dmi

People-Centred AI
UNIVERSITY OF SURREY

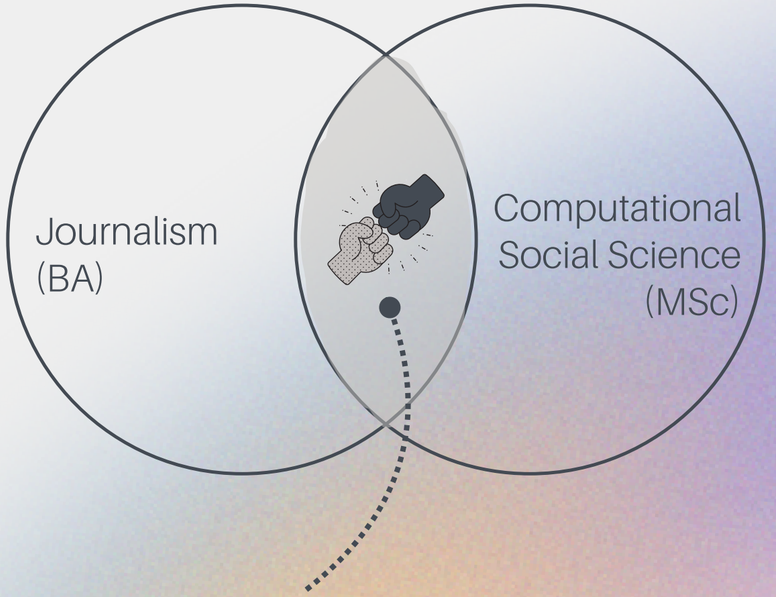
Towards Efficient, Accessible Geoparsing of Local Media

AKA **LOCATION EXTRACTION**, FROM TEXT

A BENCHMARK DATASET +
LLM-BASED APPROACH

Simona Bisiani | 5.10.2025 | Data Methods Initative Seminar

a short **history** of **me**

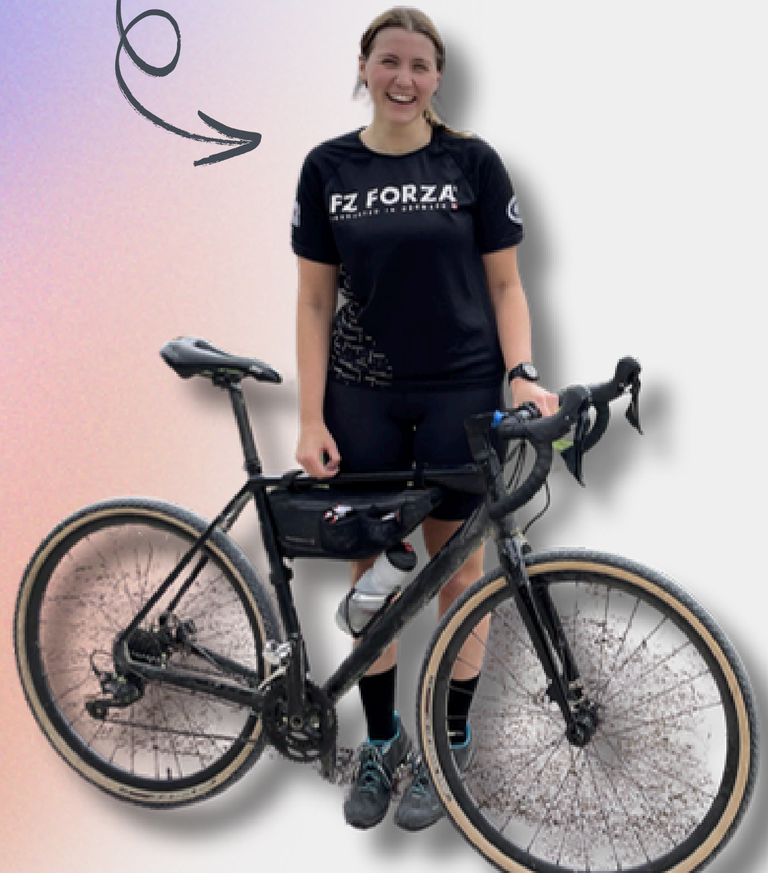


phd in **journalism**, using **css tolbox**!
based at Surrey Institute for People-Centre AI
focus on studying **local journalism in the UK**

HOW **LEARNING TO CODE** FELT LIKE

tears first

joy later



- Local media **database curator** + researcher at Public Interest News Foundation

- Come to identify myself as a '**local news cartographer**'

www.publicinterestnews.org.uk/map

simonabisiani.substack.com

where do we start...

locations inside news articles are **important**, but **extracting** them is **challenging**

table of contents

01

motivation

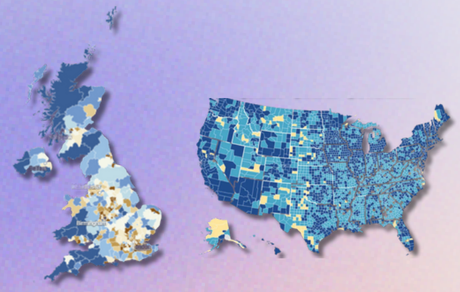
the why of going through all of this

local journalism

disruption and challenges

digital transformation

- news outlets closures[1] → *'news deserts'*[2,3]



- ⅔ jobs gone[4] → intensified work pressure[5] → 'churnalism', content recycling, minimal field reporting[6]
- mergers / acquisitions[7] → 3 companies own > 50% of the market[8]

  - newsrooms centralisation[9]
  - content syndication[11]

so what?

fewer companies → lack of pluralism / diversity[15]

no civic news → fewer voters[14]

no local information supply → proliferation of misinformation[16]

no watchdog → higher corruption[13]

local journalism

disruption and challenges

digital transformation

- news outlets closures[1] → *'news deserts'*[2,3]

→ *can we move beyond counting news outlets?*

- ⅔ jobs gone[4] → intensified work pressure[5] → 'churnalism', content recycling, minimal field reporting[6]
- mergers / acquisitions[7] → 3 companies own > 50% of the market[8]

- newsrooms centralisation[9] → *distance decay:* distance impacts relevance[10]
- content syndication[11]

→ **fewer local stories**[12] → *can we move beyond counting news outlets?*

In the new millennium, **local newspapers are local in name only**; the town or city emblazoned on the newspaper's masthead may be one of the few remaining local features of the paper.

Franklin (2006, p.xxi)

can we **empirically validate** this claim?

02

location retrieval from text

a crash course

geoparsing is the process of extracting **location references** from **unstructured text** and converting them into **structured geographic data**

"A planning application for 50 new homes in **Bramley** has been submitted to **Leeds** City Council. The development on **Station Road** would include affordable housing."

geoparsing

geoparsing

geotagging *aka* toponym recognition

Find location mentions in text

3 main approaches

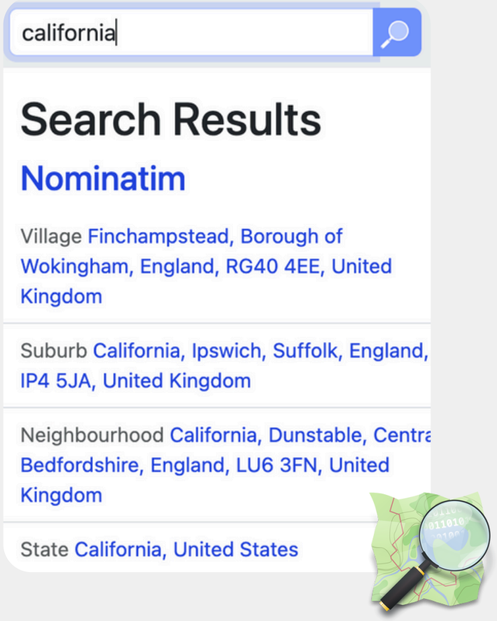
- Rule-Based
- Machine-learning Based
- Hybrid

Frameworks

- SpaCy
- NLTK
- Stanford NER

candidate selection

Search for matching real-world locations



geocoding *aka* toponym resolution

Resolve ambiguities

3 main approaches

- Rules
- Learning & ranking
- Learning & classification

Heuristics

- Population
- Proximity to other locations in the same text
- *One-sense-per-referent*

geocoding is ***so* not fun**

best performance requires **complex**, specialised ML models

high technical barriers to implementation

limited generalisability across domains and regions

no knowledge of method performance on local news

03

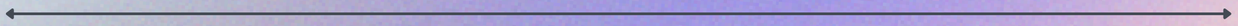
wait... can i use LLMs for that?

a glimmer of hope

key challenge is

context understanding

worse — *better*



traditional toponym resolution

LLMs

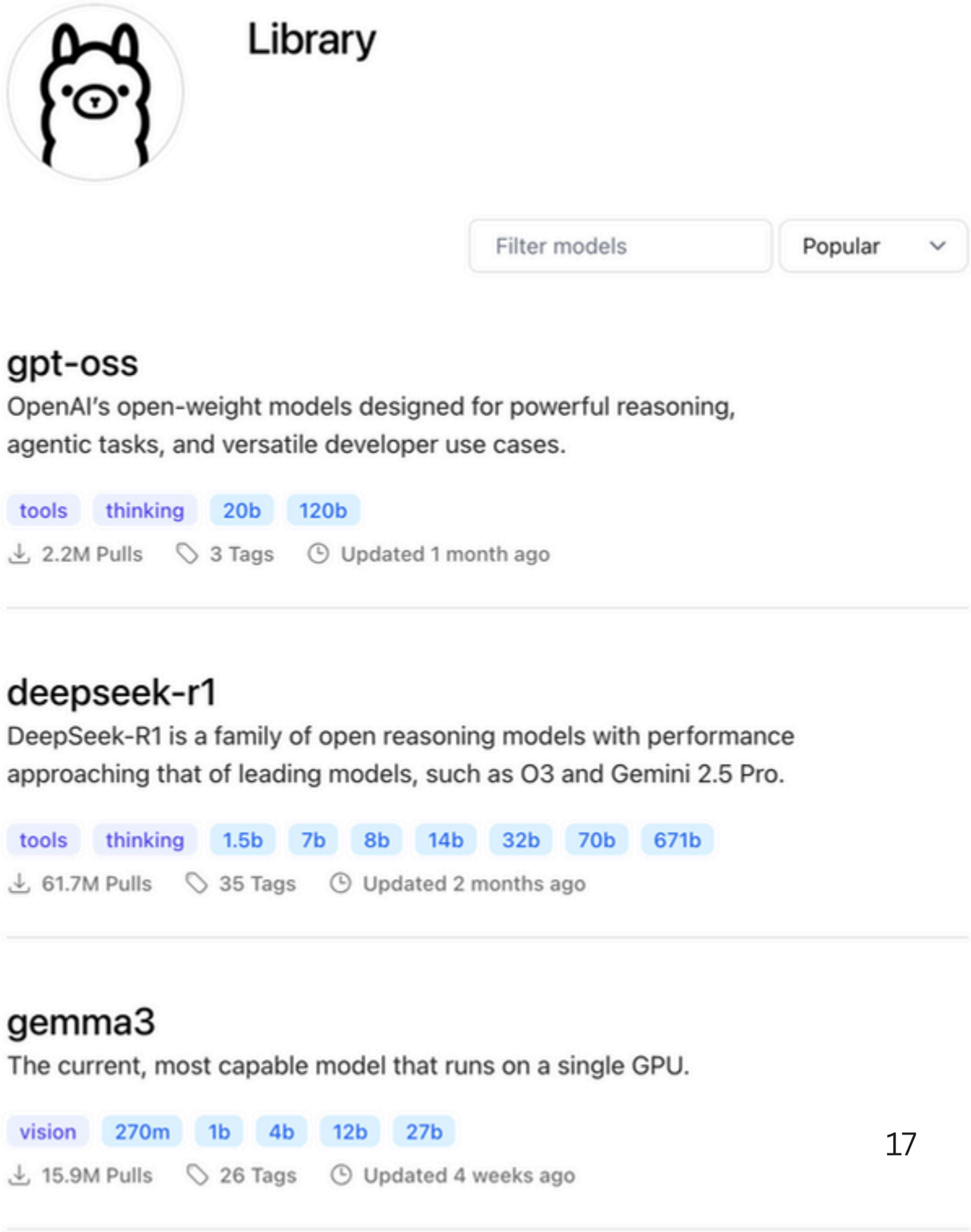
OLLAMA

LLMs, made simple (possible?)

- privacy: your data stays on your machine
- no API costs: run models for free locally
- *you pay for the energy you consume
- models are 'lightweight' (quantized)
- offline: works without internet connection

Resources:

- Ollama install page
- Matt Williams YouTube channel



full pipeline

data

toponym recognition

candidate retrieval

spatial calculations

toponym resolution

200 articles from UKTwitNewsCor

SpaCy
F1 score: 0.94

2 x human annotators
https://prodi.gy/

Ordnance Survey + Open Street Map

Albury,
51.2198° N,
0.5002° W

Albury, Guildford

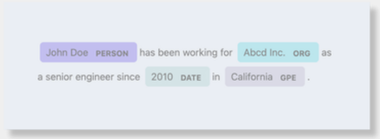
LLMs

2 x human annotators
https://labelstud.io/

gold data /
benchmark dataset
(1,314 locations)

our pipeline

toponym recognition



candidate retrieval

spatial calculations

toponym resolution

---

state-of-the-art (Hu et al., 2025)

toponym recognition

Albury

fine-tuned LLM

Albury, United Kingdom

candidate retrieval

pick n. 1 candidate, based on string similarity and population

higher technical difficulty

problematic for local media

prompt engineering

same as human annotators!

classification

Map the entity (a toponym) to the correct Local Authority District (LAD) from the options. Instructions:

1. *Review Entity and Article context.*

2. *Check Metadata where provided:*
- Outlet coverage LAD, Domain, and other Entities and candidates (for hints on location).

3. *Choose the best option:*
- Select a LAD from the list provided, or choose from "LAD not in options," "Entity is not a location," "Entity is outside the UK," "Entity spans multiple districts," or "Unsure."

4. *Format response as JSON:*
{ 'chosen_option': 'Your choice', 'reasoning': 'Your reasoning' }

how much does it know geography?

zero-shot

Your goal is to deduce in which UK's Local Authority District (LAD) the entity (a toponym) in question is situated. Instructions:

1. *Look at the entity provided and read the article* carefully to understand the context.

2. *Check Metadata:* the publisher's domain is provided as it may provide geographic context.

3. *Determine* which Local Authority District should be associated to the entity.

4. Format response as JSON:
{ 'chosen_option': 'Your choice', 'reasoning': 'Your reasoning' }

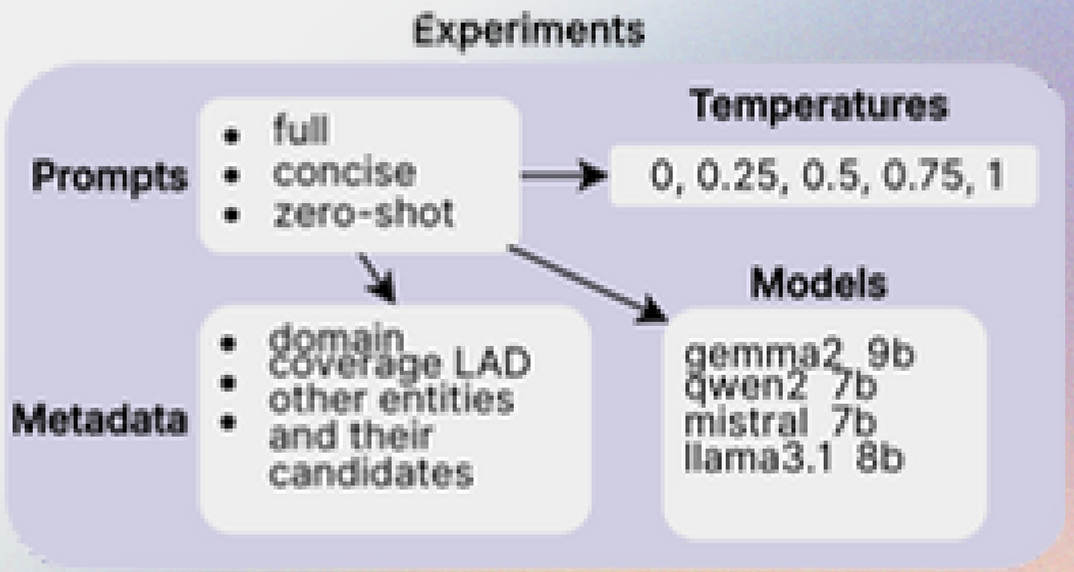
Chain of Thought (CoT)

few-shot prompting

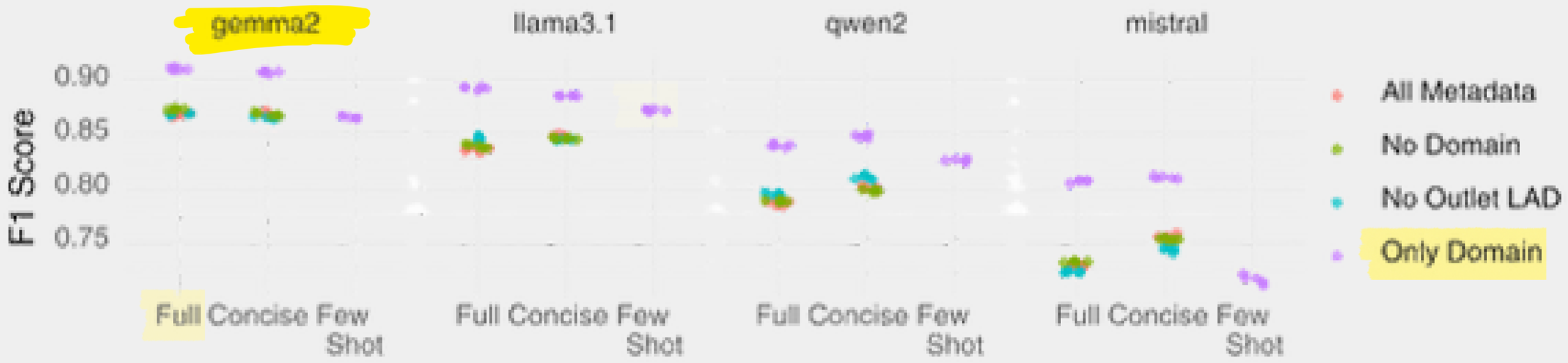
Example 1:
Entity: King's Head pub.
Article: *Incident outside the King's Head pub on Main Street, Guildford.*
Domain: guildforddragon.co.uk.
Output: {'chosen_option': 'Guildford', 'reasoning': 'The article mentions Main Street, Guildford.'}.

Example 2:
Entity: Dublin.
Article: *Dublin has experienced a lot of rain lately.*
Domain: belfasttelegraph.co.uk.
Output: {'chosen_option': 'Entity is a location outside the UK', 'reasoning': 'Dublin is located in Ireland, not in the UK.'}.

llm experiments



results (1/2)

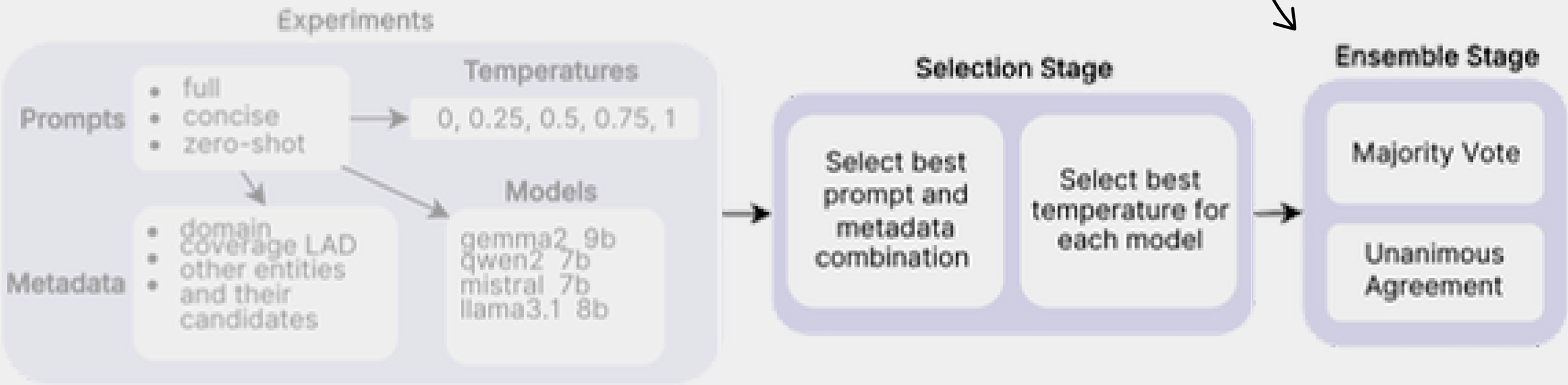


WHAT DRIVES PERFORMANCE?

Model used | Metadata | Prompt | Temperature

llm experiments

LLM-as-judge: ask different models, make a vote. Select majority choice.



results (2/2)

Select a LAD from the list provided, or choose from "LAD not in options," "Entity is not a location," "Entity is outside the UK," "Entity spans multiple districts," or "Unsure."

	Classification Metrics		Spatial Metrics	
	Task (model output)	**System** (pipeline as a whole)	**A@161**	**Mean Error (km)**
Baseline Most-voted option; random if no majority	**.87**	**.78**	**.86**	2667
Majority Keep only instances where at least 2+ models agree (**90%**)	.93	.85	.89	2035
Unanimous Keep only instances where all models agree (**68%**)	.97	.97	**1**	75
Benchmark *Hu et al. (2024) fine-tuned LLM*		**.80**	**.99**	22

where can we do better?

candidate retrieval

Why are some correct options missing?

toponym resolution

Can we get even better results with different models or prompts?

04

scaling up

challenges and considerations

here's the ugly truth

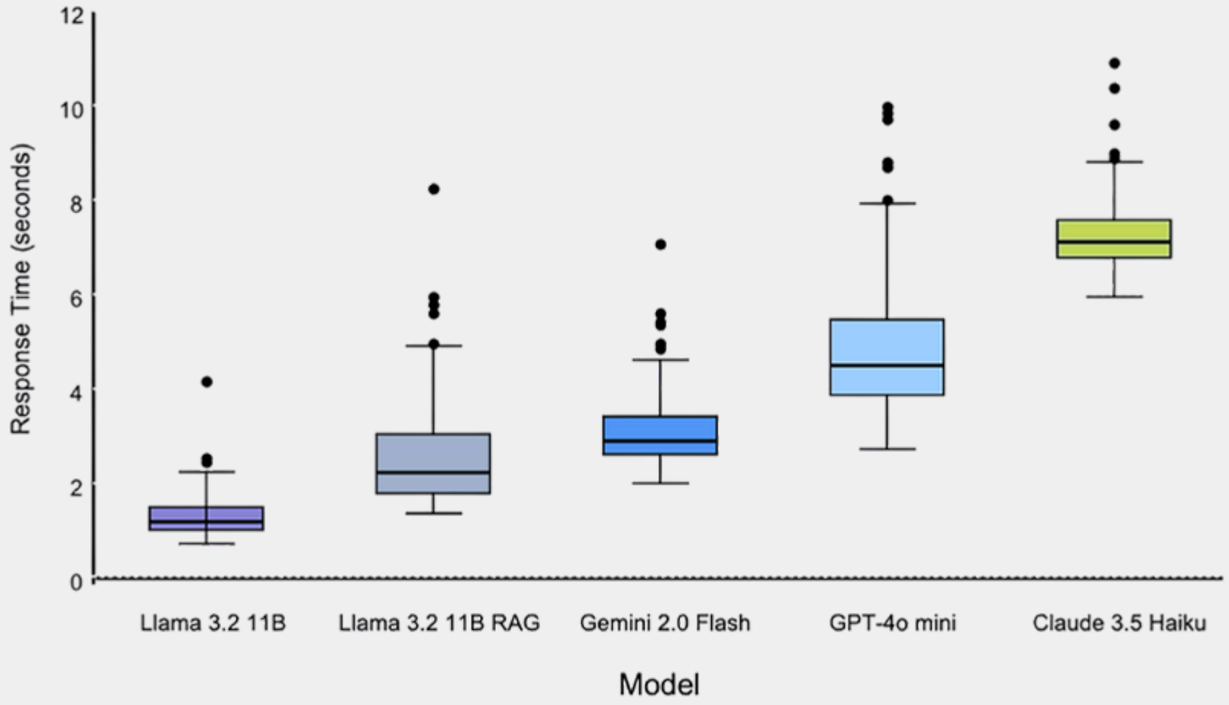
LLM inference,
how quickly a model processes a request and generates a response, **isn't fast.**

BASICS

response times

Factors

- Model size (7B vs 70B parameters)
- Hardware (GPU vs CPU, GPU memory)
- Prompt length
- Response length
- Whether model is already loaded in memory



https://www.nature.com/articles/s41746-025-01802-z

BASICS

how ollama works

api call

a structured way to send a request to a server, and send back a response.

port

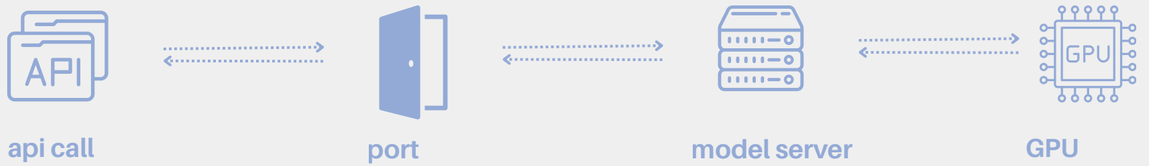
a "door" on your computer that programs use to talk to each other.

model server

a program running in the background that loads the model and waits for requests

GPU

the workhorse, holds the model in memory and processes the request

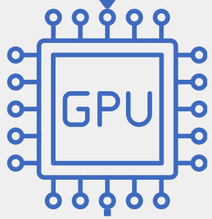


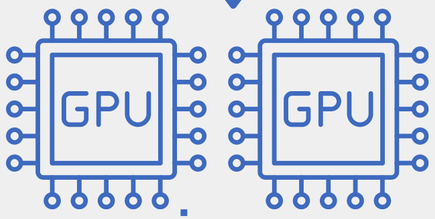
ADVANCED

parallelisation

a task is ***embarrassingly parallel*** if it's individual units don't depend on each other

how many **GPUs** do you have?





One model, tune **OLLAMA_NUM_PARALLEL** based on GPU performance

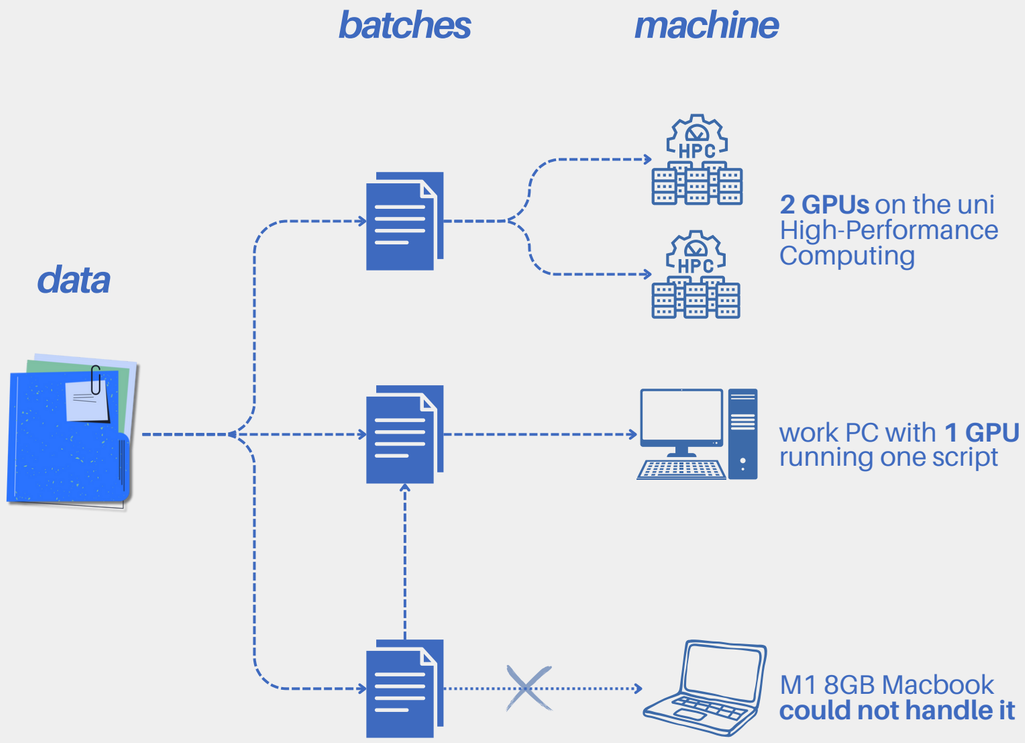
Multiple models on same GPU

- memory issue
- only different models allowed

One server per GPU, with a unique port

trial and error...

what worked for me



Deployed this procedure to **~400.000 news articles.**

trial and error...

how long this whole thing took

toponym recognition

A couple of days on the uni cluster, on *1.6m articles, not locations*

candidate retrieval

A couple of days, on *1.6m articles, not locations*

toponym resolution

One month on *400k articles (18% of data), 1.6m locations*

\+

countless days spent studying, debugging....

TIPS

lessons learned

01

pilot on a sample

02

measure total time

03

plan

04

batch your data

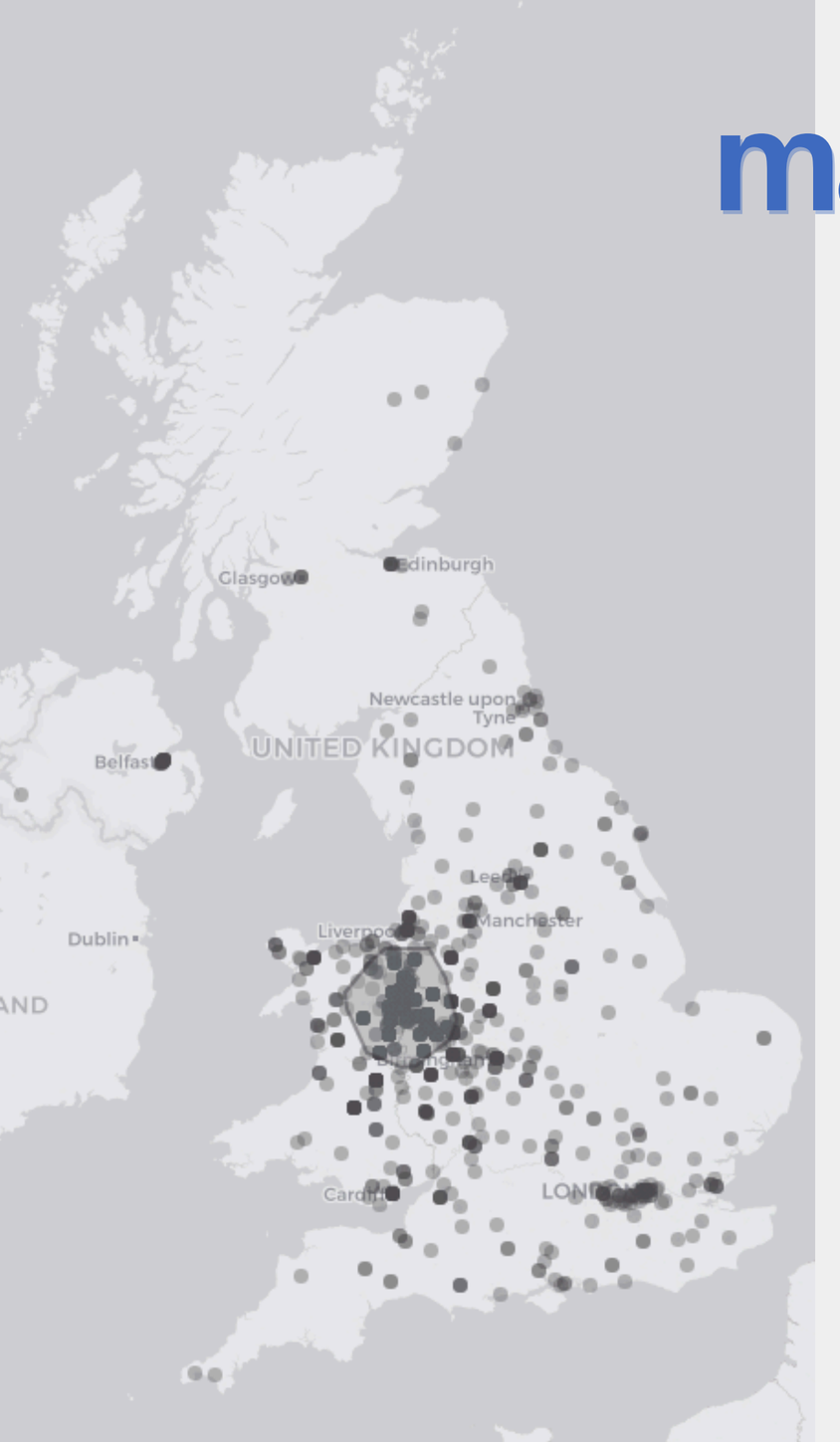
05

monitor progress

Batch	N_data_points	Finished	Remaining	Run where	File_name
	1,657,425	1,657,425	0		100%
28	25000	yes	.db	local_machine	llm_results_23_32.db
29	25000	yes	.db	local_machine	llm_results_23_32.db
30	25000	yes	.db	local_machine	llm_results_23_32.db
31	25000	yes	.db	local_machine	llm_results_23_32.db
32	25000	yes	.db	local_machine	llm_results_23_32.db
33	5000	yes	.rds	HPC	batch_{number}_combined.rds
34	5000	yes	.rds	HPC	batch_{number}_combined.rds
35	5000	yes	.rds	HPC	batch_{number}_combined.rds
36	5000	yes	.rds	HPC	batch_{number}_combined.rds

https://simonabisiani-local-news-map-explorer.share.connect.posit.cloud/

mapping local news coverage

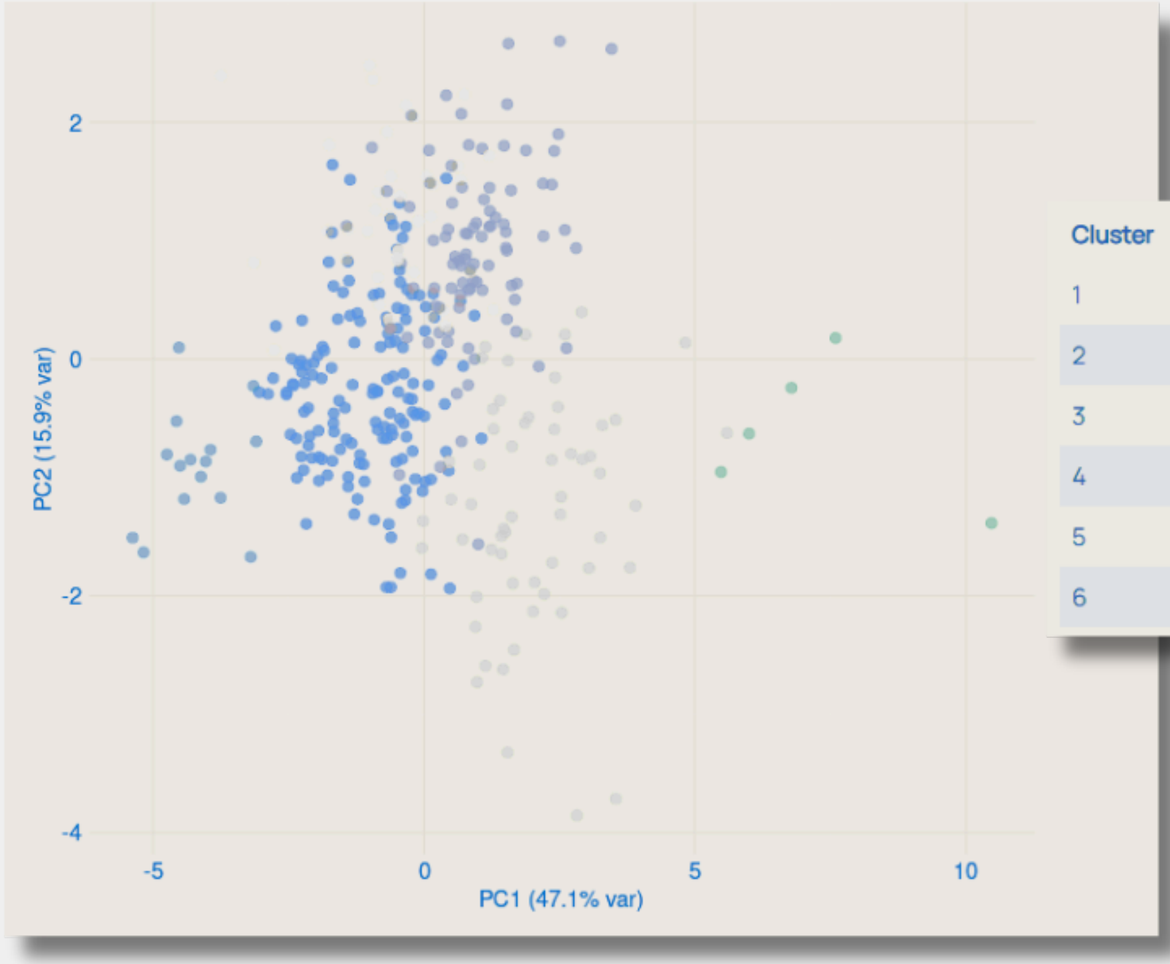


We constructed a four-dimensional framework to describe news coverage based on its spatial properties

Dimension	Metric	Definition	Interpretation
Spatial Extent	Area (km^2)	Area of the minimum convex polygon enclosing 75% of locations, by frequency	Larger area → broader territorial coverage within the core region
Administrative Reach	Districts	Administrative districts (N) covered by outlet's location mentions	More districts → wider cross-boundary coverage
Spatial Heterogeneity	Entropy	Shannon entropy of mention proportions across districts	Higher values → more balanced coverage across districts
Distance Decay	Within 10km (%)	Fraction of all mentions falling within 10km of the outlet primary location	Higher values → more localised coverage

data-driven classification of subnational media

https://simonabisiani-clustering-analysis-dashboard.share.connect.posit.cloud

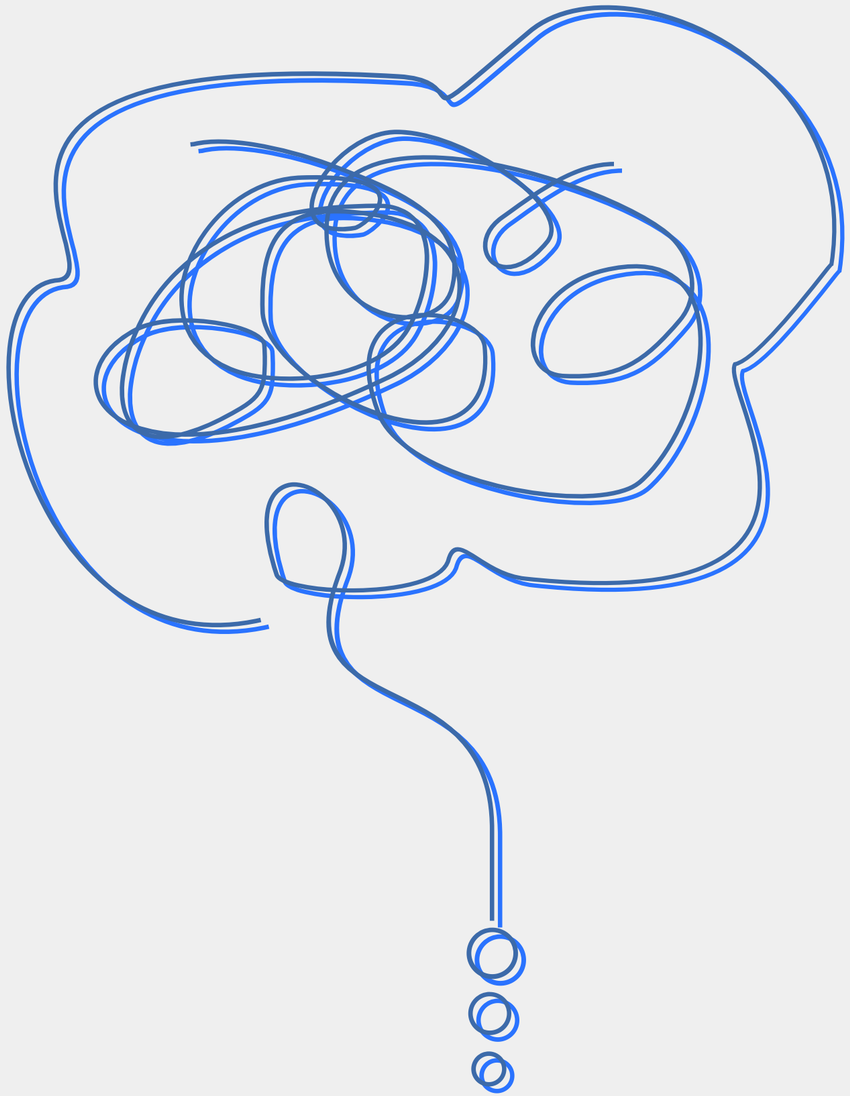


Cluster	Label	Size	Area
1	National	5	79013.873
2	Local-Regional	152	1215.73
3	Metropolitan	90	5768.969
4	Hyperlocal	15	32.957
5	Sub-Regional	58	7388.752
6	Rural	38	2715.293

ongoing

- studying the **relationship between ownership** structure **and** local news coverage **proximity**

- producing an **assessment of local news provision** in the UK that uses coverage distribution instead of outlets location



thanks!

references

1. Turner, A. (2022, July 26). UK local newspaper closures: Launches in digital and print balance out decline. Press Gazette. https://pressgazette.co.uk/news/uk-local-newspaper-closures-2022/
2. Bisiani, S., & Mitchell, J. (2024). UK Local News Report April 2024. Public Interest News Foundation.
3. Metzger, Z. (2024). The State of Local News—2024 Report. Local News Initiative, Medill School of Media, Journalism, and Integrated Marketing Communications, Northwestern University.
4. Ponsford, D. (2024, February 15). Colossal decline of UK regional media since 2007 revealed. Press Gazette. https://pressgazette.co.uk/publishers/regional-newspapers/colossal-decline-of-uk-regional-media-since-2007-revealed/
5. Davies, N. (2011). Flat earth news: An award-winning reporter exposes falsehood, distortion and propaganda in the global media. Random House.
6. Champion, K. (2015). Measuring content diversity in a multi-platform context. The Political Economy of Communication, 3(1).
7. Doyle, G. (2002). Media Ownership: The Economics and Politics of Convergence and Concentration in the UK and European Media.
8. Media Reform Coalition. (2025). Who Owns the UK Media? Goldsmiths Leverhulme Media Research Centre.
9. Sharman, D. (2021). Reach plc to close all bar 15 of its newspaper offices—Journalism News from HoldtheFrontPage. HoldtheFrontPage. https://www.holdthefrontpage.co.uk/2021/news/publisher-to-close-all-bar-15-offices-leaving-dailies-without-base-on-patch/
10. Karlsson, M., & Rowe, E. H. (2019). Local Journalism when the Journalists Leave Town. Nordicom Review, 40(s2), 15–29.
11. Garz, M., & Ots, M. (2025). Media consolidation and news content quality. Journal of Communication, jqae053.
12. Franklin, B. (Ed.). (2006). Local journalism and local media: Making the local news. Routledge.
13. Usher, N., & Kim-Leffingwell, S. (2023). How Loud Does the Watchdog Bark? A Reconsideration of Losing Local Journalism, News Nonprofits, and Political Corruption. International Journal of Press/Politics.
14. Rubado, M. E., & Jennings, J. T. (2020). Political Consequences of the Endangered Local Watchdog: Newspaper Decline and Mayoral Elections in the United States. Urban Affairs Review, 56(5), 1327–1356.
15. Hendrickx, J., & Ranaivoson, H. (2021). Why and how higher media concentration equals lower news diversity – The Mediahuis case. Journalism, 22(11), 2800–2815.
16. Altay, S., Nielsen, R. K., & Fletcher, R. (2024). News Can Help! The Impact of News Media and Digital Platforms on Awareness of and Belief in Misinformation. The International Journal of Press/Politics, 29(2), 459–484.

code demo

01

notebook

https://simonabisiani.github.io/geoparsing-notebook/live-coding-demo.html

02

code

https://github.com/simonabisiani/geoparsing-notebook

People-Centred AI
UNIVERSITY OF SURREY

dmi

www.simonabisiani.github.io