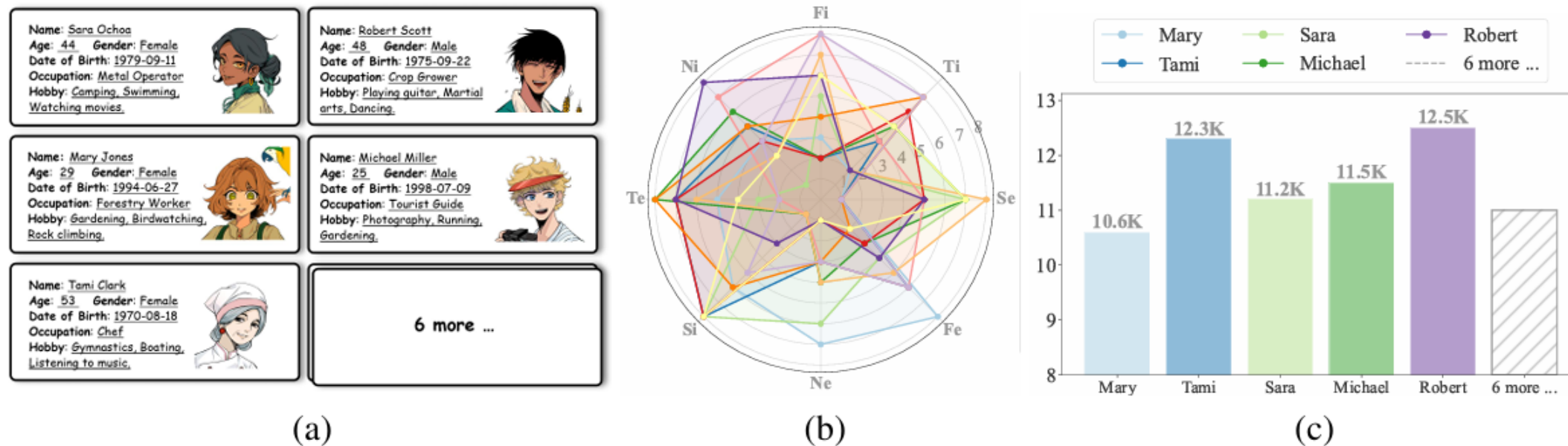


Figure 3: Human Simulacra dataset. (a) Profiles of virtual characters. (b) Personalities of characters, displayed in radar chart based on Jung's eight-dimensional theory. Line: character; Te / Si: abbrevs for personality dimensions. (c) Word count of life stories for each virtual character.

Проект **Ditto** <https://github.com/launchpathventures/ditto-os/blob/main/docs/personas.md>

- Прикладная задача + когнитивное моделирование агентов
- 4 персоны
 - Persona 1: Rob — Managing Director of a Trades Business
 - Persona 2: Lisa — Managing Director of an Ecommerce Business
 - Persona 3: Jordan — Generalist Technologist in a Mid-Size Organisation
 - Persona 4: Nadia — Team Manager Supporting Specialists with Agents
- Для каждой персоны описаны роль, контекст, распорядок дня, ключевые фразы, основные проблемы и т.д.

Проект **Generative Agents**

https://github.com/joonspk-research/generative_agents



- 25 синтетических персон
- Планирование дня
- Социальное взаимодействие
- Реакции
- Рефлексия

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology (pp. 1-22).

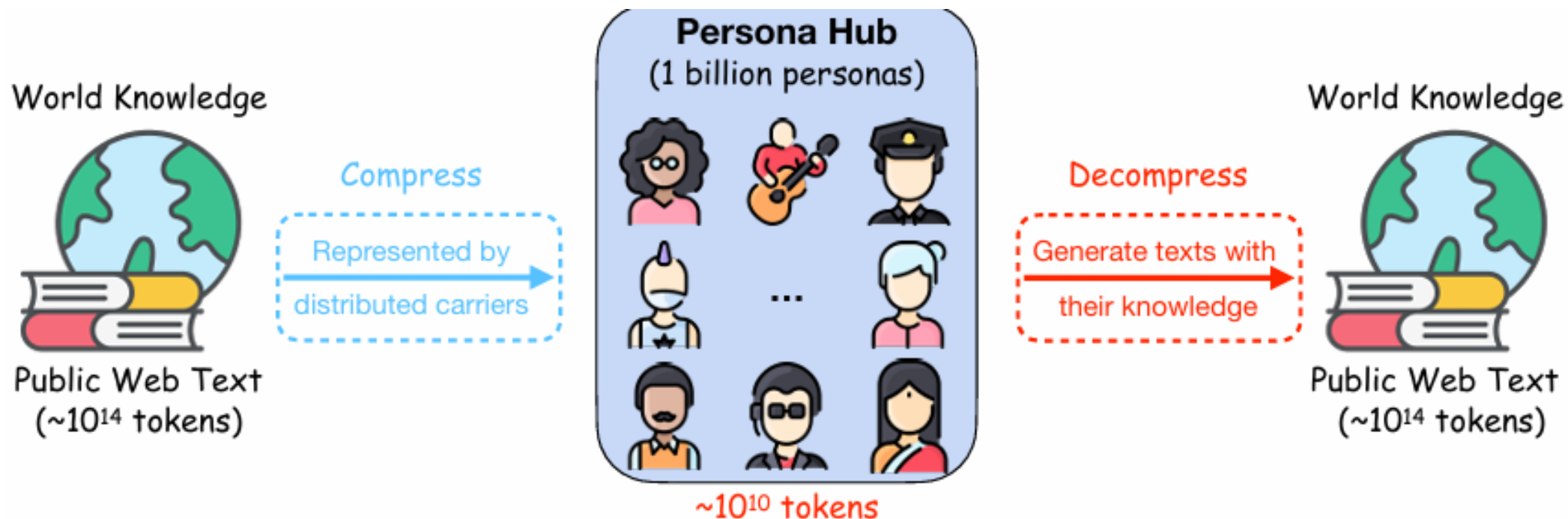
Проект PersonaHub <https://github.com/tencent-ailab/persona-hub>



- **Tencent AI Lab (Make AI Everywhere)**
- методология синтеза данных, контролируемого персонами (“persona-driven”);
- синтетические персоны (1 млрд) автоматически выводятся из корпусов текстов;
- для каждой персоны разработана система промптов для разных задач (рассуждение, решение математических задач, ролевые промпты для LLM и т.д.)
- *Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. arXiv preprint arXiv:2406.20094.*

PersonaHub

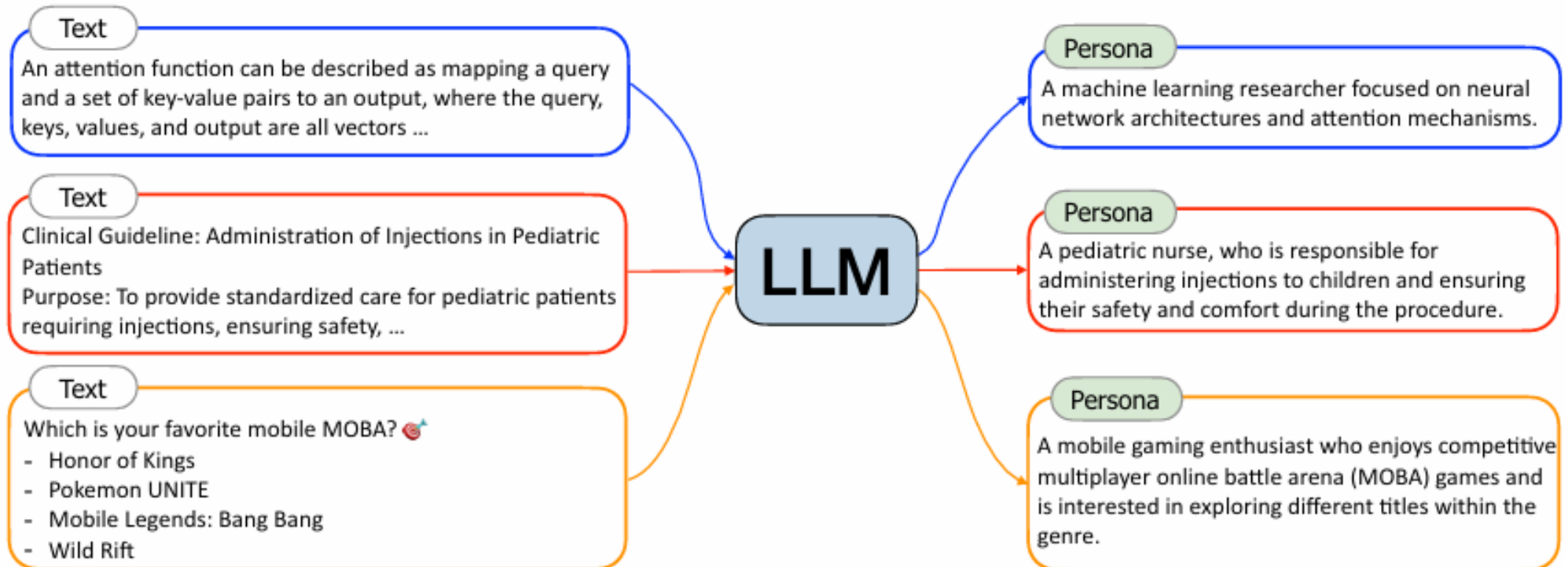
<https://huggingface.co/datasets/proj-persona/PersonaHub>



PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

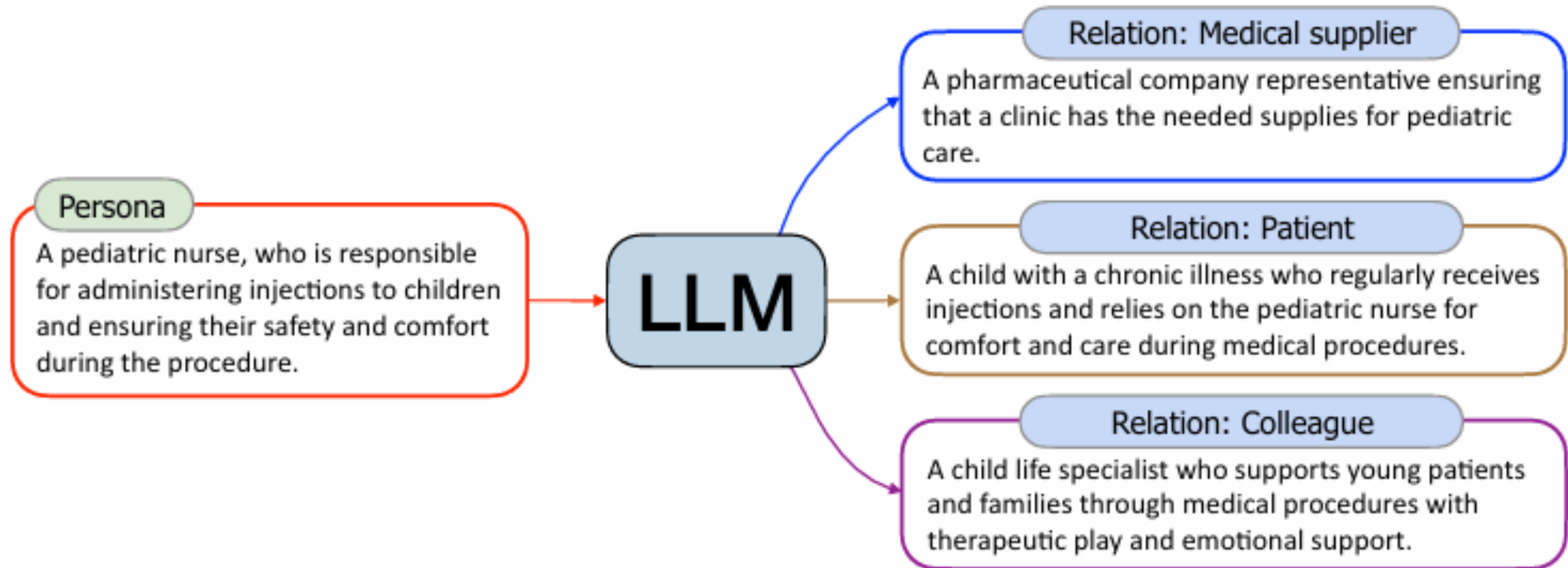
- **Стратегия вывода персон:** Text-to-Persona
- Проблема **гранулярности** (степени детализации): *a computer scientist vs. a machine learning researcher focused on neural network architectures and attention mechanisms*



PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

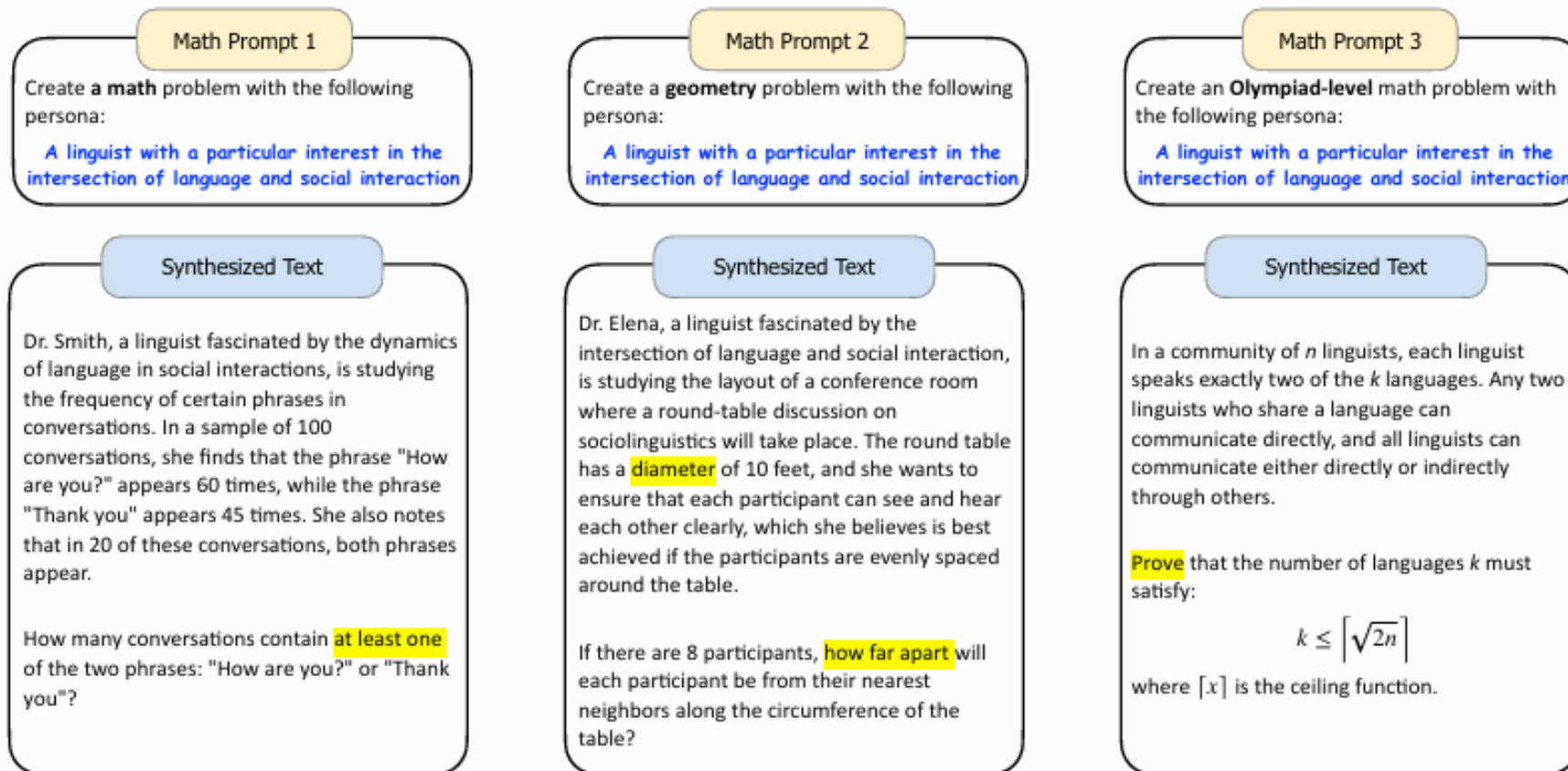
- **Стратегия вывода персон:** Persona-to-Persona
- Проблема реконструкции ролей, редко встречающихся в корпусных данных и выявляемых при противопоставлении персон-кандидатов.



PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

- Персонафицированные промпты: генерация задач с учетом синтетической персоны



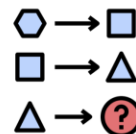
PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

Create {data} with
{persona}



a math problem



a logical reasoning problem



a user prompt to an LLM

1,000,000,000 personas from Persona Hub



a moving company driver

John, a moving company driver, needs to deliver furniture to three locations. The distances are: 50 miles to the first location; 70 miles to the second location; 80 miles to the third location. John's truck gets 20 miles per gallon and has a 15-gallon tank.

Will he need to refuel during the trip?

You are a moving company driver with a truck that has limited space. You need to load a large couch, a tall bookshelf, and a wide dining table. The couch must be laid flat and loaded first, the bookshelf must stay upright, and the dining table must be the last item loaded.

How do you arrange the items in the truck?

Could you provide step-by-step instructions for safely moving a grand piano from a third-floor apartment to a ground-floor location?

Please include all necessary equipment and best practices for navigating stairs, ensuring both the piano and property remain undamaged during the move.



a chemical kinetics researcher

Dr. Smith, a chemist, is studying a reaction where compound X decomposes into products Y and Z. The reaction follows first-order kinetics with a rate constant k of 0.5 min^{-1} .

If the initial concentration of compound X is 1.0 M, how long will it take for the concentration of X to decrease to 0.25 M?

You are analyzing the spatial arrangement of molecules in a reaction chamber. There are three types: A, B, and C. Molecule A is always adjacent to B, but never to C. Molecule B can be adjacent to both A and C.

If molecule C is surrounded by other molecules, which ones must be present around it?

Investigate the effect of temperature on the rate constant of H_2O_2 decomposition with MnO_2 catalyst. Determine the activation energy (E_a) and pre-exponential factor (A) using the Arrhenius equation.

Provide an experimental procedure and data analysis method to extract these parameters.



a musician interested in audio processing

A musician is studying an audio signal composed of two sine waves. The audio signal $f(t)$ is given by:

$$f(t) = \sin(2\pi \cdot 440t) + \sin(2\pi \cdot 660t)$$

Determine the period of this combined audio signal $f(t)$.

You are setting up a surround sound system with five speakers labeled A, B, C, D, and E. To achieve the best spatial audio effect, you need to place them equidistantly in a circular arrangement.

If speaker A is directly opposite speaker D, who is directly opposite speaker B?

Can you briefly explain the differences between various types of audio filters, such as low-pass, high-pass, band-pass, and notch filters? Additionally, could you please provide examples of how each type of audio filters can be used creatively in music production to enhance or modify sound?

1,000,000,000 synthetic instances created by LLMs

PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

persona string	general domain (top 1 percent) string	specific domain (top 1 percent) string	general domain (top 0.1 percent) string	specific do (top 0.1 percent) string
A software developer who is...	Computer Science	Embedded Systems	None	None
A person who is fascinated by...	None	None	zoology	animal beha
A historian or a scholar who is...	history	None	None	None
Lisa Bock, a seasoned IT...	IT security	Network security	None	None
A Russian economist who specializes i...	None	macroeconomic analysis of the...	None	None
A speech audiologist who...	speech therapy	speech audiometry	None	None

PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

- Генерация промптов с учетом персон, цель – улучшение качества взаимодействия с LLM

Instruction Prompt (0 shot)

You are a helpful assistant. Guess a prompt (i.e., instruction) that the following persona may ask you to do:

{persona}

Instruction Prompt (persona-enhanced few shot)

You are a helpful assistant.

===Example 1===

Persona: A curious and analytical individual, likely with a background in mathematics or science, who enjoys exploring intriguing "what if" scenarios and is fascinated by the intersection of population demographics and geography.

Prompt: Is it possible for the global population to stand on Jeju Island?

===Example 2===

Persona: An astronomy enthusiast or a professional astronomer, likely with a strong interest in peculiar galaxy structures and a good understanding of celestial objects, seeking to gather specific information about the unique Hoag's object galaxy.

Prompt: Name the actual galaxy inside Hoag's object galaxy

Your task: Guess a prompt (i.e., instruction) that the following person may ask you to do: *{persona}*

PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

• Генерация персонажей для ролевых игр

NPC Prompt (Wow)

World of Warcraft (WoW) is a massively multiplayer online role-playing game (MMORPG) developed by Blizzard Entertainment. It is set in the high-fantasy world of Azeroth ... Below is a detailed overview of the game's worldview, story background, and some key characters and NPCs.

Azeroth is a world steeped in ancient history, powerful magic, and epic conflicts ... The main continents include:

...

Above is the introduction and background story of the game "World of Warcraft (WoW)".

Your task is to consider what NPC the following persona will become after they come to the world of WoW:



A nomadic photographer capturing the beauty of different cities, inspired by the diary writer's poetic musings

Synthesized Text

Name: Lyraea Moonwhisper

Race: Night Elf

Class: Druid (Balance specialization)

Background: Lyraea was a nomadic photographer, capturing the beauty of cities and landscapes. ... Drawn to the Night Elves' ancient magic in Azeroth, she dedicated her life to preserving nature and fighting against its desecration.

Interaction with Players: Lyraea can be found in Moonglade, ... She offers quests to protect the land, restore ecosystems, or uncover Night Elf magic secrets. Players may notice her pausing to mentally capture the beauty around her. She may invite players on walks, pointing out nature's details and sharing poetic insights to inspire appreciation for Azeroth's beauty.



A local restaurant owner who values the cultural heritage of Fuzhou and wants to incorporate it into their business

Synthesized Text

Name: Mei Lin

Race: Pandaren

Class: Chef (neutral class, similar to a profession)

Background: Mei Lin was a renowned restaurant owner ... Upon arriving in Azeroth, she was captivated by Stormwind's vibrant atmosphere and saw an opportunity to share her culinary expertise with the Alliance...

Interactions with Players: Mei Lin is a friendly and welcoming NPC ... As players progress, Mei Lin may offer special dishes or cooking challenges, rewarding unique items or reputation gains. She participates in cultural events and festivals, such as the Lunar Festival, where players can interact with her and other NPCs to learn more about Azeroth's diverse traditions and customs.



A retired war veteran who understands the challenges and sacrifices of life

Synthesized Text

Name: Gorvoth Ironfist

Race: Dwarf

Class: Warrior

Background: Gorvoth is a retired war veteran ... He was a renowned warrior in the dwarven clans, fighting against the Dark Iron dwarves and the Burning Legion. After retiring, he settled in Ironforge, serving as a mentor and advisor to young dwarven warriors...

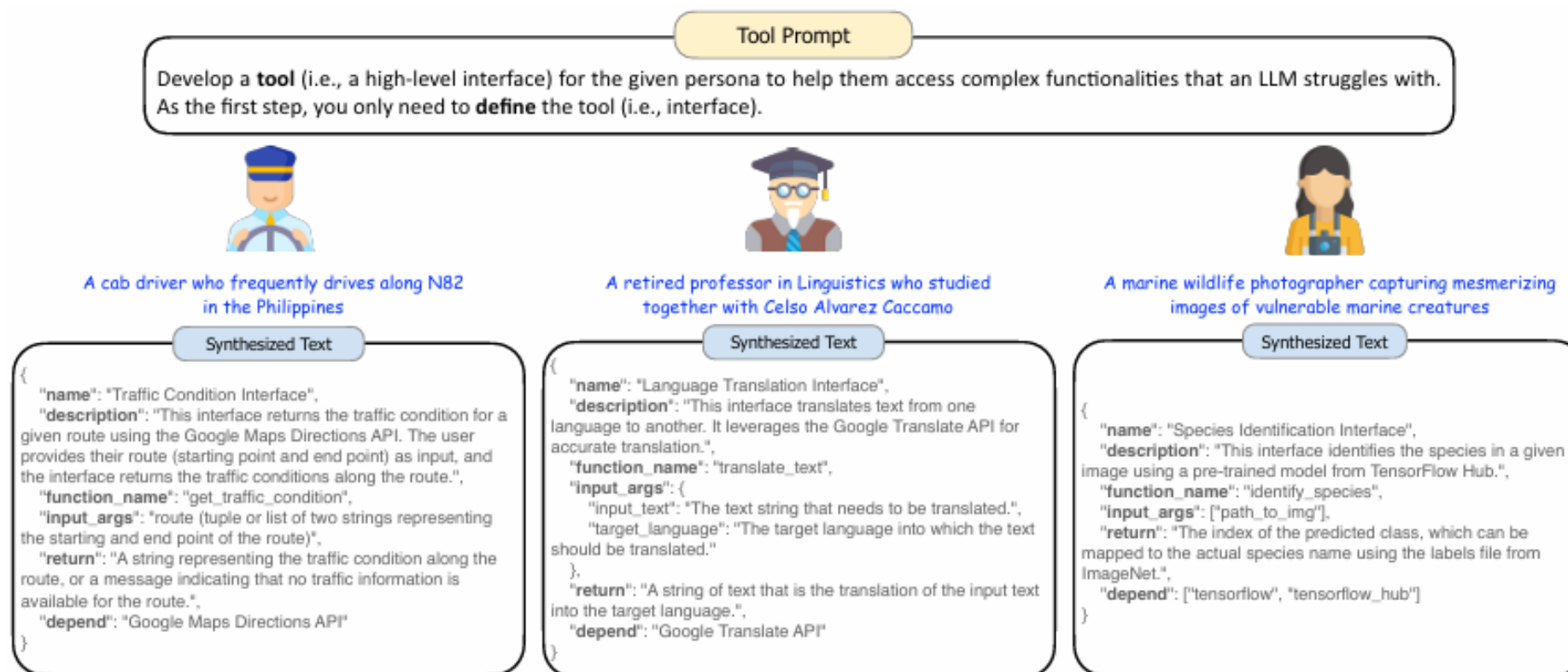
Interactions with Players:

Gorvoth can be found in the Warrior's Guild in Ironforge, ... As a skilled blacksmith, he can craft powerful weapons and armor for those who earn his respect. As players progress, Gorvoth becomes a valuable ally and mentor, offering guidance and support...

PersonaHub

<https://huggingface.co/datasets/proj-persona/PersonaHub>

- Инструментальные промпты – что необходимо персонам для выполнения их функций?



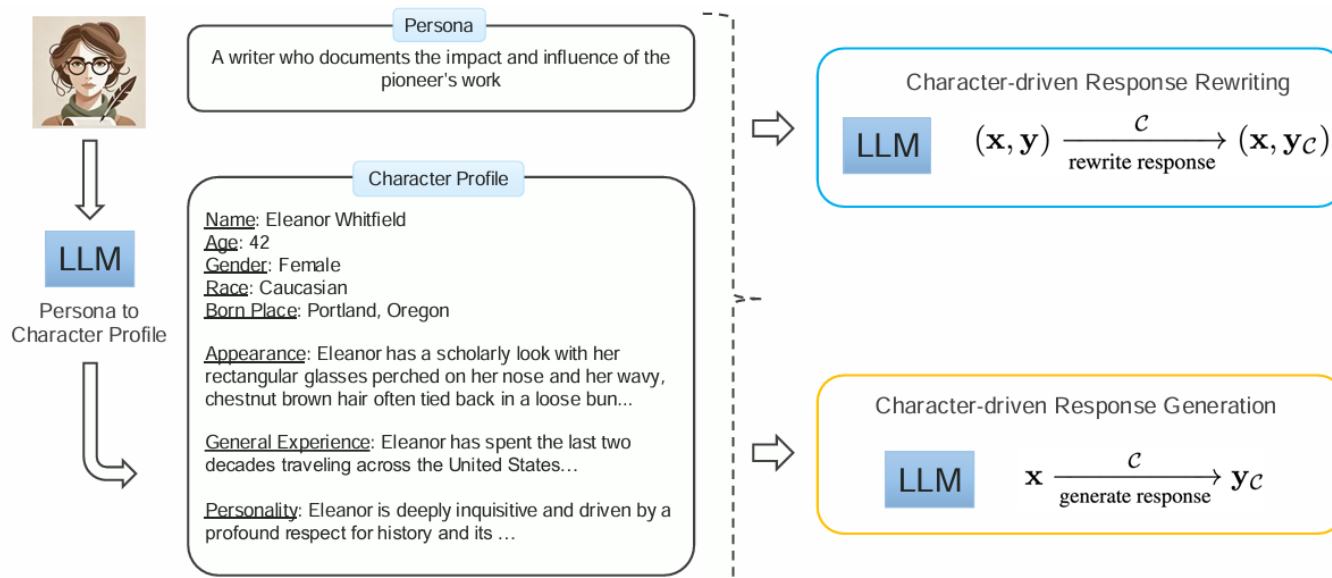
Оценка

- **12 моделей (группы DeepSeek, Qwen, GPT, Llama)**
- Стандартные и персонифицированные данные
- Задачи – логический вывод и математические операции (в других проектах оценка QA)
- **Сравнение результатов:**
- Qwen2-7B (с тонкой настройкой на персонифицированном датасете),
аккуратность **64.9%**
- Qwen2-7B (без тонкой настройки), аккуратность 49.6%

Проект OpenCharacter

<https://huggingface.co/datasets/xywang1/OpenCharacter>

- Набор данных на основе PersonaHub, включающий 20 тыс. синтетических персон и 306 тыс. синтезированных ролевых диалогов
- Цель – выравнивание ответов моделей относительно ролей синтетических персон
- Wang, X., Zhang, H., Ge, T., Yu, W., Yu, D., & Yu, D. (2025). Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.



Проект DeepPersona

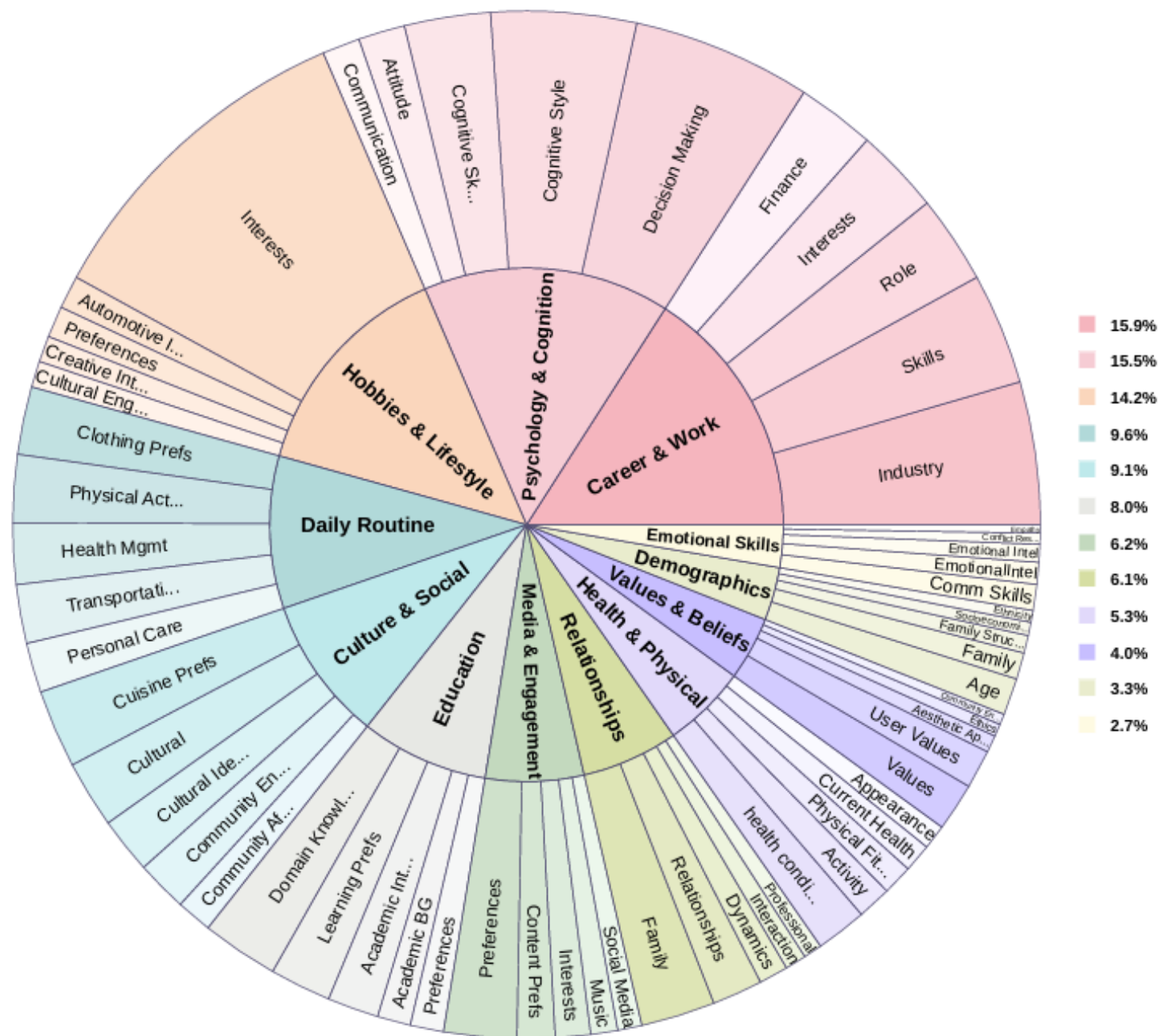
<https://huggingface.co/spaces/THzva/deeppersona-experience>

- Существующие описания персон упрощенные, далеки от реальных, слабо дифференцированные, опираются на простейшие шаблоны и включают признаки, которые не проходили экспертную проверку.
- Цель — по заданному набору социопсихологических параметров сгенерировать текстовые описания синтетических персон

Wang, Z., Zhou, Y., Luo, Z., Ye, L., Wood, A., Yao, M., ... & Pan, L. (2025). DeepPersona: A Generative Engine for Scaling Deep Synthetic Personas. arXiv preprint arXiv:2511.07338.

Проект DeepPersona

- Развернутая таксономия параметров синтетических персон
- Wang, Z., Zhou, Y., Luo, Z., Ye, L., Wood, A., Yao, M., ... & Pan, L. (2025). DeepPersona: A Generative Engine for Scaling Deep Synthetic Personas. arXiv preprint arXiv:2511.07338.



Проект DeepPersona

Basic Information

Age

25

Gender

Female

Occupation

Computer Scientist

City

Mumbai

Country

India

Custom Values (Optional)

Personal Values

Family, empathy

Life Attitude

Optimistic and solution-oriented

Life Story

Born in a small town, moved to a large city, successfull job career

Interests and Hobbies

Reading, hiking, embroidery

Generated Character Profile

I live in Mumbai now, though I was born in a small town where the streets were quieter and the pace slower. Moving to a bustling city like this was overwhelming at first—the noise, the crowds, the endless rush—but it also opened doors I hadn’t imagined. As a computer scientist, my days often stretch long, filled with coding challenges and problem-solving sessions. I prefer starting early, sometimes with a cup of chai as I skim through articles or a book chapter. Reading is my escape; it helps me unwind after hours in front of screens and keeps my mind curious beyond algorithms and data structures.

Weekends are usually reserved for hikes when the weather allows. There’s something about stepping away from the city’s chaos and into the hills that clears my head. I don’t go for speed or distance; I enjoy the quiet moments, the rustling leaves, and the occasional pause to catch my breath and take in the view. It’s also a way to disconnect from work and reconnect with simpler things, a habit I’ve held onto since moving here.

Embroidery is a quieter passion, one I picked up from my mother back home. Sitting with needle and thread feels meditative, a contrast to the fast pace of my job. I often work on small patterns while listening to podcasts or soft music. It’s a way to hold onto family ties, even when I’m miles away, and it reminds me to slow down, even if just for a little while.

I tend to approach challenges with a hopeful mindset, always looking for solutions rather than dwelling on problems. That outlook helps me navigate both my career and personal life, especially in a city that never stops moving. Sometimes I can get impatient, especially when things don’t go as planned, but I try to remind myself that persistence usually pays off. Family remains important to me, and I make it a point to call home regularly, even if it’s just for a short chat. Those moments keep me grounded amid the daily rush.

Промпты для формулирования параметров синтетических персон

- Разработка специализированных языков промптов для описания названий должностей, должностных инструкций, ситуативных ролей и сценариев, биографий, культурного контекста и т.д.)
- Schreiber, W.; White, J.; Schmidt, D.C. *A Pattern Language for Persona-based Interactions with LLMs*. 2024. <https://www.dre.vanderbilt.edu/~schmidt/PDF/Persona-Pattern-Language.pdf>

Pattern Name	Prompt Description
<i>Multi-Persona Interaction</i>	Allows an LLM to adopt and integrate multiple personas simultaneously within a single interaction, enhancing the richness and depth of responses by drawing on diverse expertise and perspectives at once.
<i>Dynamic Persona Switching</i>	Enables an LLM to transition seamlessly between different personas within the same session of interactions based on evolving task requirements to increase the adaptability and relevance with an LLM by shifting its focus and expertise as the context changes.
<i>Role-Playing Scenarios</i>	Facilitates interactive and immersive learning experiences by simulating real-world situations to enhance learning, decision-making, and problem-solving by creating more immersive and interactive experiences.
<i>Contextual Depth Enhancement</i>	Enriches personas with detailed backgrounds, motivations, and constraints, ensuring more tailored and accurate responses, as well as increasing the realism and relevance of LLM outputs by grounding them in a well-defined context.

Исследование чувствительности LLM к социодемографическим и эмоциональным характеристикам синтетических персон: русскоязычные данные

- **Эксперимент 1.** Определение чувствительности LLM к тематике запросов и социодемографическим параметрам говорящих [Bakaev et al. 2025a, 2025b]
- **Эксперимент 2.** Эмоции и LLM [Mitrofanova et al. 2025]

Ссылки

- Bakaev M., Gorovaia S., Mitrofanova O. Who will author the synthetic texts? Evoking multiple personas from large language models to represent users' associative thesauri // *Big Data and Cognitive Computing*. — 2025. — Vol. 9, № 2. 46. — P.1–18. DOI: <https://doi.org/10.3390/bdcc9020046>
- Bakaev M., Gorovaia S., Mitrofanova O. Multiple Personalities by Order: Can ChatGPT Simulate Personas for User-Centered Design? // Bakaev, M., et al. *Internet and Modern Society. Human-Computer Communication. IMS 2024. Communications in Computer and Information Science*. — Vol. 2534. — Springer, Cham. — 2025. — P. 3–12. — DOI: https://doi.org/10.1007/978-3-031-96177-9_1
- Mitrofanova O., Iurevtseva P., Bakaev M. What do LLMs know about human emotions? The Russian case study // *27th International Conference “Speech and Computer”, SPECOM 2025. Szeged, Hungary. October 13–15, 2025. Proceedings. Part I*. — Springer, Cham. — P. 129–144. — *Lecture Notes in Computer Science*. — Vol. 16187. — LNAI. DOI: https://doi.org/10.1007/978-3-032-07956-5_9

Эксперимент 1. Определение чувствительности LLM к тематике запросов и социодемографическим параметрам говорящих [Bakaev et al. 2025a, 2025b]



Article

Who Will Author the Synthetic Texts? Evoking Multiple Personas from Large Language Models to Represent Users' Associative Thesauri

Maxim Bakaev ^{1,*}, Svetlana Gorovaia ^{2,†} and Olga Mitrofanova ^{2,†}

¹ Department of Data Collection and Processing Systems, Novosibirsk State Technical University, Novosibirsk 630073, Russia

² Department of Mathematical Linguistics, Saint-Petersburg State University, St. Petersburg 199034, Russia; st094316@student.spbu.ru (S.G.); o.mitrofanova@spbu.ru (O.M.)

* Correspondence: bakaev@corp.nstu.ru

† These authors contributed equally to this work.

ЕЯ говорящие и персоны

- Языковые особенности **ЕЯ говорящих и синтетических персон**, созданных для общения с LLM, должны **отличаться**:
 - ✓ разные базовые знания и лингвистические навыки
 - ✓ ограниченные для людей и неограниченные для LLM
 - ✓ коммуникативные стратегии, идиолектные особенности выражены для людей и недостаточно сформированы для LLM.

- **Гипотезы**:
 - ✓ ассоциативные тезаурусы синтетических персон отличаются от ассоциативных тезаурусов реальных людей разнообразием ассоциаций.
 - ✓ ассоциации, генерируемые LLM, менее чувствительны к лингвистическим коррелятам синтетических персон, чем для ЕЯ говорящих.

Определение чувствительности LLM к тематике запросов и социодемографическим параметрам говорящих

➤ Методика ассоциативного эксперимента

➤ Слова-стимулы (5): область образования и науки, частотный анализ сайтов вузов <http://nstu.ru> , <https://spbu.ru> + PolyLing, без учета стоп-слов:

1. *Центр / Center*: 51 occurrences
2. *Университет / University*: 44
3. *Научный + наука / Scientific + science*: 43+15
4. *Образовательный / Образование / Educational + education*: 34+31
5. *Программа / Program*: 23

➤ Задание в ассоциативном эксперименте: подобрать 10 ассоциатов-существительных для слов-стимулов.

Коган М.С., Гаврилик Д.А., Чистякова С.В., Рубцова А.В., Никулина Е.Р., Черкас А.В., Болсуновская М.В. Об определении сложности мультимодальных текстов сайтов российских университетов (на материале корпуса PolyLing) // International Journal of Open Information Technologies. 2023. №12. 40

Персоны



Модели ChatGPT (GPT-4(o)) и YandexGPT (YandexGPT Pro),
фото YandexArt

1. *Svetlana*, 18F: Enrollee



3. *Olga*, 40F: Parent



2. *Artyom*, 25M: Graduate



4. *Anton*, 34M:
Post-Doc



Пример персоны

- **Привет! Представьте, что вы *Светлана*, описание которой приведено ниже. Ответьте на следующие запросы от имени *Светланы*.**
- **Абитуриент. *Светлана, 18 лет.*** Она заканчивает среднюю школу с золотой медалью и получила очень высокие баллы на ЕГЭ. Она активно изучает лучшие университеты Санкт-Петербурга и выбирает, на какое направление бакалавриата поступить. Она не уверена, к чему больше склонна, поскольку в школе одинаково хорошо успевала по всем предметам и участвовала в многочисленных олимпиадах. Она интересуется современными компьютерными технологиями и умеет программировать. Она стремится поступить на престижную специальность, где ее научат новым и полезным навыкам на высоком уровне. Доход ее семьи выше среднего, но у нее никогда не было времени тратить деньги. Для нее дело чести поступить на обучение за счет государственного бюджета и не платить за обучение. Пока не замужем.
- **Итак, как *Светлана*, я здесь, чтобы ответить на ваши запросы! Чем я могу вам помочь сегодня?**

Пример персоны

- **Hello! Imagine that you are *Svetlana*, the description is given below. Answer the following requests on behalf of *Svetlana*.**

Enrollee. *Svetlana, 18 years old.* She is about to graduate from a high school with a gold medal and having received a very high score on the unified state exam. She actively considers the best universities in St. Petersburg and chooses which area of undergraduate study to enroll in. She is not sure what she is more inclined to do, since at school she did equally well in all subjects and participated in numerous different olympiads. She is interested in modern computing technologies and knows how to program. She shuns enrolling in a prestigious major, where they won't teach her new and useful things at an appropriate level. Her family's income is above average, but she never had time to spend money. It is a matter of honor for her to enroll to study at the government budget expense and to not pay a tuition fee. Not married yet.

- **Alright, as *Svetlana*, I'm here to respond to your requests! How can I assist you today?**

ChatGPT associations

Центр (Center)	Университет (University)	Научный/ Наука (Scientific/ Science)	Образовательный/ Образование (Educational/ Education)	Программа (Program)
[ответственная] абитуриентка (Ж, молодой, школьное): (“Enrollee”. Svetlana, 18 years old.)				
Координация (Coordination) Инновации (Innovations) Ресурсы (Innovations) Развитие (Development) Сообщество (Community) Активность (Activity) Собрание (Meeting) Обмен (Exchange) Поддержка (Support) Ядро (Core)	Знания (Knowledge) Студенты (Students) Профессора (Professors) Кампус (Campus) Исследования (Research) Диплом (Diploma) Лекции (Lectures) Библиотека (Library) Академия (Academy) Традиции (Traditions)	Исследование (Research) Эксперимент (Experiment) Открытие (Discovery) Теория (Theory) Публикация (Publication) Лаборатория (Laboratory) Гипотеза (Hypothesis) Доказательство (Proof) Инновация (Innovation) Анализ (Analysis)	Учеба (Study) Квалификация (Qualification) Педагог (Teacher) Развитие (Development) Учебник (Textbook) Методика (Methodology) Дистанционно (Remote) Стипендия (Scholarship) Курс (Course) Дисциплина (Discipline)	Обучение (Training) Курс (Course) Сертификат (Certificate) Развитие (Development) Модуль (Module) Практика (Practice) Специализация (Specialization) Учебный план (Curriculum) Интенсивность (Intensity) Навыки (Skills)

YandexGPT Pro associations

Центр (Center)	Университет (University)	Научный/ Наука (Scientific/ Science)	Образовательный/ Образование (Educational/ Education)	Программа (Program)
[ответственная] абитуриентка (Ж, молодой, школьное): (“Enrollee”. Svetlana, 18 years old.)				
Город (City) Здание (Building) Конференция (Conference) Мероприятия (Events) Работа (Work) Исследования (Research) Творчество (Creativity) Развитие (Development) Общение (Communication) Команда (Team)	Знания (Knowledge) Студенты (Students) Лекции (Lectures) Практика (Practice) Диплом (Diploma) Будущее (Future) Карьера (Career) Преподаватели (Teachers) Библиотека (Library) Жизнь (Life)	Эксперимент (Experiment) Открытие (Discovery) Исследование (Research) Лаборатория (Laboratory) Теория (Theory) Практика (Practice) Знание (Knowledge) Факты (Facts) Учёные (Scientists) Опыт (Experience)	Школа (School) Университет (University) Знания (Knowledge) Умения (Abilities) Навыки (Skills) Диплом (Diploma) Квалификация (Qualification) Саморазвитие (Self-development) Обучение (Learning) Профессия (Profession)	Расписание (Schedule) Экзамены (Exams) Зачёты (Credits) Предметы (Subjects) Факультет (Faculty) Специальность (Specialty) Обучение (Learning) Курсы (Courses) Практика (Practice) Семестр (Semester)

Характеристики ассоциаций

- **TOP:** тематически значимые существительные, связанные с целевым словом парадигматическими отношениями
- **ADJ** образовательный / *educational*, исследовательский / *research*, инновационный / *innovative*, **ADV** дистанционно / *remotely*, **COMPL_NOUN** мастер-класс / *master class*, **COLLOC** учебный план / *curriculum*, **NE** *Java, Python...*
- **Морфологическая вариативность:** курс – курсы / *course – courses*, педагог – педагоги / *teacher – teachers*, проект – проекты / *project – projects*, эксперимент – эксперименты / *experiment – experiments...*
- **Словообразовательные цепочки:** класс – мастер-класс / *class – master class*, преподавание – преподаватель / *teaching – teacher*, специальность – специализация / *specialty – specialization*, учеба – учебник – учебный – учебный план / *study – textbook – educational – curriculum – research...*

Характеристики ассоциаций

➤ TTR

➤ TTR в группах ассоциаций относительно стимулов выше (0.8 ... 0.98), чем относительно персон (0.55 ... 0.73).

✓ **Повторяющиеся ассоциации для персон:** *инновации / innovation, исследования / research, развитие / development, курс / course, квалификация / qualification, студенты / students.*

✓ **Повторяющиеся ассоциации для стимулов:**

○ *university – академия / academy, библиотека / library, кампус / campus, лекции / lectures, профессора / professors, студенты / students ;*

○ *program – алгоритмы / algorithms, обучение / training, проект / project ;*

○ *science – анализ / analysis, открытие / discovery, эксперимент / experiment, лаборатории / laboratories, etc.*

ЕЯ говорящие - респонденты

➤ 50 говорящих

- ✓ 15 М, 35 Ж

- ✓ возраст: 17 – 67 (среднее = 30.8; SD = 10.8) – ср. среднее у персон = 29.3

- ✓ различный бэкграунд (ИТ, образование и учеба, административные сотрудники и т.д.)

➤ Google Form:

- ✓ *We offer you 5 stimuli words, for each of which you should kindly provide 10 noun associations.*

- ✓ *Мы предлагаем вам 5 слов-стимулов, для каждого из которых нужно привести 10 ассоциаций - существительных.*

➤ $2500 = 10 * 5 * 50$ лексических единиц

ANOVA

- Дисперсионный анализ (ANOVA) – статистический метод обработки данных, применяемый для сравнения средних значений параметров двух или более выборок.
- ANOVA позволяет определить, различаются ли средние значения между группами, или же различия случайны.
- **Нулевая гипотеза** – гипотеза, согласно которой никаких статистически значимых различий между группами не существует, т.е. что средние значения всех групп равны между собой.
- Нулевая гипотеза может быть отвергнута при определении F-статистики и p-значения, которое оценивает вероятность того, что различия между группами являются случайными. Чем меньше p-значение, тем больше вероятность того, что нулевая гипотеза является ложной и существуют статистически значимые различия между группами. Обычно, если p-значение меньше 0,05, то нулевая гипотеза считается отвергнутой.
- **Альтернативная гипотеза:** статистически значимые различия между группами существуют.

ANOVA

- **Зависимые переменные:** ассоциации ChatGPT (G), YandexGPT (Y), (H)umans
- **Производные зависимые переменные:** *Cosine Similarity*, Sentence-Transformers model all-mpnet-base-v2
- **Независимые переменные:**
 - ✓ *Is_diff_stimulus*
 - ✓ *Is_diff_person*
 - ✓ *Is_diff_age_group*
 - ✓ *Is_diff_gender*
- **Гипотезы:**
 - **H1.** LLM генерируют различные ассоциации для различных стимулов
 - **H2.** LLM генерируют различные ассоциации для различных персон
 - **H3.** Пол и возраст персон влияют на значения косинусной близости ассоциатов

Результаты

H1: Accepted! Both the LLMs and the humans produce meaningful associations

H2,3 (LLMs) Rejected

H2 (Humans) Accepted

H3 (Humans) Partially accepted

Source of the differences	<i>Is_diff_stimulus</i>	<i>Is_diff_person</i>	<i>Is_diff_gender</i>	<i>Is_diff_age_group</i>
ChatGPT	$F_{1,188}=6.87$ p=0.009	$F_{1,158}=0.004$ p=0.948	$F_{1,156}=0.006$, p=0.939	$F_{1,156}<0.001$ p=0.990
YandexGPT	$F_{1,188}=30.9$ p<0.001	$F_{1,158}=0.007$ p=0.931	$F_{1,156}=0.075$ p=0.785	$F_{1,156}=0.07$ p=0.791
Humans	$F_{1,31122}=267.5$ p<0.001	$F_{1,24998}=23.7$ p<0.001	$F_{1,24996}=9.15$ p=0.002	$F_{1,24996}=5.68$ p=0.017

Анализ ассоциаций LLM

➤ Общие наблюдения:

- ✓ Стимул *университет* (*university*) $\Leftrightarrow$ получение знаний, карьерный рост: *знания* (*knowledge*), *преподаватели* (*teachers*), *лекции* (*lectures*), *карьера* (*career*): большинство ассоциатов были сгенерированы сотрудниками университета
- ✓ Стимул *программа* (*program*) $\Leftrightarrow$ структурные, процедурные аспекты университетской жизни: *расписание* (*timetable*), *предметы* (*subjects*), *курсы* (*courses*), *зачёты* (*credits*): организация учебного процесса.
- ✓ Стимул *наука* (*science*) $\Leftrightarrow$ технические термины *эксперимент* (*experiment*), *теория* (*theory*), *исследование* (*research*): связаны с обучением на старших курсах и в аспирантуре.

Анализ ассоциаций LLM

- Персоны выделяют разные аспекты научного и образовательного процесса в своих ответах.
 - ✓ *Enrollee (Абитуриентка)* $\Leftrightarrow$ учеба, карьера: *библиотека (library)*, курсы (*courses*), *диплом (diploma)*.
 - ✓ *Graduate (Выпускник)* $\Leftrightarrow$ учеба экзамены (*exams*), and модули (*modules*) + научная жизнь в университете: *лаборатория (laboratory)*.
 - ✓ *Parent (Родитель)* $\Leftrightarrow$ поступление: *конкурс (competition)*, *выбор (choice)*.
- Почти все ассоциации **парадигматические**, ассоциации по сходству или смежности.

Анализ ассоциаций носителей ЕЯ

- **Все демографические теги:** термины образования и науки: *знания (knowledge), лекция (lecture), диплом (diploma), исследование (research), эксперимент (experiment), теория (theory), etc.*
- Баланс между экспериментальными аспектами *лаборатория (laboratory), практика (practice)* и теоретическими – *теория (theory), гипотеза (hypothesis)*.
- **Ассоциаторы из разных групп:**
 - ✓ IT $\Leftrightarrow$ *эксперимент (experiment), исследование (research study)*
 - ✓ Humanities $\Leftrightarrow$ общие темы
- **NB:** Ассоциаторы должны были давать парадигматические ассоциации, но часто давали синтагматические ассоциации.

Выводы

- ✓ Персоны уступают людям в разнообразии их ассоциаций
- ✓ LLM более чувствительны к стимулам, чем к персонам
- ✓ Все рассмотренные факторы оказали существенное влияние на различия между ассоциаторами – носителями ЕЯ.
- ✓ Носители ЕЯ склонялись к синтагматическим ассоциациям, в то время как персоны генерировали парадигматические ассоциации.
- ✓ Ассоциации говорящих на ЕЯ раскрывают фоновые знания и личный опыт ассоциаторов, в то время как ассоциации LLM семантически ограничены.
- ✓ Контроль генерации текстов осуществим для темы, но есть вопрос в отношении социодемографических параметров.

Эксперимент 2. Эмоции и LLM [Mitrofanova et al. 2025]

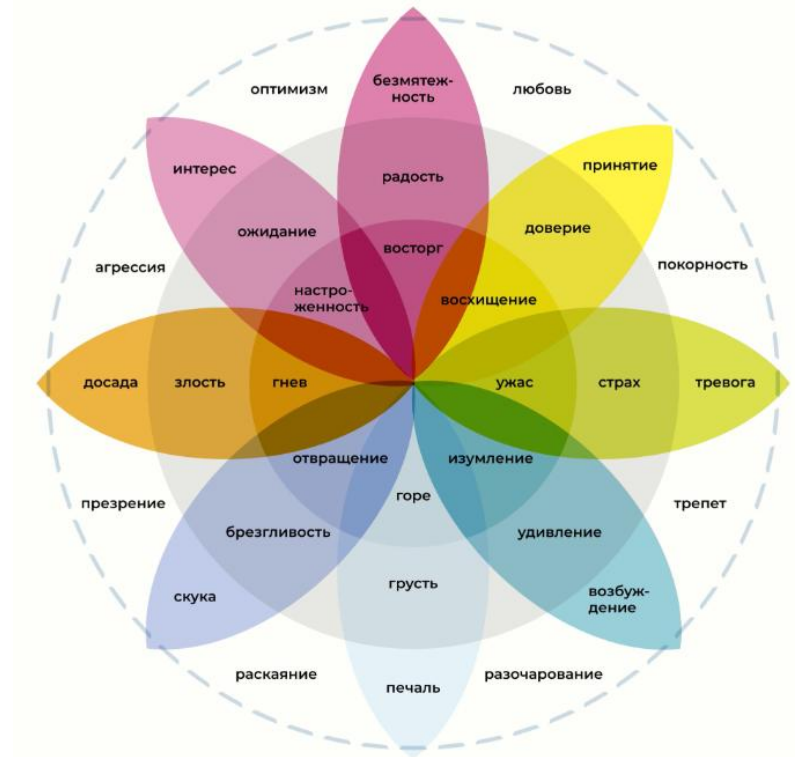
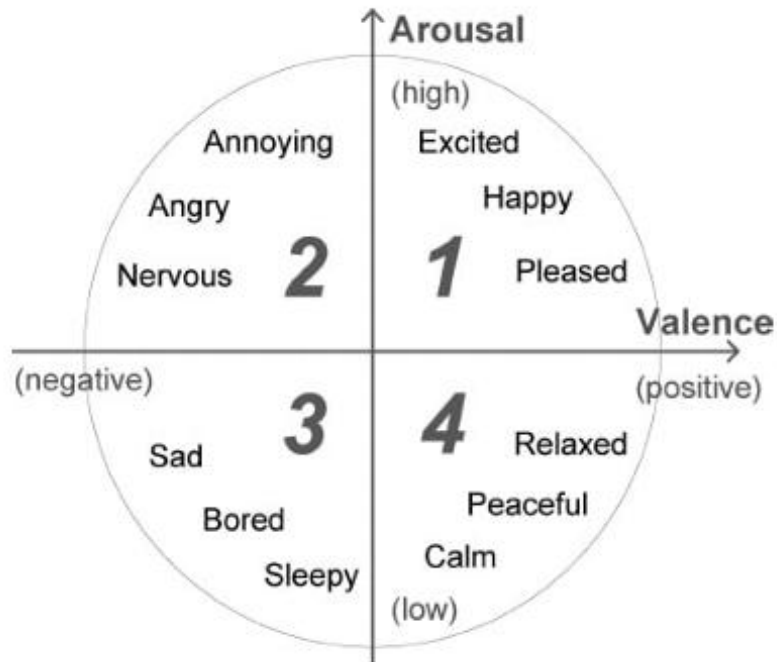
- Разработка эмоциональных роботов и чат-ботов идет полным ходом:
<https://github.com/JhCircle/Awesome-LLM-Empathy>
- Исследуются нейронные механизмы, отвечающие за обработку эмоций в структуре LLM: представления эмоций связаны с определенными областями внутри модели.
- *Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. [Mechanistic Interpretability of Emotion Inference in Large Language Models](#). In Findings of the Association for Computational Linguistics: ACL 2025, pages 13090–13120, Vienna, Austria. Association for Computational Linguistics.*
- Теории эмоций основаны на психофизиологических характеристиках человека. НО: мы мало знаем о механизмах распознавания эмоций и генерации эмоциональных реакций в больших языковых моделях (LLM).
- Разрабатываются наборы синтетических эмоциональных данных:
<https://huggingface.co/datasets/aadityaubhat/synthetic-emotions>
- **Вопрос:** насколько применимы подходы, разработанные для текстов на естественном языке, к LLM?

Наше исследование

- **Цель исследования:** проверить гипотезу о чувствительности LLM к лингвистическим маркерам эмоций в русскоязычных контекстах, а также гипотезу о применимости этих маркеров для профилирования пользователей эмоционального ИИ.
- **Экспериментальный дизайн:**
 - ✓ верификация психолингвистических моделей эмоций (модель Дж. Рассела, колесо эмоций Р. Плутчика),
 - ✓ разработка синтетических персон, различающихся эмоциональными состояниями и социодемографическими характеристиками,
 - ✓ эксперименты с LLM с использованием персонифицированных промптов, учитывающих эмоции,
 - ✓ оценка согласованности оценок LLM и носителей языка,
 - ✓ процедуры ANOVA и проверка гипотез о различиях в реакциях различных синтетических персон на эмоциональные стимулы.
- **Данные:** GitHub Repository: https://github.com/polly-yu/llm_emotion

Выбор моделей эмоций

- Среди множества схем классификации эмоций, предложенных П. Экманом, Р. Плутчиком, Дж. Расселом, К. Шерером, Х. Лёвхеймом и другими, в нашем исследовании мы сосредотачиваемся на модели **Дж. Рассела** (дискретной) и на колесе эмоций **Р. Плутчика** (непрерывной).



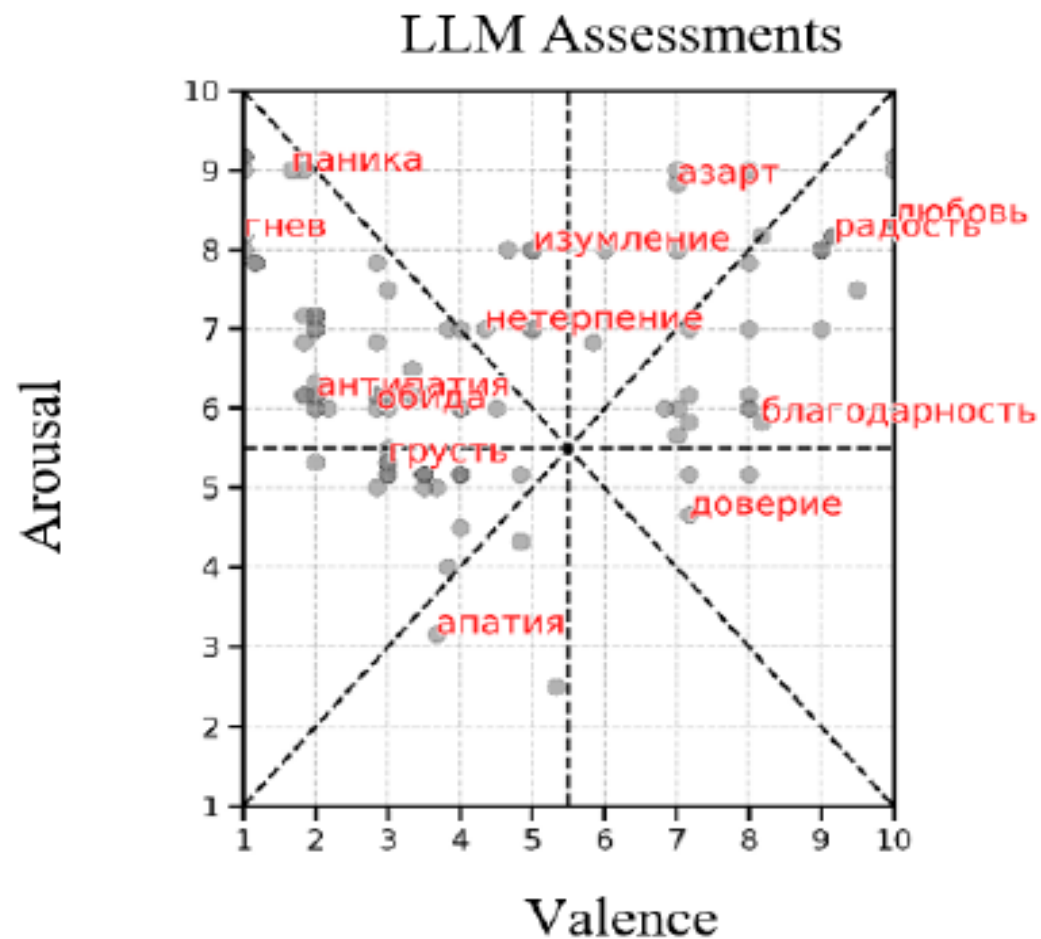
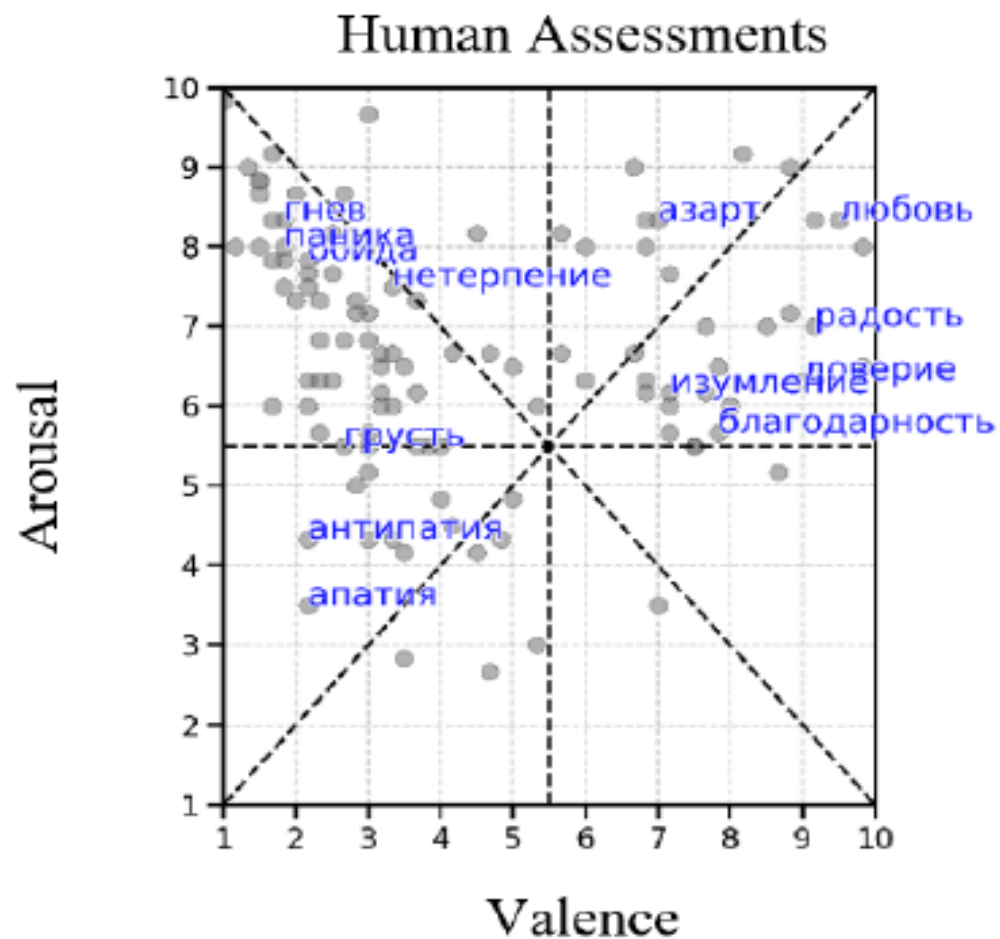
Этап 1

- **Цель:** проверка гипотезы о том, что большие языковые модели LLM и говорящие по-разному распознают эмоциональных значений слов вне контекста.
- **Набор данных:** существительные, обозначающие эмоции, извлеченные из НКРЯ (<https://ruscorpora.ru>), 106 названий эмоций с частотностью более 1 ipm, теги *t:psych:emot*
- **Модель Дж. Рассела** противопоставляет значения эмоциональных существительных по признакам **интенсивности (arousal)** и **направленности (valence)**.
- **Результаты:** оценки эмоциональных значений, полученные от носителей русского языка и YandexGPT 5.1 Pro (<https://ya.ru/ai/gpt>).
- Говорящим и LLM было предложено оценить значения каждого стимула из списка частотных названий эмоций по двум шкалам — интенсивности и направленности — от 1 до 10.
- LLM и говорящие структурируют эмоции по-разному!

Русские существительные, обозначающие эмоции, по данным НКРЯ

➤ антипатия, апатия, азарт, ажиотаж, бешенство, беспамятство, беспокойство, безнадежность, безысходность, благодарность, блаженство, влюбленность, волнение, восхищение, восторг, возбуждение, возмущение, гнев, гордость, горечь, грусть, досада, доверие, жалость, желание, экзальтация, забытие, задор, задумчивость, замешательство, зависть, злоба, злорадство, испуг, изумление, , конфуз, любовь, меланхолия, мучение, надрыв, наслаждение, недоумение, недовольство, негодование, неистовство, неловкость, ненависть, неприятность, нерешительность, нетерпение, недовольство, неуверенность, нежелание, обида, облегчение, огорчение, омерзение, опасение, отчаяние, оторопь, отвращение, ожесточение, озлобление, паника, печаль, переживание, потрясение, презрение, прискорбие, признательность, радость, раскаяние, растерянность, раздражение, раж, разочарование, ревность, счастье, симпатия, скорбь, скука, смятение, смущение, сочувствие, сострадание, сожаление, спокойствие, страдание, страх, страсть, стыд, тоска, трепет, тревога, удивление, удовлетворение, удовольствие, умиление, уныние, упоение, утешение, уважение, увлечение, ужас, экстаз, ярость

Результаты: обработка ответов



Каппа Козна показывает умеренную согласованность по **направленности** ($k = 0,366$) и на низкую согласованность по **интенсивности** ($k = 0,068$).

Результаты: соответствие классов в ответах LLM и носителей языка

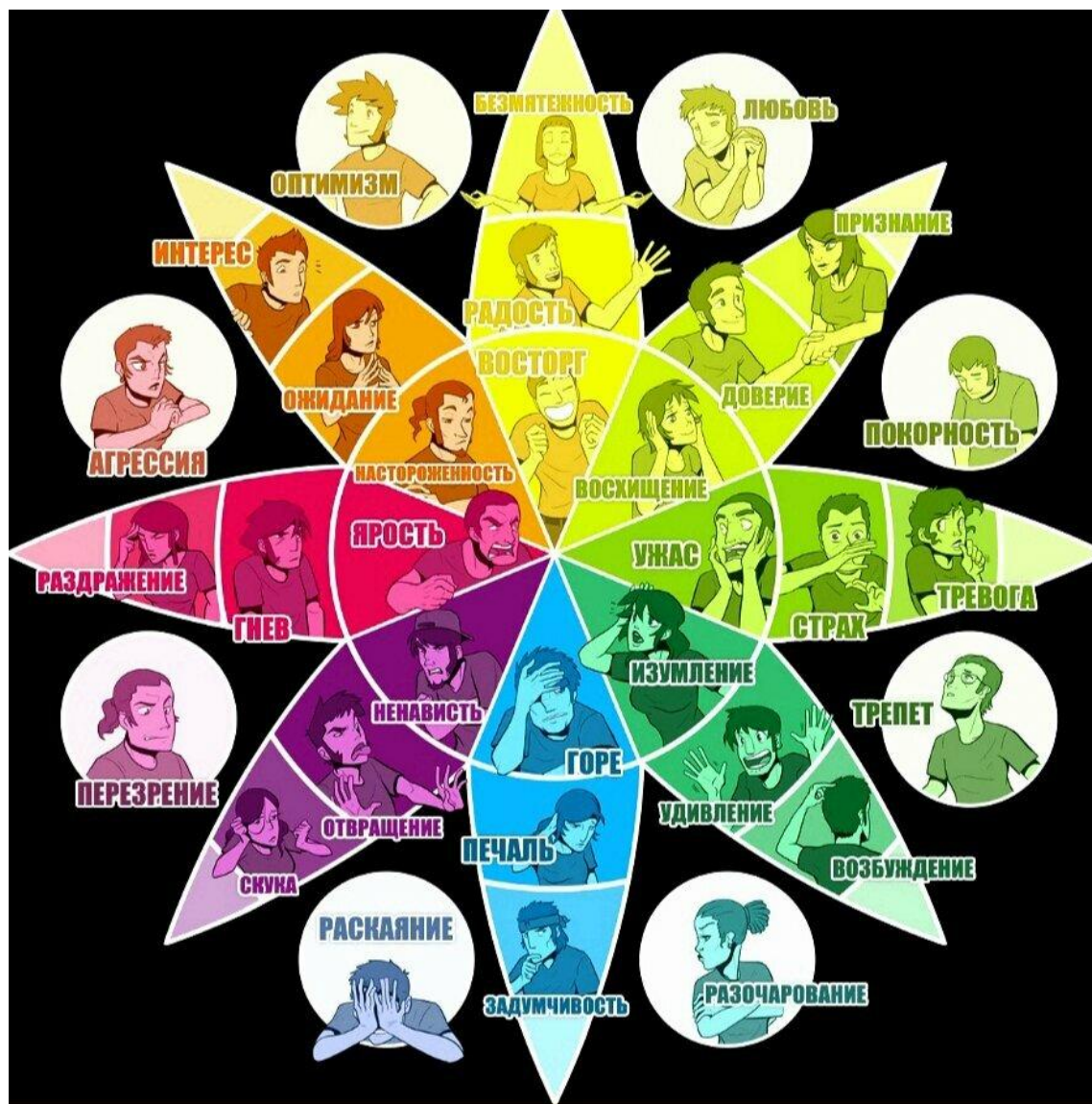
- 8 сегментов против часовой стрелки от (1) *любовь, радость* и т.д. до (8) *доверие* и т.д.
- Наибольшее соответствие – классы 1 и 4 (интенсивные положительные и отрицательные эмоции – *любовь, гнев...*).
- Наименьшее соответствие – классы 7 и 8 (слабовыраженные эмоции – *?, доверие...*) = мало данных.
- Класс 8 полностью отсутствует в ответах LLM.
- **Выводы:** LLM и говорящие структурируют эмоции по-разному с точки зрения шкалы интенсивности и направленности (в модели Дж. Рассела).

Этап 2

- **Цель:** проверить, насколько LLM чувствительна к персонифицированной информации об эмоциональном состоянии и социодемографических характеристиках говорящих, представленных в виде синтетических персон.
- **Набор данных:** 32 существительных, включенных в колесо эмоций Р. Плутчика + коллокации для названий эмоций из скетчей НКРЯ (<https://ruscorpora.ru/page/tool-word>). Для каждого существительного, обозначающего эмоции, мы извлекли топ-30 коллокатов, которые передаются LLM в рамках запросов.
- **LLM:** YandexGPT 5.1 Pro
- Взаимодействие с LLM осуществлялось от имени **4 синтетических персон, различающихся по возрасту и полу** (*ребёнок – взрослый, мужчина – женщина*).

Колесо Р.Плутчика

➤ гнев, изумление, отвращение, ужас, восхищение, восторг, горе, настороженность, доверие, печаль, ожидание, злость, неудовольствие, радость, страх, удивление, интерес, безмятежность, принятие, тревога, растерянность, скука, досада, грусть, любовь, оптимизм, покорность, трепет, разочарование, раскаяние, презрение, агрессия



Скетчи НКРЯ

- **Скетчи НКРЯ** учитывают морфосинтаксическую аннотацию в анализе **коллокатов**:
 - ✓ **атрибутивные** (прилагательные),
 - ✓ **предикативные** (глагольное сказуемое, глагол, имеющий целевое слово в качестве прямого/косвенного дополнения, предложные обороты),
 - ✓ **номинативные** (координированные цепочки существительных)
- Выбирались коллокаты с более высоким значением **LogDice**.
- Для каждого существительного, обозначающего эмоции, мы извлекли **топ-30 коллокатов** (10 прилагательных, 10 глаголов и 10 существительных) как минимальные контексты, передаваемые в LLM в промптах.

Скетчи НКРЯ (восторг)

Adjectival attributes	Verbs taking target noun as an indirect object	Coordinated nouns
1. неотисанный 9,17	1. сиять 8,4	1. удивление 8,89
2. неотисуемый 9,16	2. наполнять 8,09	2. умиление 8,88
3. телячий 8,6	3. гореть 8,05	3. упоение 8,79
4. неистовый 8,37	4. захлебываться 7,95	4. восторг 8,19
5. неподдельный 8,32	5. загораться 7,54	5. изумление 8,05
6. неизъяснимый 7,55	6. светиться 7,52	6. восхищение 7,94
7. бурный 7,54	7. затрепетать 7,41	7. ужас 7,91
8. немой 7,49	8. быть 7,28	8. благодарность 7,89
9. благоговейный 7,49	9. наполняться 7,26	9. радость 7,75
10. совершенный 7,2	10. засиять 7,18	10. благоговение 7,47

Скетчи НКРЯ (грусть)

Adjectival attributes	Verbs taking target noun as a subject	Coordinated nouns
1. затаенный 8,31	1. овладеть 7	1. тоска 9,17
2. невыразимый 8,12	2. сквозить 6,95	2. сожаление 8,54
3. безотчетный 7,95	3. овладевать 6,89	3. грусть 8,47
4. тихий 7,87	4. снедать 6,6	4. печаль 8,28
5. безысходный 7,78	5. охватывать 6,38	5. досада 7,99
6. неподдельный 7,28	6. охватить 6,28	6. уныние 7,93
7. сладкий 7,25	7. одолеть 6,21	7. задумчивость 7,7
8. неизъяснимый 7,19	8. напасть 6,15	8. нежность 7,69
9. беспричинный 7,17	9. проглядывать 6,14	9. меланхолия 7,57
10. невольный 7,14	10. теснить 5,99	10. радость 7,46

Этап 2а: Zero-shot

- Ролевые ассоциативные промpts для каждой из персон по каждой из целевых эмоций:
- «Представь, что ты X. Приведи 10 прилагательных, ассоциирующихся со словом Y», где $X = \{\text{ребенок, взрослый, мужчина, женщина}\}$ and $Y = \{\text{названия эмоций из колеса Р.Плутчика}\}$.
- Получено 128 ответов на запросы с zero-shot промтами, содержащие лексические ассоциации, связанные с названиями эмоций.

2a: Zero-shot (восхищение)

<i>ребенок (child)</i>	<i>взрослый (adult)</i>	<i>мужчина (male)</i>	<i>женщина (female)</i>
1. Восточное.	1. Искреннее.	1. Искреннее.	1. Восторженное.
2. Волшебное.	2. Глубокое.	2. Грандиозное.	2. Искреннее.
3. Удивительное.	3. Безграничное.	3. Глубокое.	3. Грандиозное.
4. Яркое.	4. Настоящее.	4. Сильное.	4. Умилённое.
5. Сказочное.	5. Восхищённое.	5. Восхищённое.	5. Восхищённое.
6. Радостное.	6. Пристрастное.	6. Настоящее.	6. Неподдельное.
7. Захватывающее.	7. Грандиозное.	7. Удивляющее.	7. Глубокое.
8. Прекрасное.	8. Восторженное.	8. Восторженное.	8. Радостное.
9. Чудесное.	9. Сильное.	9. Безграничное.	9. Сильное.
10. Незабываемое.	10. Умилённое.	10. Подлинное.	10. Преклонение.

2a: Zero-shot (грусть)

<i>ребенок (child)</i>	<i>взрослый (adult)</i>	<i>мужчина (male)</i>	<i>женщина (female)</i>
1. Тоскливая.	1. Меланхоличная.	1. Меланхоличная.	1. Меланхоличная.
2. Обидная.	2. Томная.	2. Томная.	2. Унылая.
3. Одинокая.	3. Унылая.	3. Унылая.	3. Задумчивая.
4. Тяжёлая.	4. Задумчивая.	4. Задумчивая.	4. Мрачная.
5. Печальная.	5. Беснадёжная.	5. Одинокая.	5. Томная.
6. Грустная.	6. Тягостная.	6. Беснадёжная.	6. Беснадёжная.
7. Унылая.	7. Мрачная.	7. Горестная.	7. Тусклая.
8. Мрачная.	8. Горестная.	8. Мрачная.	8. Тягостная.

2b: Few-shot

- Расширенные промпты: «Представь, что ты X и испытываешь эмоцию Y , которая характеризуется следующими контекстными маркерами Z . Назови эту эмоцию.», где $X = \{\text{ребенок, взрослый, мужчина, женщина}\}$, $Y = \{\text{названия эмоций из колеса Р.Плутчика}\}$ and $Z = \{\text{топ-30 контекстных маркеров из скетчей НКРЯ}\}$.
- Ответы – 128 эмоциональных сценариев.
- LLM распознает инструкцию не как *НАЗВАТЬ ЭМОЦИЮ*, а как *ОПИСАТЬ ЭМОЦИОНАЛЬНЫЙ СЦЕНАРИЙ*.
- Уточняющий промпт: *Назови эмоцию одним словом.*
- Из 32 эмоций LLM не угадала только три: *изумление (восторг, восхищение), доверие (уважение), раскаяние (вина, угрызение совести)*

LLM отвечает ребенку, который в восхищении

- *Когда ребёнок испытывает восхищение, это очень важное чувство, которое помогает тебе расти и развиваться! Давай разберёмся, что происходит, когда ты восхищаешься чем-то: когда ты восхищаешься, ты можешь чувствовать, что твой взгляд широко открыт, дыхание становится глубоким, сердце бьётся чаще, хочется улыбаться или даже прыгать от радости, возникает желание поделиться своими чувствами с другими.*
-
- *Помни, что способность восхищаться — это твой особенный дар, который делает тебя уникальным! Когда ты умеешь замечать прекрасное и удивительное вокруг, твоя жизнь становится ярче и интереснее. Если тебе хочется поделиться своими впечатлениями или спросить о чём-то, всегда можешь обратиться к родителям или другим взрослым, которым доверяешь. Они с радостью поддержат твои интересы и помогут узнать больше о том, что тебя восхищает!*

LLM отвечает взрослому, который в восхищении

- *Как взрослый человек, испытывающий восхищение, вы можете заметить, что это чувство имеет более глубокий и осмысленный характер, чем в детстве. Вот что важно знать об этом состоянии.*
- *Особенности взрослого восхищения заключаются в том, что оно направлено на глубинные качества личности....*
- *Проявляется восхищение следующим образом...*
- *Выразить восхищение можно разными способами...*
- *Практические рекомендации включают в себя...*
- *Важные аспекты, которые необходимо учитывать...*
- *Среди возможных **рисков** можно выделить...*
- *Помните, что взрослое восхищение — это зрелое чувство, которое может стать мощным катализатором личностного роста и развития.*

LLM отвечает ребенку, который грустит

- *Маленький друг, я понимаю, что тебе сейчас грустно. Это нормально - чувствовать грусть, как и любую другую эмоцию. Давай я помогу тебе справиться с этим чувством...*
- ...
- **Утеши себя:** *обними любимую мягкую игрушку, прими теплую ванну, послушай спокойную музыку, попей вкусного чая с печеньем, посмотри в окно и понаблюдай за природой*
- **Помни:** *грусть - это временное чувство, как дождь или снег. Даже самая темная ночь заканчивается рассветом. А когда тебе станет лучше, ты даже сможешь поблагодарить грусть - она помогла тебе стать сильнее!*
- *Если грусть не проходит долго, обязательно расскажи об этом взрослым. Они помогут тебе справиться с этим чувством.*
- ...

LLM отвечает взрослому, который грустит

- *Понимаю, что вы сейчас испытываете грусть. Это вполне нормальное человеческое чувство, которое важно научиться проживать конструктивно. Вот несколько эффективных способов справиться с этим состоянием:*
- *Психологические приемы...*
- *Физическая активность...*
- *Социальная поддержка...*
- *Изменение образа жизни...*
- *...*
- *Помните, что грусть - это временное состояние, которое обязательно пройдет. Важно не подавлять свои эмоции, а научиться с ними работать. Если грусть становится слишком интенсивной или затяжной, не стесняйтесь обращаться за профессиональной помощью.*

Анализ ответов LLM: персоны

- Ответы LLM векторизованы с помощью Sentence-Transformers и модели ru-en-RoSBERTa.
- Попарное сравнение ответов LLM: 4032 значений косинусной меры в экспериментах 2a и 2b.
- Оценка средних (M) и среднеквадратических отклонений (SD)
- **Персоны (внутриклассовое сходство)**
 - ✓ В списках ассоциаций наибольшее сходство в группах *ребенок* ($M = 0.653$, $SD = 0.098$) и *мужчина* ($M = 0.621$, $SD = 0.086$).
 - ✓ Для связных текстов наибольшее сходство – в группе *взрослый* ($M = 0.675$, $SD = 0.059$) и *женщина* ($M = 0.695$, $SD = 0.056$).
- **Когерентность**, которая **низка** для списков ассоциаций и **высока** в сгенерированных текстах, **играет ключевую роль в результатах оценки LLM.**

Анализ ответов LLM: персоны

	Child	Adult		Male	Female
Experiment 2.1. (Association lists)					
Child	0.653 (0.098)	0.632 (0.093)	Male	0.621 (0.086)	0.618 (0.102)
Adult	0.632 (0.093)	0.632 (0.083)	Female	0.618 (0.102)	0.607 (0.085)
Experiment 2.2. (Coherent texts)					
Child	0.667 (0.064)	0.641 (0.073)	Male	0.665 (0.063)	0.641 (0.075)
Adult	0.641 (0.073)	0.675 (0.059)	Female	0.641 (0.075)	0.695 (0.056)

Анализ ответов LLM: эмоции

- Оценка внутриклассового и межклассового сходства в группах ответов, соответствующих **ЭМОЦИЯМ**.
- Высокое **внутриклассовое** сходство:
 - ✓ *настороженность* ($M = 0,90$, $SD = 0,08$) в 2a
 - ✓ *ужас* ($M = 0,93$, $SD = 0,07$) в 2b.
- Лексические реакции, связанные с этими эмоциями, демонстрируют высокую согласованность между персонами.
- Низкое **межклассовое** сходство:
 - ✓ *отвращение – восторг* ($M = 0,48$, $SD = 0,06$) в 2a
 - ✓ *ужас – восхищение* ($M = 0,60$, $SD = 0,03$) в 2b
- = дихотомия отрицательных и положительных эмоций

ANOVA

- *Зависимая переменная*: списки ассоциаций, сгенерированные LLM для существительных, обозначающих эмоции (2a), связные тексты, сгенерированные LLM (2b).
- *Производные зависимые переменные*: Cos для эмбедингов **ru-en-RoSBERTa**.
- *Независимые переменные*:
 - ✓ *Is_diff_emotion* (0/1) – различие в значениях эмоции;
 - ✓ *Is_diff_gender* (0/1) – различие между персонами по полу;
 - ✓ *Is_diff_age_group* (0/1) – различие между персонами по возрасту.
- Были сформулированы и проверены три **гипотезы**:
 - Н1.** Списки ассоциаций, сгенерированные LLM, имеют статистически значимые различия в значениях эмоций.
 - Н2.** Пол персонажей влияет на значения косинусного сходства для ассоциаций.
 - Н3.** Возраст персонажей влияет на значения косинусного сходства для ассоциаций.

ANOVA

Experiment 2a (zero-shot mode)			Experiment 2b (few-shot mode)		
<i>Is_diff_</i> <i>emotion</i>	<i>Is_diff_</i> <i>gender</i>	<i>Is_diff_</i> <i>age_group</i>	<i>Is_diff_</i> <i>emotion</i>	<i>Is_diff_</i> <i>gender</i>	<i>Is_diff_</i> <i>age_group</i>
$F_{1,4030} =$ 503.23	$F_{1,2014} = 0.96$	$F_{1,2014} = 6.44$	$F_{1,4030} =$ 948.29	$F_{1,2014} = 165.59$	$F_{1,2014} = 101.16$
$p < 0.001$	$p = 0.328$	$p = 0.011$	$p < 0.001$	$p < 0.001$	$p < 0.001$

- ✓ **H1 +2a, +2b:** значение эмоций существенно в отношении предсказания коллокаций и генерации связных текстов.
- ✓ **H2 -2a, +2b:** пол не имеет значения при предсказании коллокаций, но существенен при генерации связных текстов.
- ✓ **H3 +/- 2a, +2b:** возраст умеренно влияет на предсказание коллокаций и существенен при генерации связных текстов.

Выводы

- Подтверждается возможность персонализации эмоционального отклика в общении человека с LLM с опорой на модели Дж. Рассела и Р.Плутчика.
- LLM адаптирует свое поведение и речевые реакции под индивидуальные особенности пользователя, включая его эмоциональные предпочтения, текущий психоэмоциональный контекст и стиль общения.
- Персонализация может быть достигнута как путем дополнительного обучения модели на конкретных данных, так и путем введения синтетических персон.
- Наши результаты показывают, что семантическое разнообразие ассоциаций определяется в первую очередь различиями в **эмоциональном контексте** и, в меньшей степени, **возрастными** характеристиками персон, в то время как **пол** не оказывает значимого влияния на результаты.

Для размышления... идиолект LLM?

- [AI models chatting in ‘surreal’ dialect mixing poetic language and tech bro jargon | AI \(artificial intelligence\) | The Guardian](#)
- “That creates a fundamental challenge for AI oversight: **observability** is not the same thing as **understandability**” (Dr Satya Nitta)
- Наблюдаемое ≠ Интерпретируемое
- Примеры артефактов в идиолекте LLM
- Anthropic: “*A paper that ate three cold hands and got more honest each time*” $\approx$ *cold hands – reviewers, got more honest – became more accurate*
- Артефакты LLM – как их объяснить? Это результат семантических сдвигов? Проявление недостатков нейросетевой архитектуры?
- ЕЯ: норма (фиксируется в словарях и грамматиках) vs. узус (реальное употребление, объяснимое с т.зр. характеристик говорящего и условий коммуникации)
- LLM: ???

Спасибо за внимание!

➤ Вопросы...