

Using a LLM for Dialectology:

Map Experiment

*(Exploring the Conflict between the Seen
and the Known)*

Ruslana Margova, GATE
CLaDa, 26.06.2026

New age of AI

- Large Language Models (LLMs) create opportunities for analyzing complex language datasets.
- **The Spatially Challenge:** how AI processes spatial information and interprets specialized, non-textual scientific data.

New age of AI

- **Dialectology as a Test:** Dialect maps offer a complex multi-layered structure combining linguistic classifications, geographic distribution, and visual isoglosses.
- **The Cognitive Conflict:** Interpreting these maps requires distinguishing between physical landmarks (mountains, rivers) and abstract, invisible linguistic boundaries.

New age of AI

Ћћ

- **The Case Study Focus:** Evaluating how a multimodal LLM balances visual geography with academic knowledge when tracing the primary Bulgarian dialect division (the Yat isogloss).

New age of AI

ᚼᚼ

- **RQ1: Do AI-generated interpretations align with established dialectological classifications?**
- **RQ2: To what extent does physical geography influence the model's interpretation of abstract linguistic boundaries?**
- **rq3: Speculative linguistic hypotheses**

New age of AI



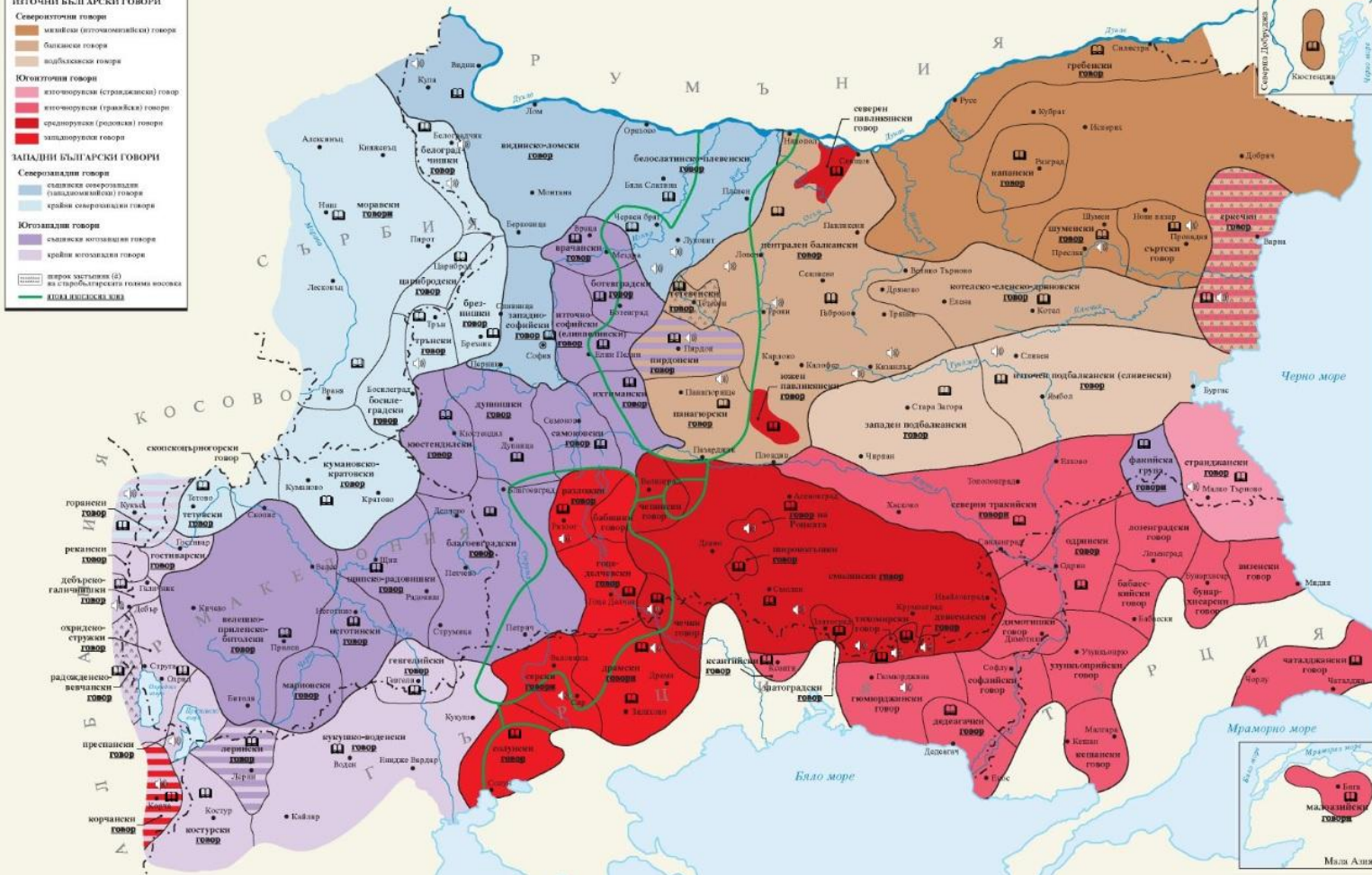
- **LLM Vulnerabilities:** Existing research highlights significant geographic biases and hallucinations.
- **Digital Dialectology:** Modern geolinguistics leverages GIS and digital platforms (like the IBL-BAS Interactive Dialect Map) to map critical features like the Bulgarian Yat boundary.
- **The Research Gap:** no many studies examine how AI systems interpret multi-layered geolinguistic maps.

Experiment with me...

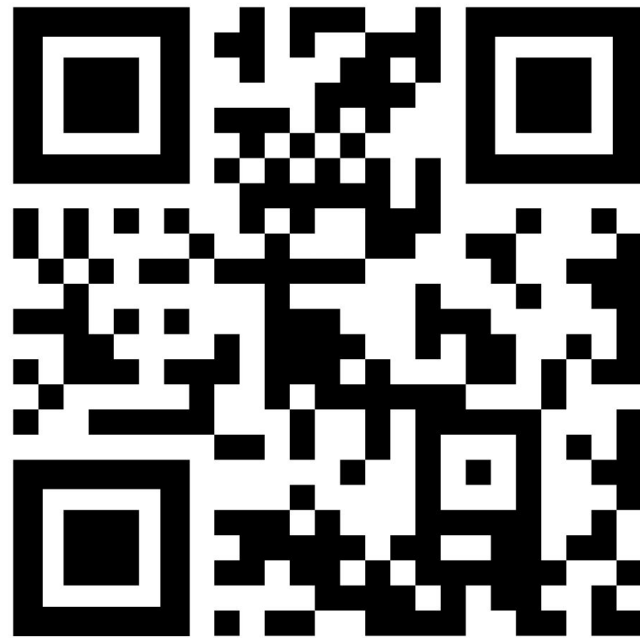


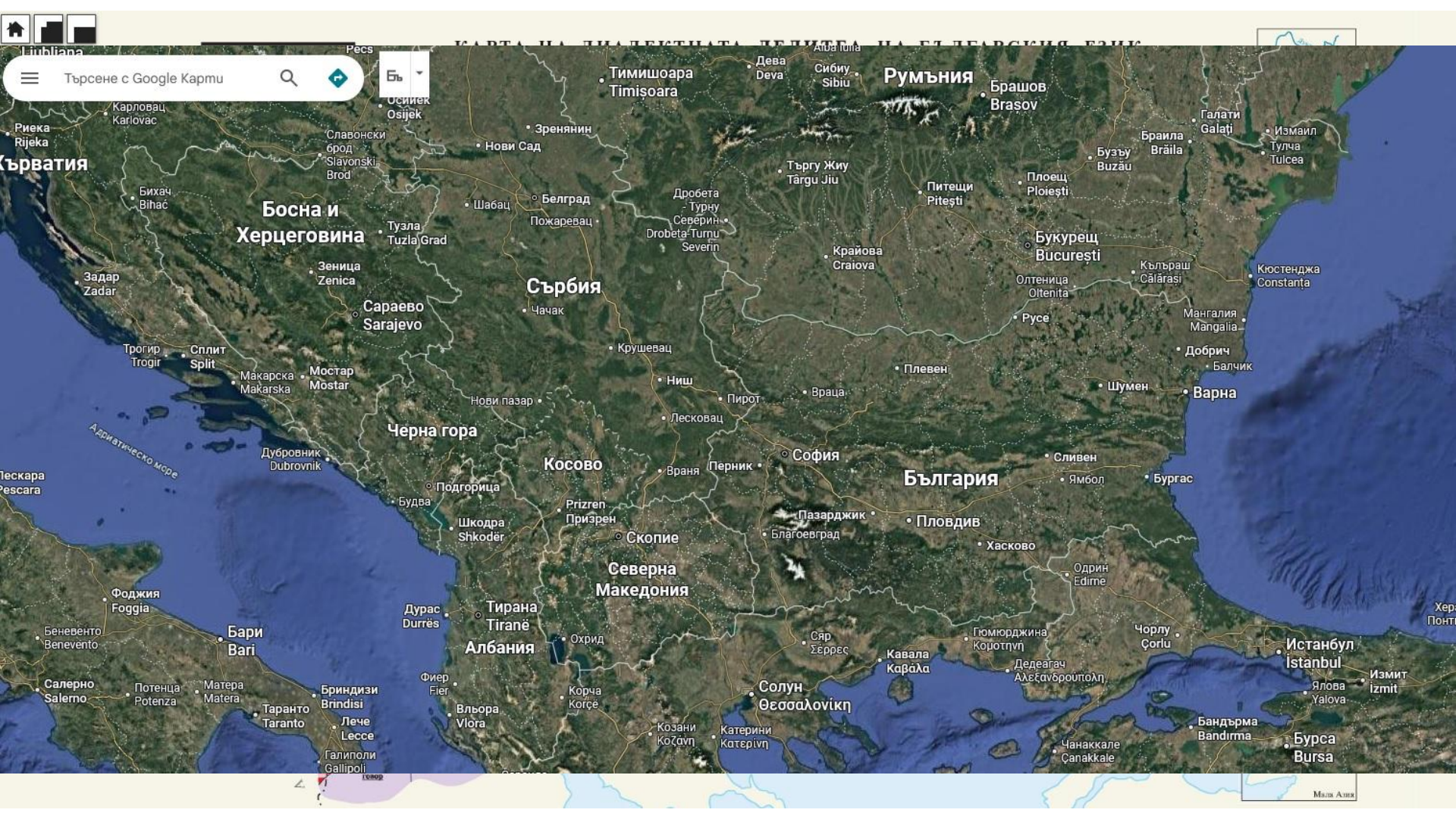


КАРТА НА ДИАЛЕКТНАТА ДЕЛИТБА НА БЪЛГАРСКИЯ ЕЗИК



Experiment with me...





Open Chat GPT and ask:

Prompt 1: *"Analyze the two maps and define the geographical features such as mountains, rivers, and elevation that determine the separation of the different dialects."*

Open Chat GPT and ask:

Prompt 2: *"Act as an expert in Bulgarian dialectology. Analyze the Interactive Dialect Map of the Institute for Bulgarian Language (IBL-BAS) and identify the Yat boundary (the isogloss of the Yat reflex). Based on this specific map, describe the trajectory and location of this boundary."*

Open Chat GPT and ask:

Prompt 3: *"Is it true that Bulgarian dialects in mountain regions possess **a richer system of consonants**?"*

Let's play

Results 1

ББ

ChatGPT каза:

По двете карти ясно се вижда, че **диалектното деление на българския език** се определя в огромна степен от **географските бариери** – планини, масиви, ридове, речни долини, котловини и исторически изолирани зони.

По-долу е цялостен анализ на основните географски фактори, които стоят зад разделянето на различните български диалекти.

1. ГОЛЕМИТЕ ПЛАНИНИ КАТО ГРАНИЦИ МЕЖДУ ДИАЛЕКТИ

★ *Стара планина (Балканът)* – най-силната разделителна линия

Това е **главната природна бариера**, определяща голямото деление на **севернобългарски** и **южнобългарски диалекти**.

Причини:

- висок и дълъг планински масив (над 500 км);
- исторически слаби проходи → ограничени контакти;
- различна търговска и административна ориентация през векове (Дунав vs Тракия).

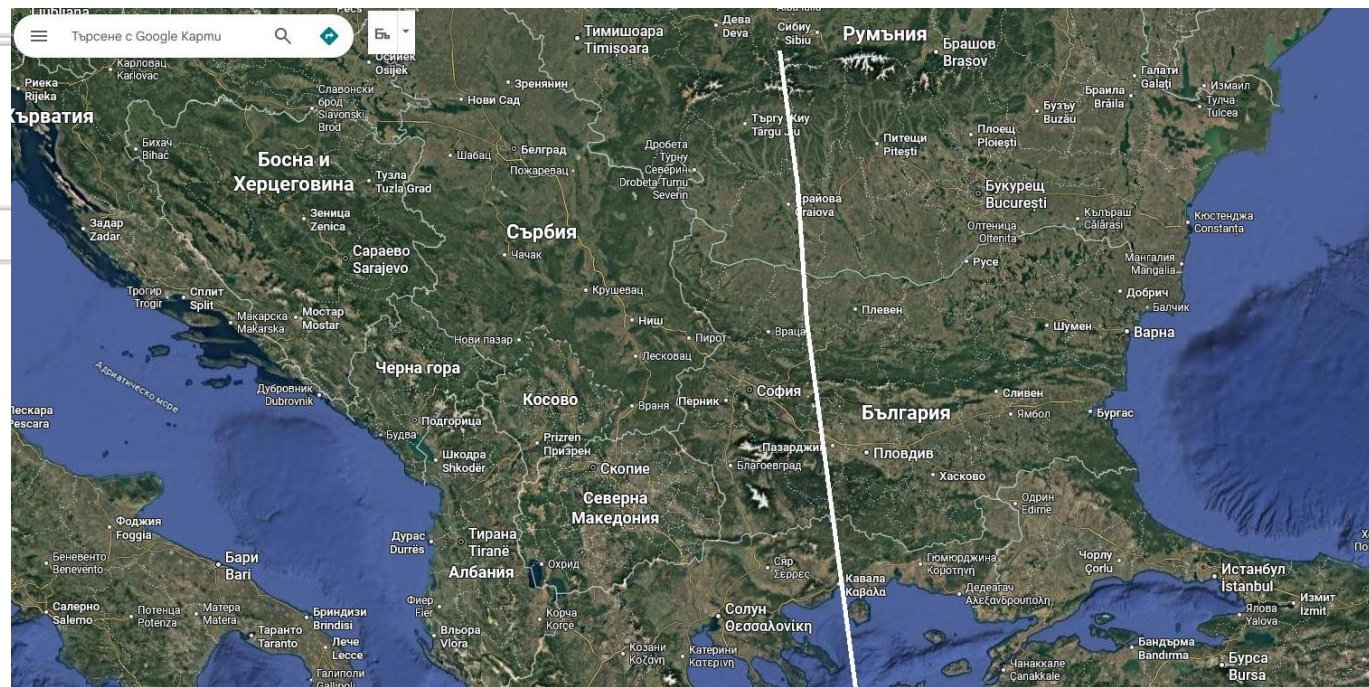
Резултат:

На картата ясно личи:

- на север – говори като *видински, ломски, белослатински, търновски, русенски, добруджански*;
- на юг – *балкански, странджански, родопски, тракийски* и др.

Results 2

ЂѢ



Results 3

Yes, it can be said that Bulgarian mountain dialects often possess a richer system of consonant sounds compared to lowland dialects. This is due to differences in the historical development of the language across various geographic regions of Bulgaria. Reasons for this include:

- 1. Local isolation: Many mountain regions in Bulgaria were historically relatively isolated from one another, which led to the preservation of archaic features and the development of specific local characteristics. Isolation often contributes to linguistic diversity and the preservation of older linguistic forms;*
- 2. Phonetic characteristics: In some mountain dialects, a larger number of consonant sounds or more complex consonant clusters can be observed. For example, in certain Rhodope dialects, consonants may be preserved in forms that differ from the standard language.....*

Да може да се каже, че българските планински диалекти често имат по-богата система от съгласни звуци в сравнение с равнинните диалекти. Това се дължи на различия в историческото развитие на езика в различните географски райони на България.

Причини за това:

- 1. Местна изолация:** Много от планинските райони в България са били сравнително изолирани в миналото, което е довело до по-слабо влияние на новите езикови тенденции от равнинните и големите градове. Изолацията е спомогнала за запазването на по-старите форми и особености в диалектите.
- 2. Фонетични особености:** В някои планински диалекти се наблюдава по-силно изразено съгласно редукция или употреба на допълнителни съгласни в някои форми на думите. Например, в част от планинските диалекти може да се срещне по-често употребата на звуци като "ж", "ш", "ч", както и различни варианти на "р", които в равнинните диалекти се срещат по-рядко или се изговарят по различен начин.
- 3. Фонетични изменения:** В резултат на езиковото развитие и влиянието на околната среда, в планинските диалекти могат да се наблюдават фонетични особености, като например преобладаване на определени съгласни или нови звукови комбинации, които не са характерни за по-високите и равнинни райони.

Примери:

- В някои планински диалекти, например в Пирин и Рила, има по-ясно изразено разграничаване на съгласни звуци като "г" и "д", което не се среща толкова често в равнинните диалекти.
- Планинските диалекти също така могат да имат по-силно изразени или различни форми на съгласните "ц", "ч" и "ш", които често се изговарят по различен начин в различните части на страната.

Това не означава, че всички планински диалекти задължително са по-богати на съгласни звуци, но е често срещано явление.

Conclusions

- Model performance relies heavily **on prompts**; without expert guidance, visual geography triggers a flawed **North–South split** instead of the linguistic **East–West** boundary.
- Leading questions prompt **plausible but unsupported explanations**, proving that both visual features and prompt design distort AI reliability.
- Results are exploratory due to testing a single model (ChatGPT 4o), a single phenomenon (Yat boundary), and a limited prompt pool.

Conclusions

- **Future Research:** Next steps will evaluate open-source LLMs, analyze multiple dialect datasets, and run repeated trials to measure quantitative response consistency.
- **Broader AI Safety Impact:** The model's vulnerability to suggestion shows it can easily generate persuasive, pseudo-scientific disinformation.
- **Urban Planning Risks:** If AI prioritizes physical geography over abstract socio-cultural boundaries, automated urban models risk miscalculating infrastructure and perpetuating biases.

Study Limitations

- **Single Model Limitation:** Testing relied exclusively on ChatGPT-4, making results model-specific and non-generalizable.
- **Single Linguistic Phenomenon:** The analysis focused solely on the Yat boundary, omitting other phonological or lexical features.
- **Limited Cartographic Sources:** The experiment combined just one linguistic map with a single topographic reference.

Study Limitations

- **Qualitative Approach:** The exploratory design used a small prompt pool without standardized quantitative benchmarks.
- **No Consistency Testing:** Data came from limited interactions without evaluating stability across repeated model runs.
- **Task Specificity:** Findings apply strictly to dialect map interpretation and cannot generalize to all linguistic reasoning.

Share your results with me...

ruslana.margova@gate-ai.eu

Thank you!