Thesis Defends

Employing Large Language Models for Systematic Extraction of Nanoparticle Designs from Scientific Literature

Presented by: Halidu Abdulai

Supervisors: Cristina Suemay Manresa Yee, Sebastien Lafond, and Hergys Rexha

Date; 26th June 2025

EDISS
MASTER'S PROGRAMME

Why does this Research Work Matter?

- **179+ million** Europeans (**1 billion** people globally) live with brain diseases [1]
- EU healthcare spends **€800 billion every year** in the fight against these diseases [2]
- Their effective treatment is hindered by the **blood-brain barrier (BBB)**, with **less than 5%** of candidate drugs for brain diseases capable of **effectively penetrating** the BBB [3]

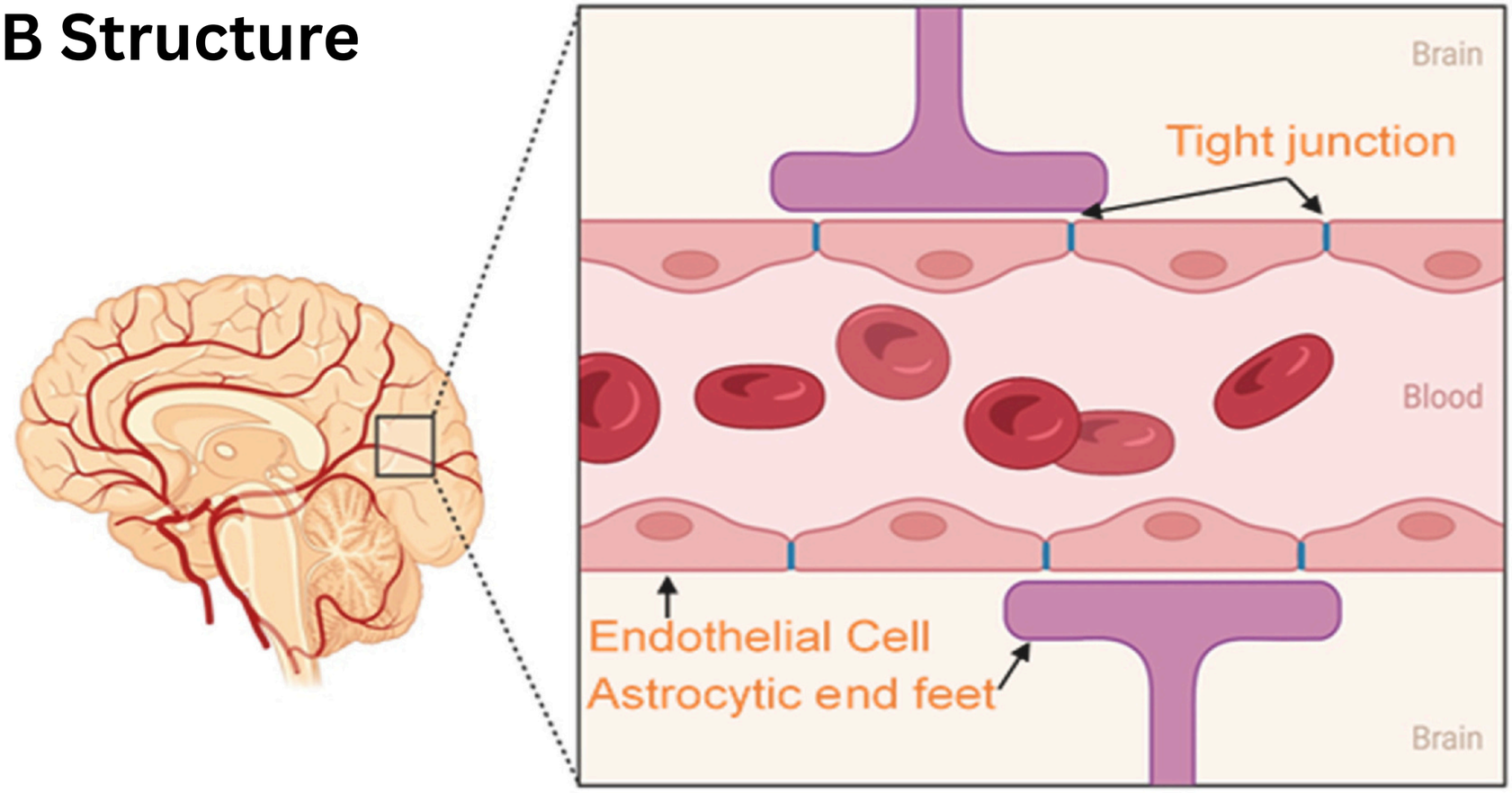
[1] *Banks W (2016); Dong X (2018)*

[2] *DiLuca M, Olesen J (2014)*

[3] *Partridge B, et al. (2022)*

The Brain Blood Barrier (BBB)

The BBB Structure



SRC: Ivanov, A., Zhang, J., & Smith, B. (2024). Brain barrier challenges for advanced drug delivery.

Use of Nanoparticles in Drug Delivery

- Nanoparticles (NPs) offer a **solution for drug delivery across the BBB**, yet translational research has been limited [4]
- **30+ years** of research → **1500+** published reports
- **Fewer than 30** NP systems have reached clinical trials → **<5%** of all clinical trials on nanomedicine.
- **850.000+** animals are used **annually** in the EU in nervous-system related research [5]

[4] *Mohapatra P, et al. (2024); Janjua TI, et al. (2023)*

[5] *Home | ClinicalTrials.gov*

The Solution: NAP4DIVE

NAP4DIVE

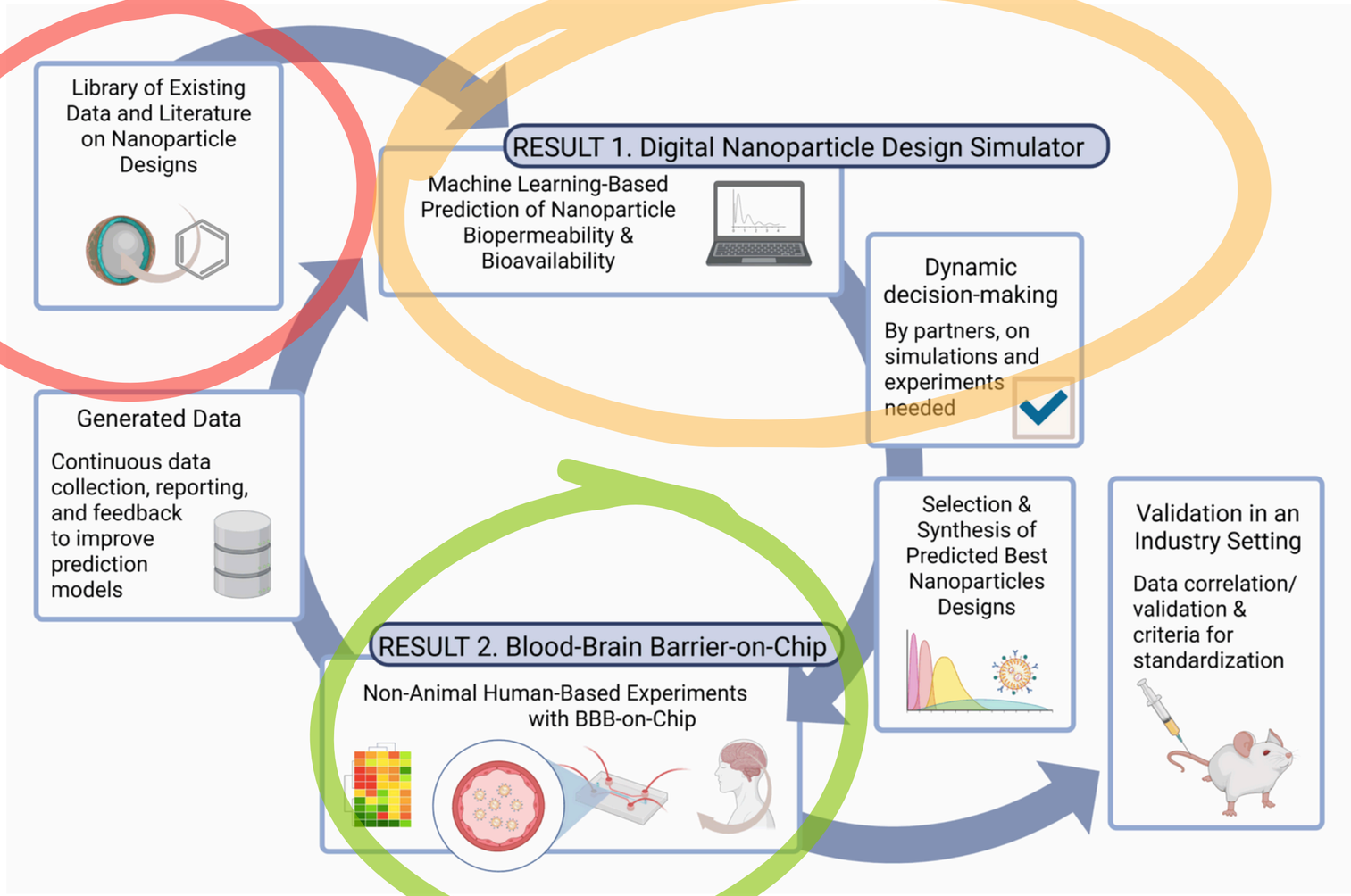
NON-ANIMAL PLATFORM FOR NANOPARTICLE-BASED DELIVERY ACROSS THE BLOOD-BRAIN BARRIER INTERFACE WITH VEHICLE EVOLUTION

What?

NAP4DIVE aims to reduce animal use in Central Nervous System drug development by up to 95% while saving 30 % of costs.

Aims of NAP4DIVE

Develop two complementary platforms:



Extracting NP Design Details from Articles

We want to extract NP design information, but…

- Reported inconsistently across publications
- Time-consuming and labor-intensive

Credit: This image was generated using Dalle-E 3

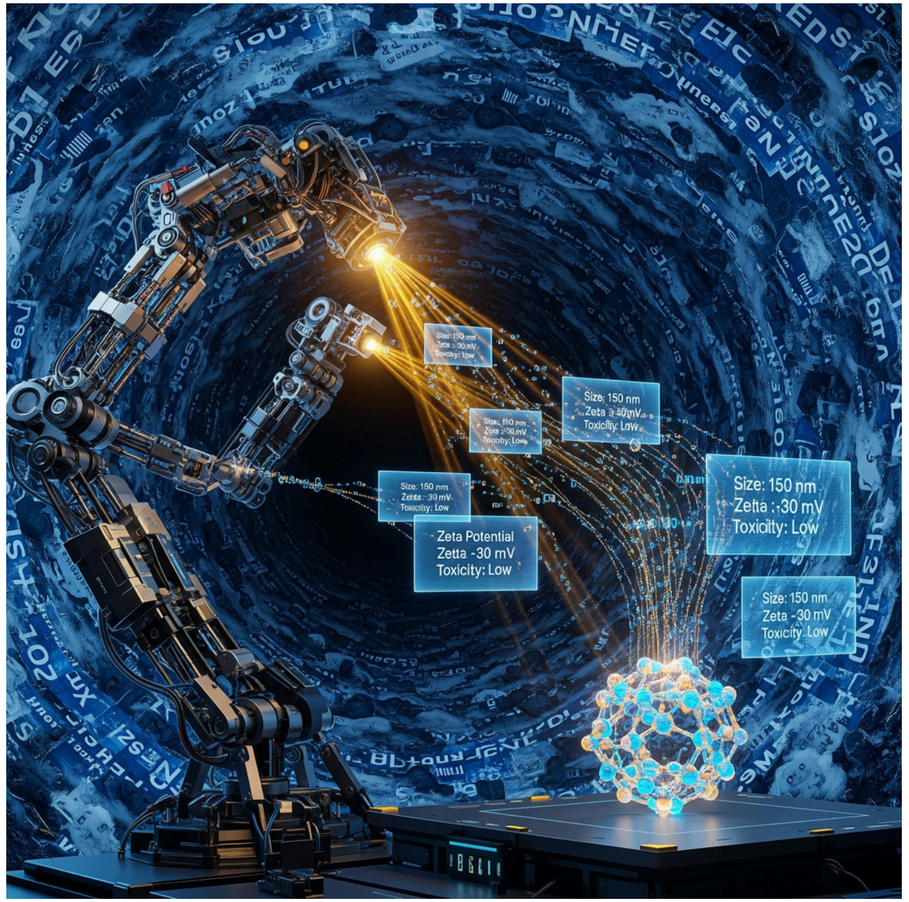
Solution: LLMs to the Rescue

“Employing Large Language Models for Systematic Extraction of Nanoparticle Designs from Scientific Literature”

LLMs:

- Are powerful for text processing
- Generalizes well to other domains

Leverage these capabilities to extract NP design parameters



Credit: This image was generated using Dalle-E 3

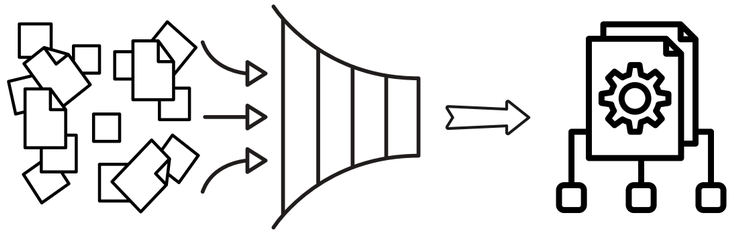
Limitations of Existing tools

Existing tools have proven efficient but...

- **No explicit work** on extraction of nanoparticle design parameters
- Extract **limited subsets** of data (e.g., bulk modulus, ligand names) without relational linking
- Depends on **Proprietary** LLMs (e.g., GPT, Claude)
- **Not optimized** for resource-constrained environment

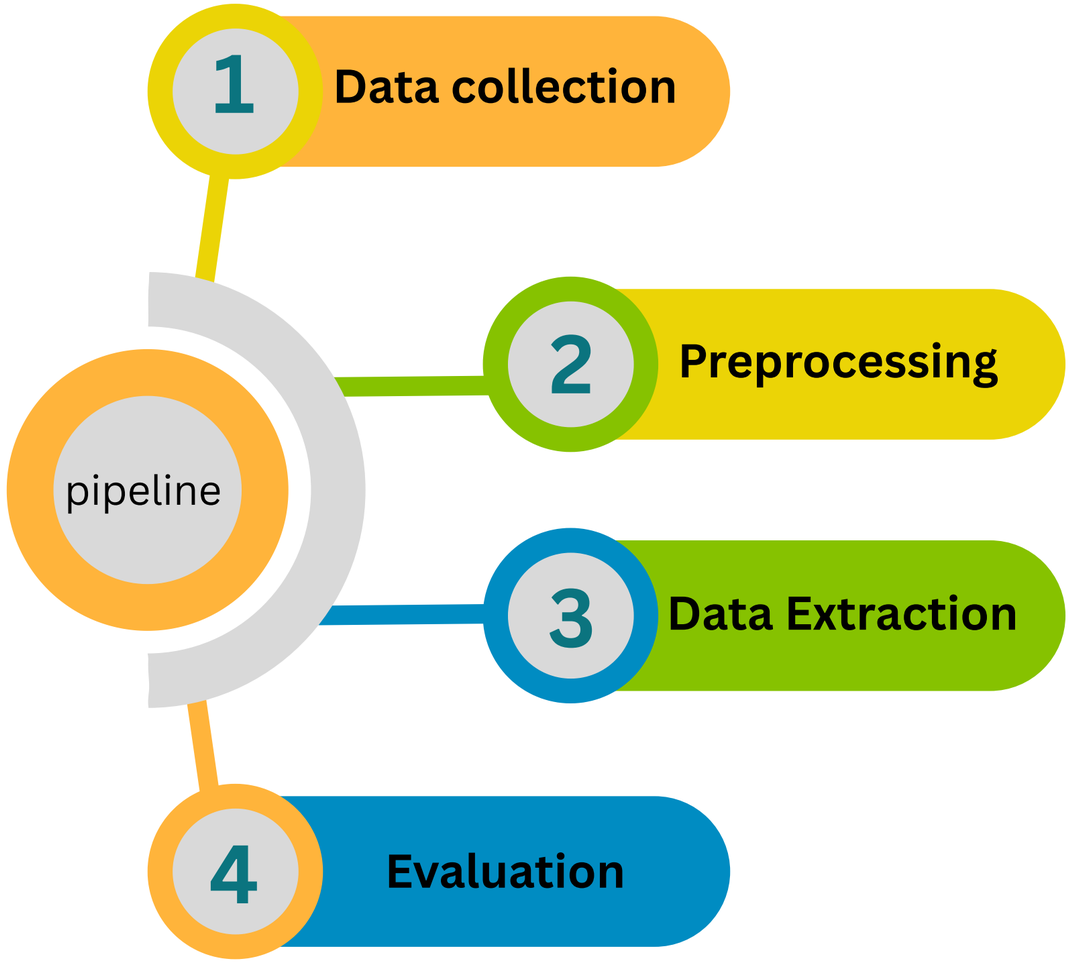
Addressing these Limitations

This thesis work addresses these concerns!

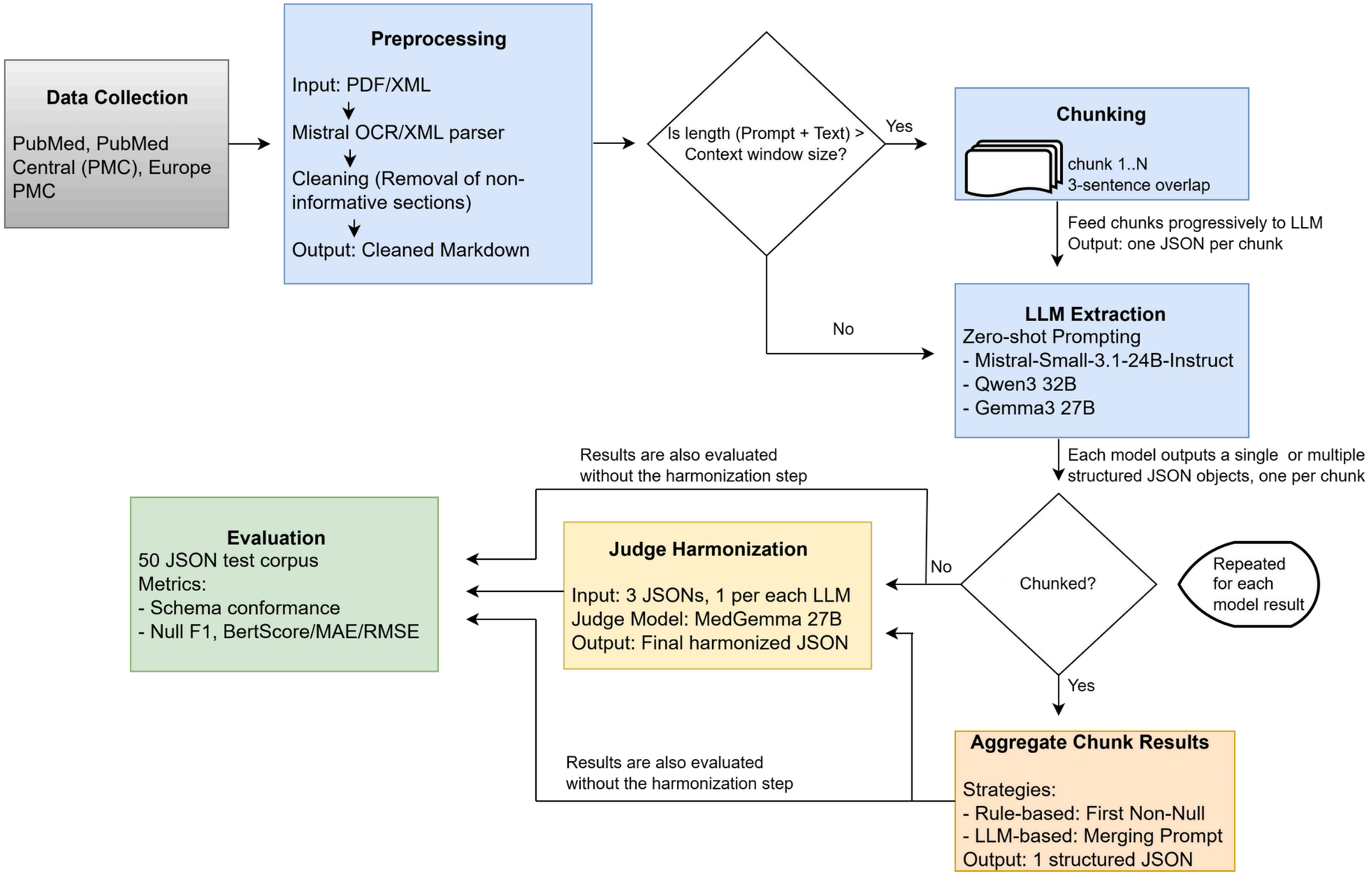


- Extract **comprehensive design** profiles
- Operates with **open-weight LLMs** deployable on **consumer-grade** hardware
- Tailored for an **underexplored** neurotherapeutic application domain

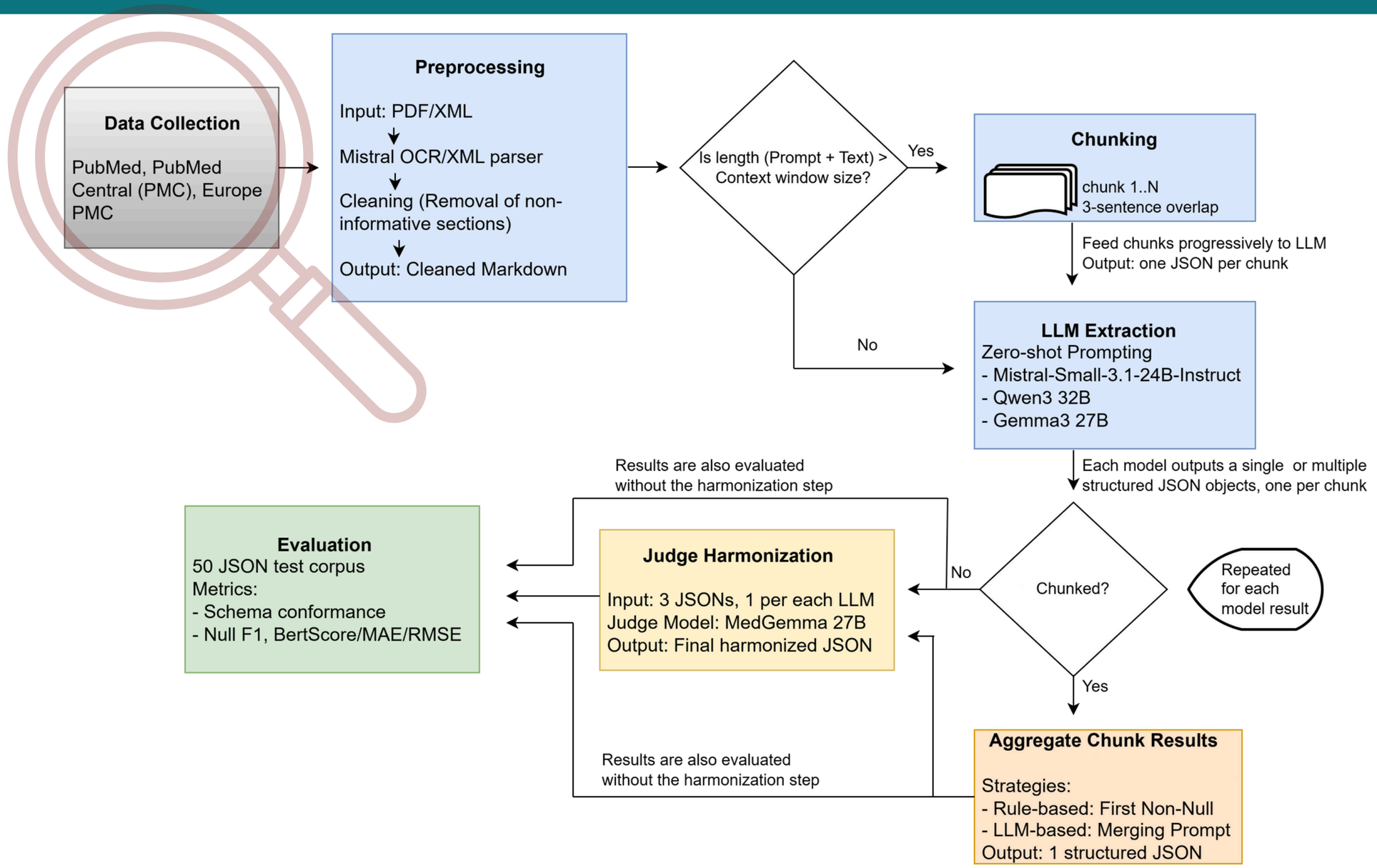
Pipeline Architecture



Pipeline Architecture



Data Collection: PubMed

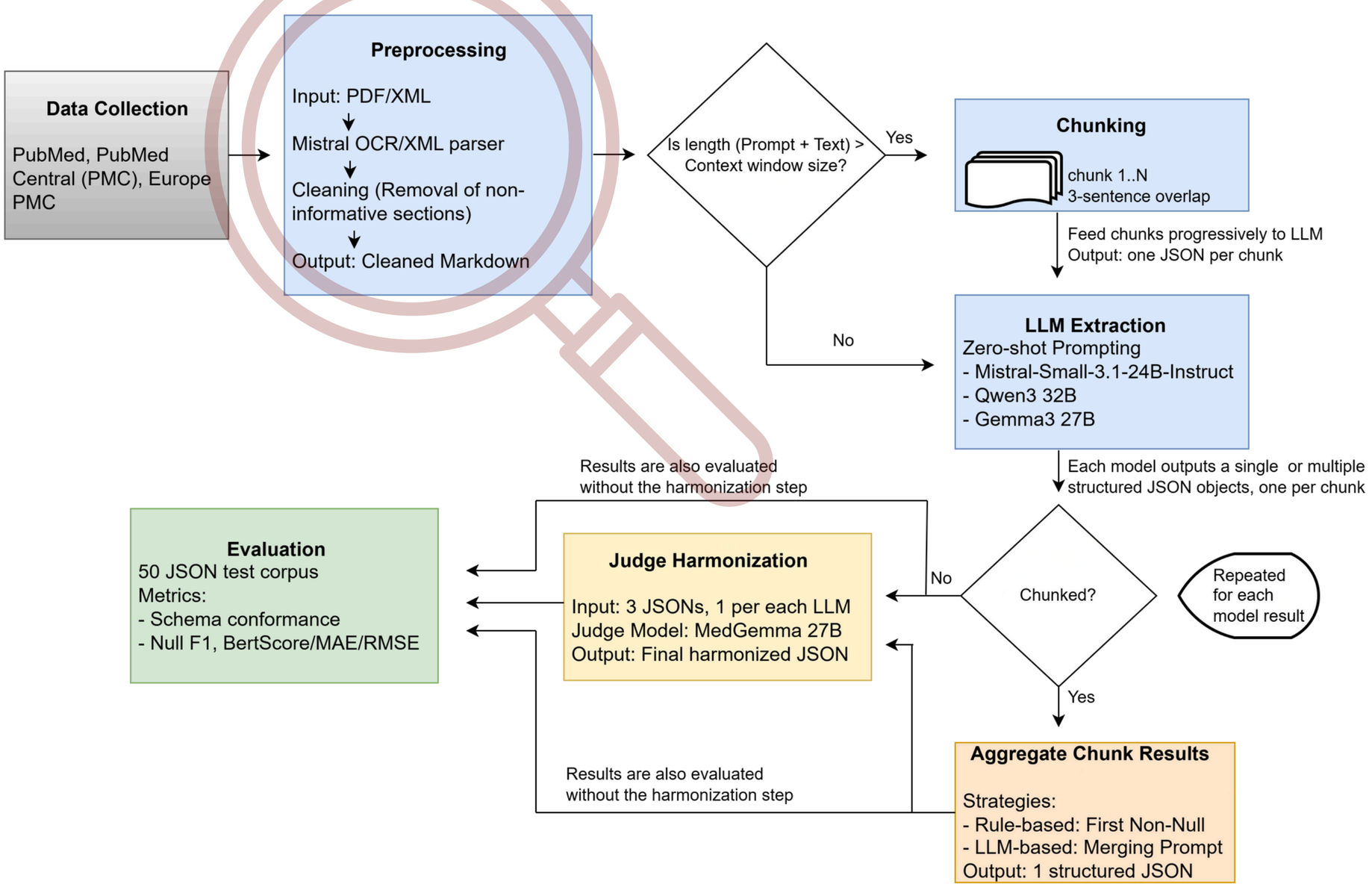


Data Collection: PubMed

- **Article Type:** Full-text research articles
- **Six nanoparticle types:** MSNs, EVs, LNPs, MOFs, PMs, and SNAs.
- **Central nervous system (CNS) conditions:** (e.g., Parkinson, Alzheimer)
- **Language:** English-based articles
- **Date range:** 2010 onwards

Total files retrieved: **528**
Automatically retrieved: **142 (~27%)**
Manually retrieved: **386 (~73%)**

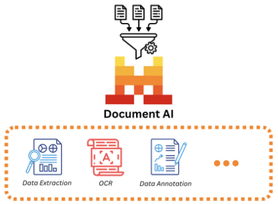
Data Preprocessing

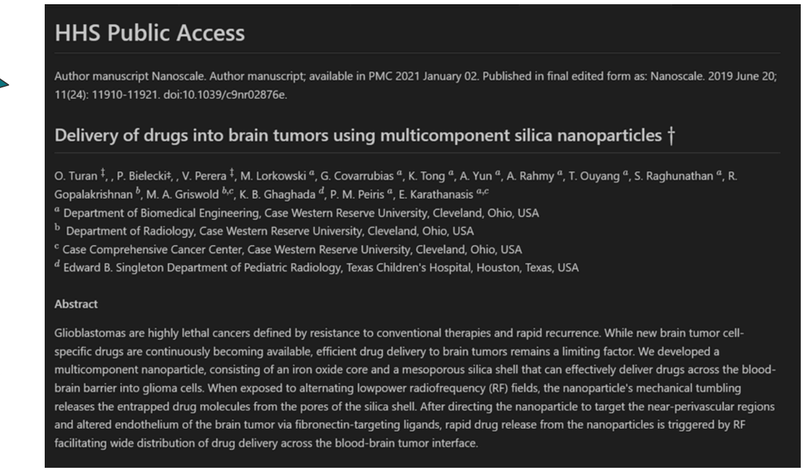


Data Preprocessing

PDF files

XML files





HHS Public Access

Author manuscript Nanoscale. Author manuscript; available in PMC 2021 January 02. Published in final edited form as: Nanoscale. 2019 June 20; 11(24): 11910-11921. doi:10.1039/c9nr02876e.

Delivery of drugs into brain tumors using multicomponent silica nanoparticles †

O. Turan [‡], , P. Bieleckit‡, , V. Perera [‡], M. Lorkowski [a], G. Covarrubias [a], K. Tong [a], A. Yun [a], A. Rahmy [a], T. Ouyang [a], S. Raghunathan [a], R. Gopalakrishnan [b], M. A. Griswold [b,c], K. B. Ghaghada [d], P. M. Peiris [a], E. Karathanasis [a,c]

[a] Department of Biomedical Engineering, Case Western Reserve University, Cleveland, Ohio, USA

[b] Department of Radiology, Case Western Reserve University, Cleveland, Ohio, USA

[c] Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio, USA

[d] Edward B. Singleton Department of Pediatric Radiology, Texas Children's Hospital, Houston, Texas, USA

Abstract

Glioblastomas are highly lethal cancers defined by resistance to conventional therapies and rapid recurrence. While new brain tumor cell-specific drugs are continuously becoming available, efficient drug delivery to brain tumors remains a limiting factor. We developed a multicomponent nanoparticle, consisting of an iron oxide core and a mesoporous silica shell that can effectively deliver drugs across the blood-brain barrier into glioma cells. When exposed to alternating lowpower radiofrequency (RF) fields, the nanoparticle's mechanical tumbling releases the entrapped drug molecules from the pores of the silica shell. After directing the nanoparticle to target the near-perivascular regions and altered endothelium of the brain tumor via fibronectin-targeting ligands, rapid drug release from the nanoparticles is triggered by RF facilitating wide distribution of drug delivery across the blood-brain tumor interface.

2. Experimental section

2.1. Synthesis of Fe@MSN nanoparticles

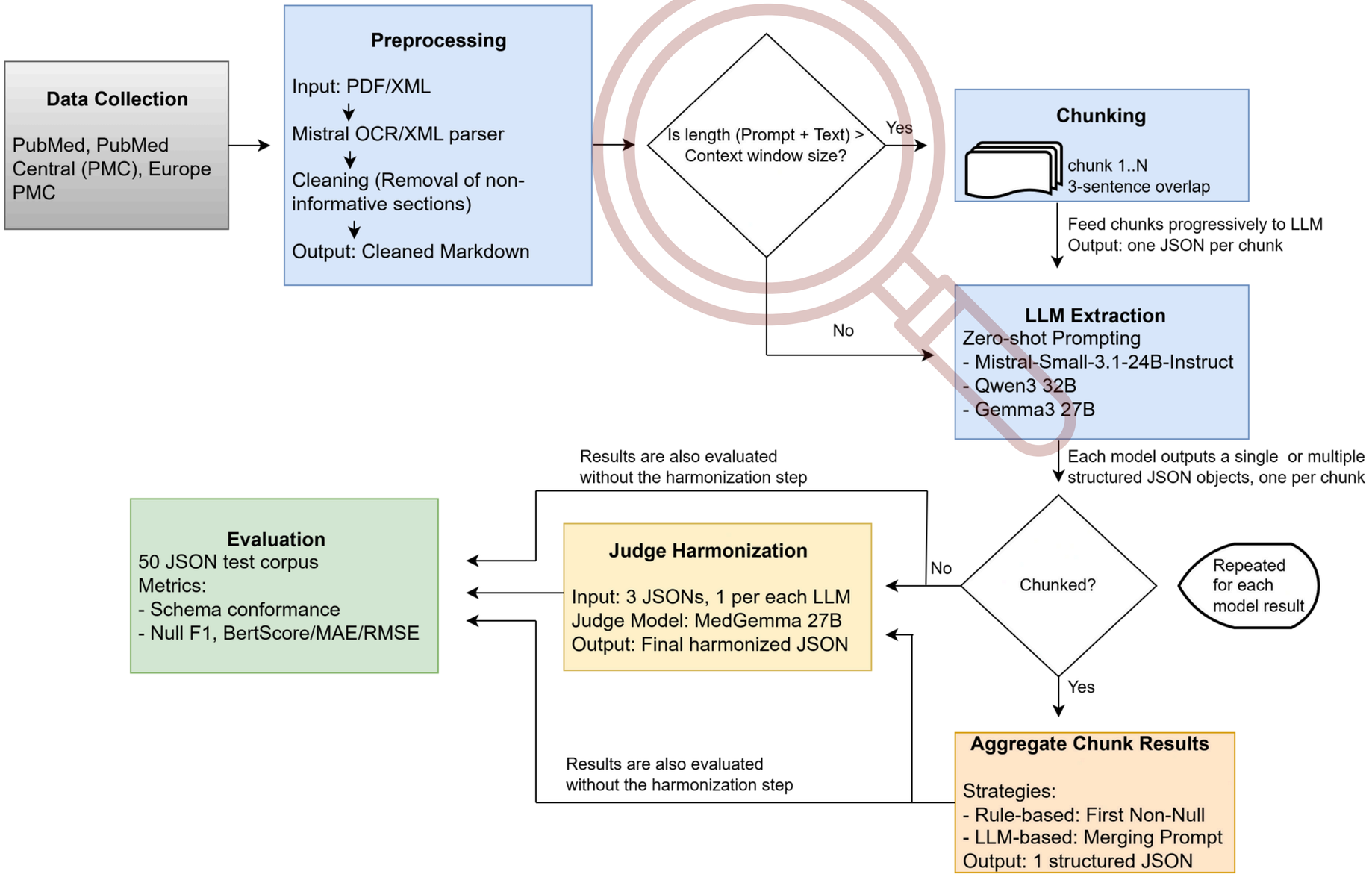
We synthesized the iron oxide cores using a coprecipitation method. Briefly, $FeCl_3 \cdot 6H_2O$ and $FeCl_2 \cdot 4H_2O$ were dissolved in deoxygenated water followed by the addition of the iron precursor solution at $80^{\circ}C$ under argon. After magnetic separation and washing steps, the iron oxide cores were coated with citric acid. The silica shell was synthesized using a basecatalyzed sol-gel process with modifications. Upon dispersing the iron oxide cores in a solution of CTAB, tetraethylorthosilicate (TEOS) was added followed by phosphonate functionalization and silane-PEG- NH_2. By adjusting the pH according to the pK_a of the drug, drugs were conveniently loaded into the nanoparticles via co-incubation for 12 h under mild mixing. 1400 W was loaded into Fe @MSN nanoparticles in PBS at a pH of 8 . DOX was loaded at a pH of 7.4. Any unbound drug was removed from the particles by washing with PBS several times and repeated centrifugation. To evaluate drug loading capacity, the residual drug was measured after the loading procedure. The washing solutions were collected and the residual drug content was measured by UV-Vis absorption spectroscopy at $\lambda = 480\ nm$ for DOX or an HPLC assay for 1400 W . For the concentration of 1400 W , an HPLC assay was run in a C18 5μ reverse-phase column (isocratic; mobile phase: 50% water, 25% methanol and 25% acetonitrile; flow rate of $0.5 ml min^{-1}$; detection at 254 nm). Infrared analyses for DOX and 1400 W were obtained using a Thermo Nexus 870 FTIR spectrometer with an attenuated total reflection (ATR) accessory. Spectra over the 4000-500 cm^{-1} range were obtained by the co-addition of 64 scans with a resolution of 4 cm^{-1}. The content of Fe and Si in the Fe @MSN nanoparticle was determined via ICP-OES (Optima 7000 DV; PerkinElmer). First, the sample was digested with hydrofluoric acid (HF) in a 50 mL polyethylene tubes at room temperature. The sample was held at room temperature for about 6 h . Then, 25% by mass fraction of tetramethyl ammonium hydroxide (TMAH) was

added to neutralize the solution. Finally, an aliquot of $0.2 M HNO_3$ was added for iron digestion. The sample was then analyzed with ICP-OES.

The fibronectin-targeting peptide CREKA [15,16] was conjugated on the surface of the Fe @ MSN particle via its distal end of PEG- NH_2 using standard conjugation chemistry. Briefly, the thiol of the cysteine residue on the peptide was conjugated to the amine of PEGNH_{2} (2)

Output: Clean Markdown file

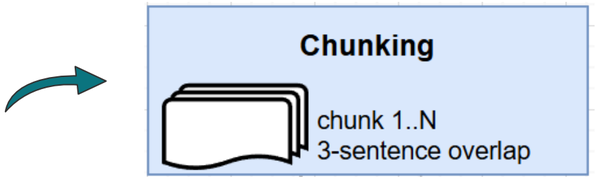
Data Preprocessing

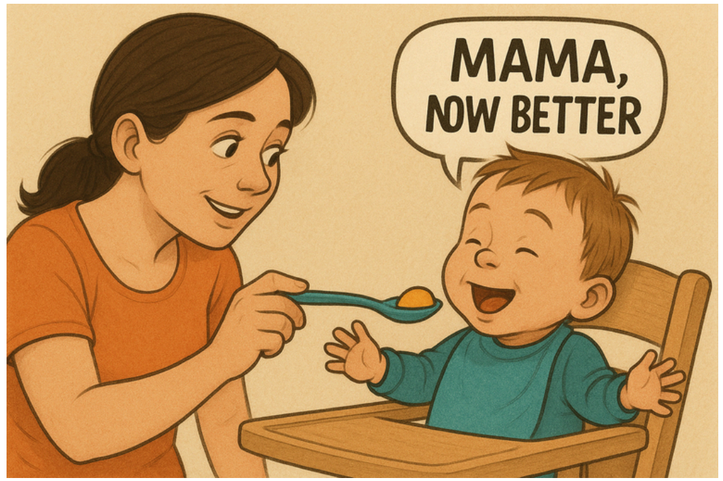


Data Preprocessing

Is len(text + prompt) > context window size?

Trying to feed a very long document to an LLM

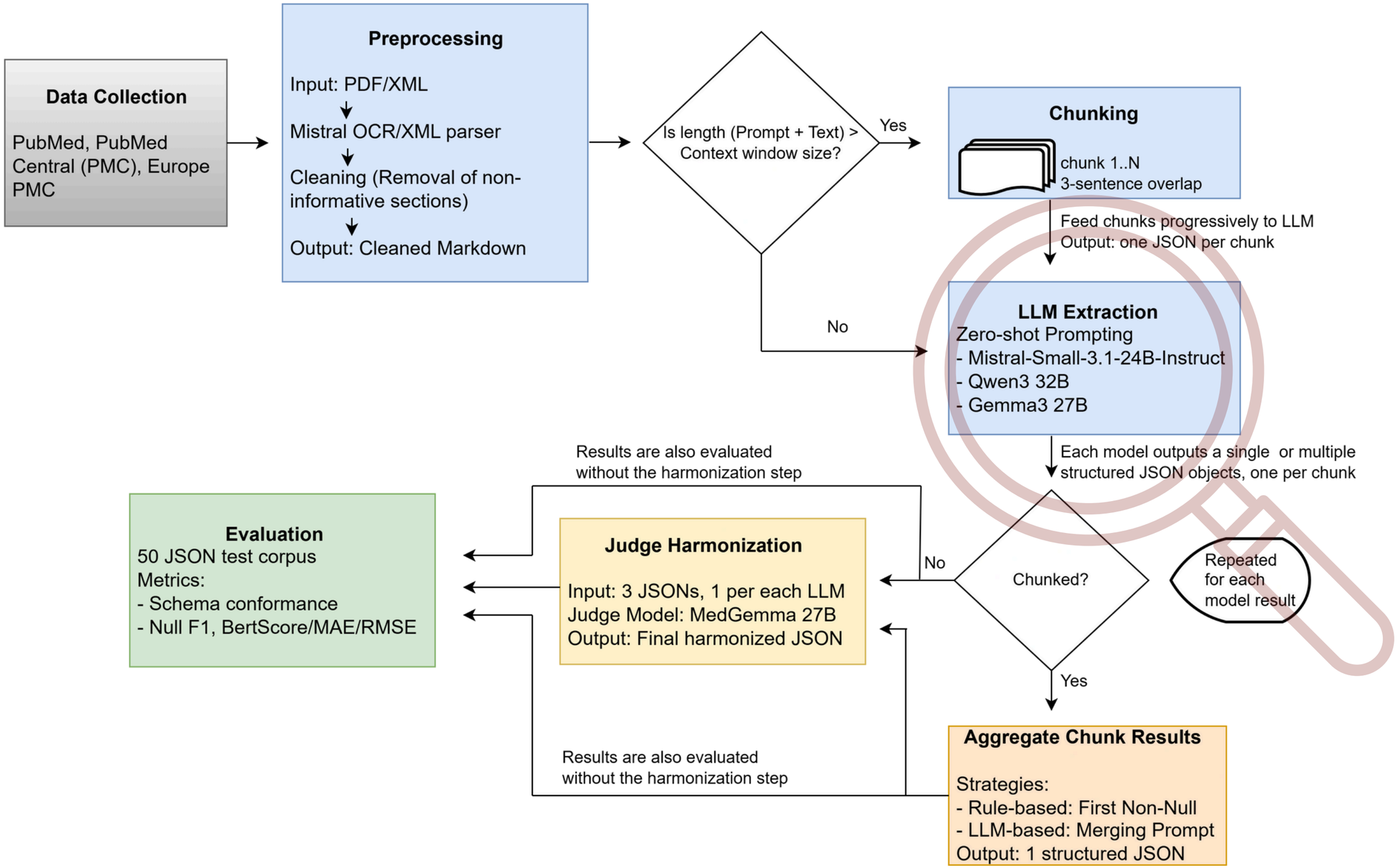




Long docs chunked and fed to LLM

Credit: Images generated Using Dalle-E 3

Data Extraction



Data Extraction

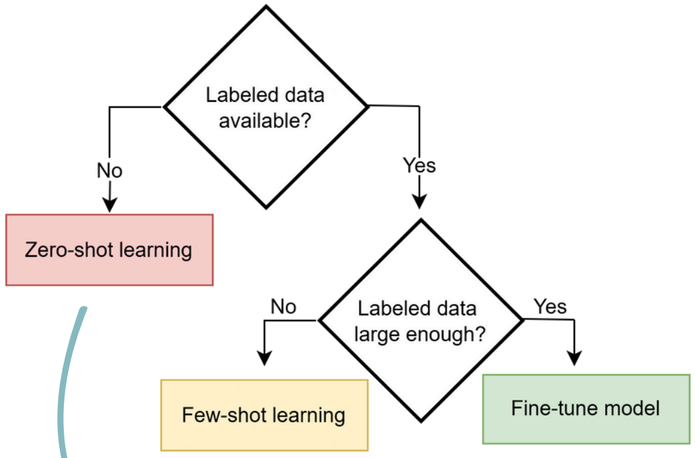
Technical specifications

Model Name	Params	Context Len.	Min. VRAM (FP16/BF16)
Mistral Small 3.1 24B Instr.	24B	128k	~55 GB
Qwen3 32B	32.8B	32k native / 131k (YaRN)	~65.6 GB
Gemma3 27B-it	27B	128k	~54 GB

DeepSeek R1 → ~671 B params (Mixture of Experts, active ~37 B)

32B (~65.6GB) << 671B (~720 GB) !!!

Data Extraction



For each file:

+

Schma + Description

Text + Taxonomy → prompt:
Chunked if length > LLM's context window

Aggregate results for chunked docs:

- *First Non-Null*
- *Merging Prompt*

Hugging Face

Gemma 3 27B Google

Mistral Small 3

Qwen-3 by Alibaba

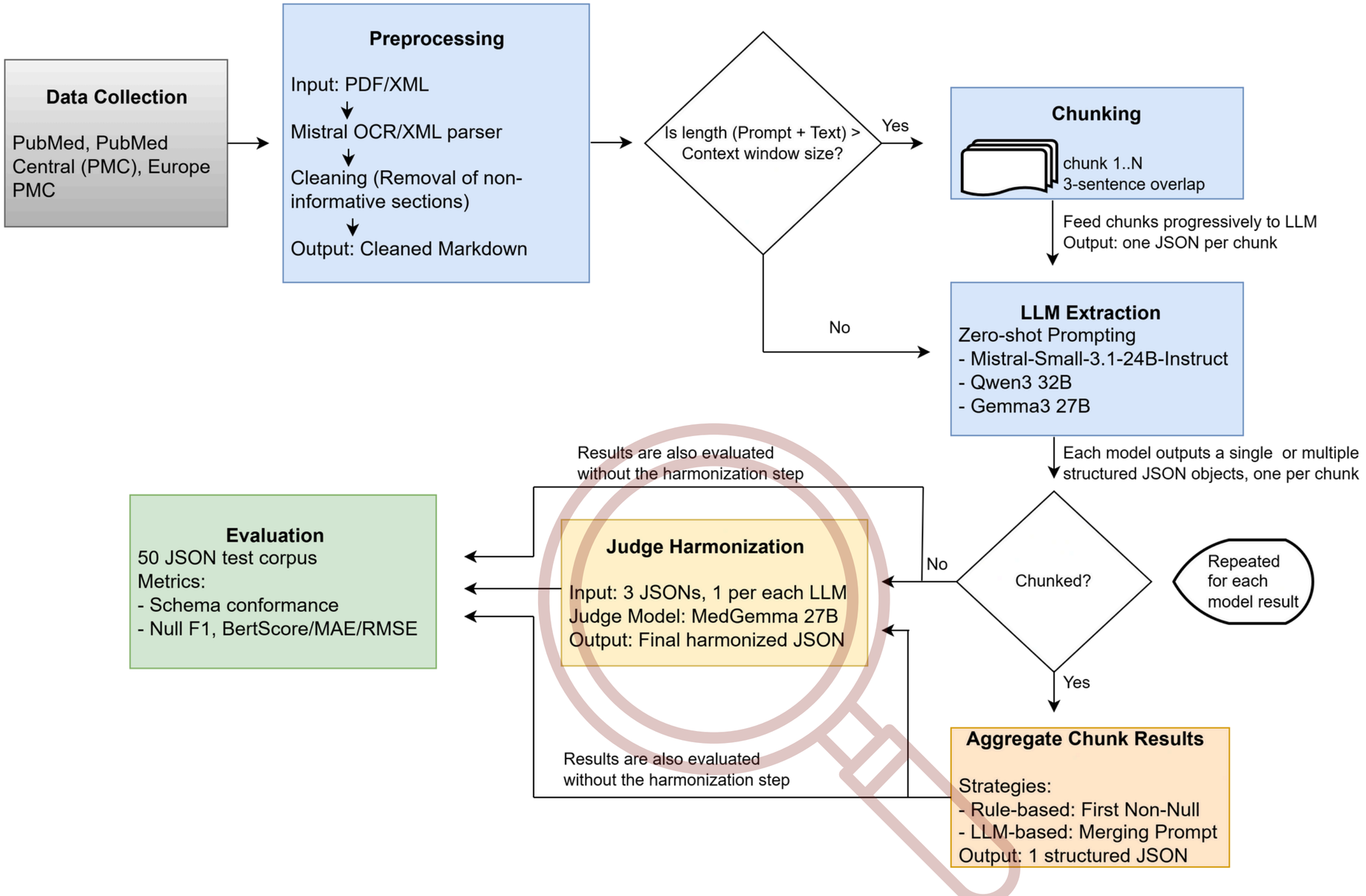
One JSON output per LLM, multiple outputs, one per chunk if doc is chunked

If single output:

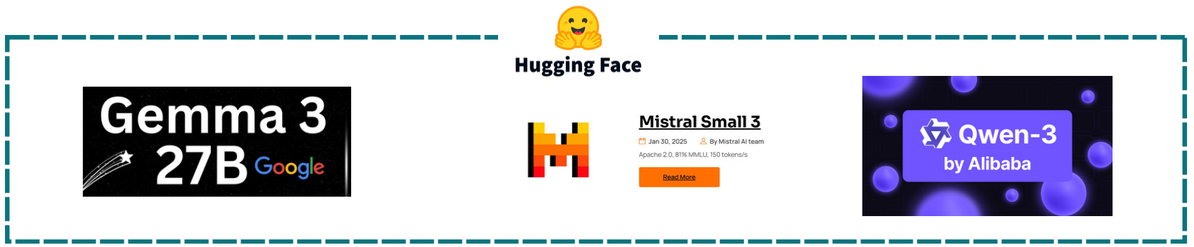
Output: *One JSON per LLM*

```
{
 "Nano_material_name": "PLGA-EV",
 "particle_size": "150 ± 20 nm",
 "zeta_potential": "-32.5 mV",
 "surface_modification": "PEGylated with anti-TfR antibody",
 ...
}
```

Data Harmonization



Data Harmonization



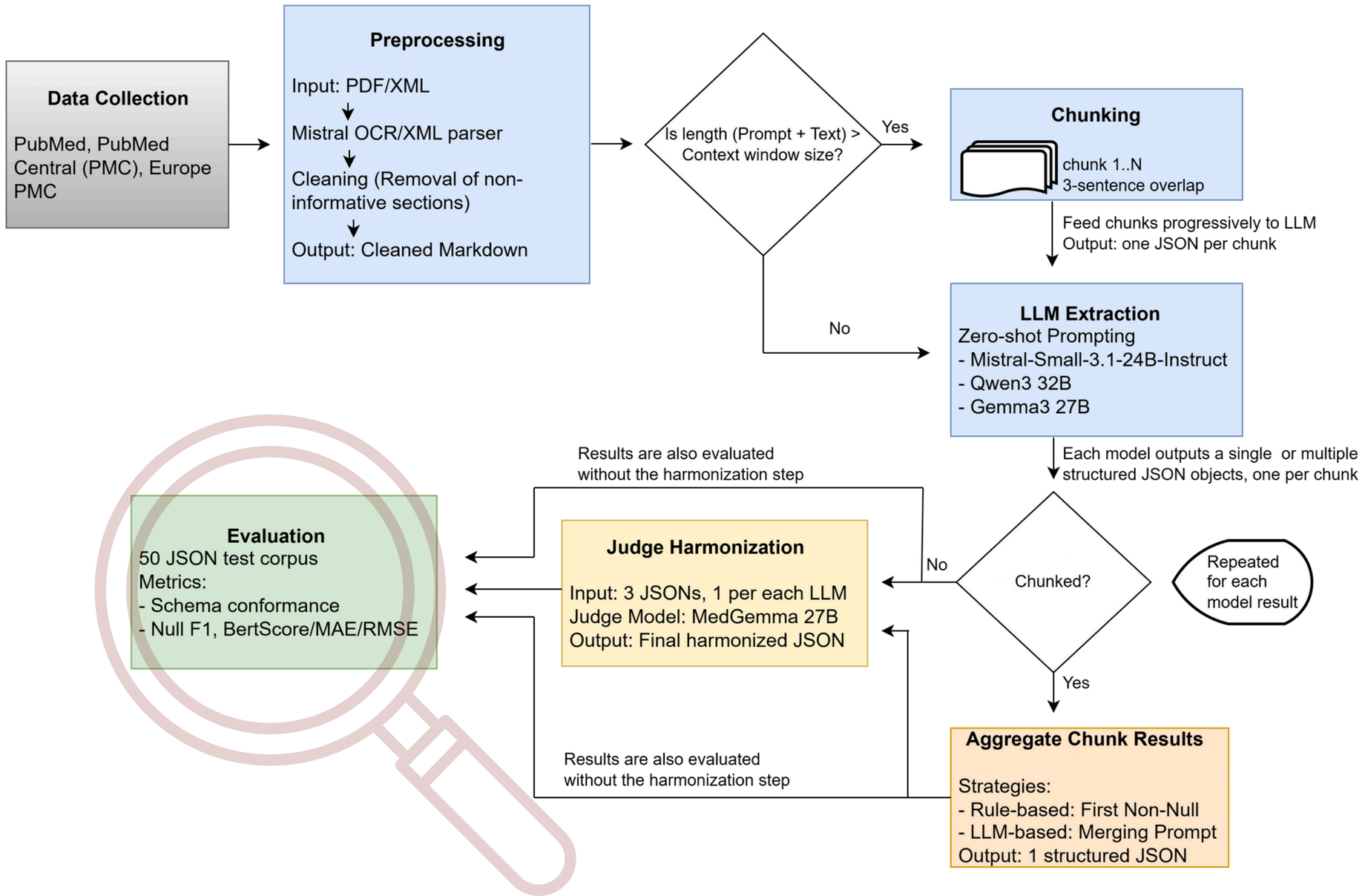
One JSON output per LLM

MedGemma

A domain-adapted variant of Gemma3 27B

Final output: *An aggregated JSON*

Data Harmonization



Performance Evaluation

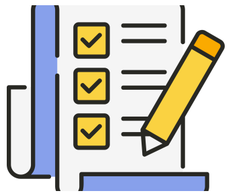
One JSON output per LLM

MedGemma

Final output: *An aggregated JSON*

Evaluate each LLM's output

50 GT corpus

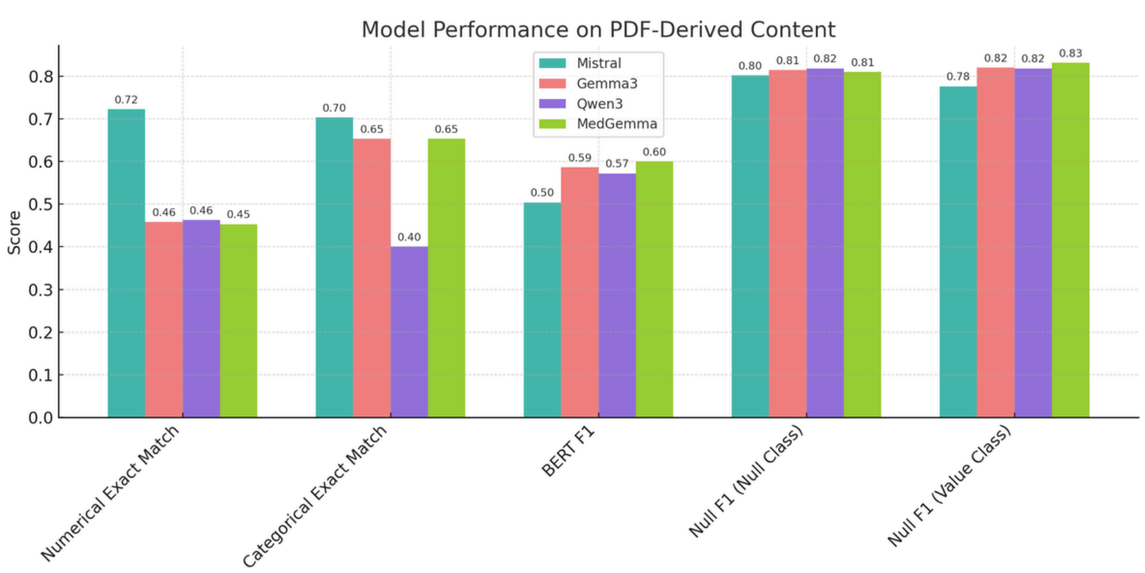


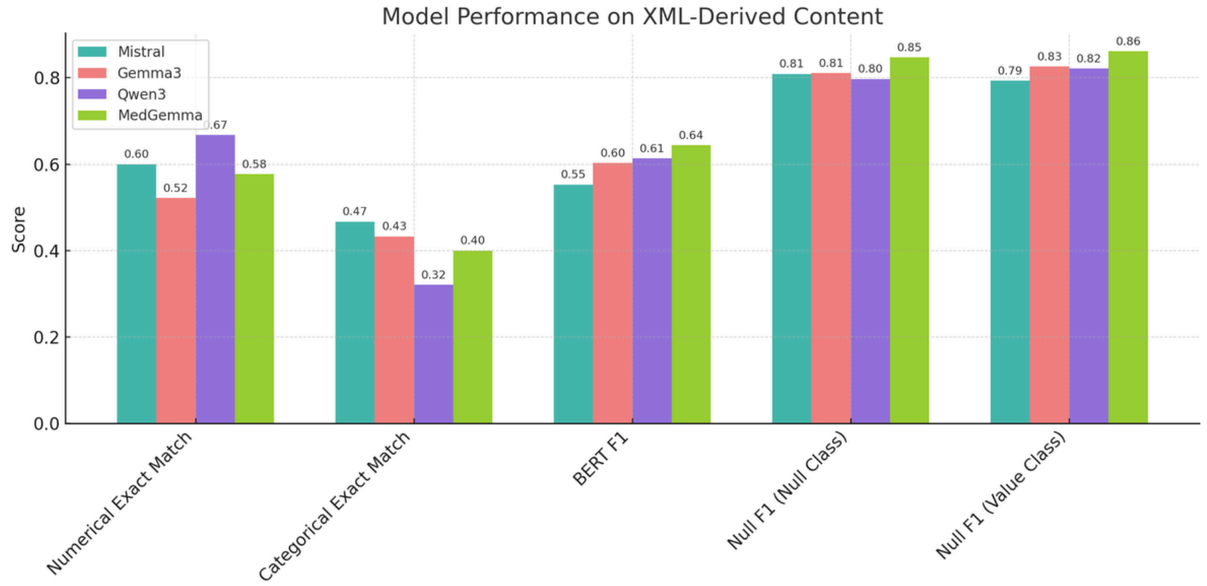
Evaluate judge LLM's output

Metrics:

- **Schema Conformance:** 0-1
- **Null F1:** 0-1
- **BertScore:** 0-1
- **Categorical Exact Match:** 0-1
- **Numerical Exact Match:** 0-1
- **MAE/RMSE:** 0-∞

Results





Attention: Bigger does not mean better!

Error Analysis

Field Confusion: "22 ± 1 nm" extracted as a monodispersity value, despite no mention → confused size with homogeneity.

Prefix Omission: Skipped critical prefix like "TMZ-loaded", returning incomplete material names

Loss of Structure / Detail: Reduced detailed material composition to a flat list of chemicals, missing core-shell info, MIL type, and drug-loading

OCR Artifacts: Misread "1.6 × 10^2 nm" as "16102 nm"—artifact from faulty scientific notation recognition.

Study Limitations

- Lack of inter-annotator agreement (IAA)
- Preprocessing Challenges → OCR artifacts
- Limited Capabilities of Zero-Shot Approach
- Small Evaluation Corpus Size

Conclusion & Future Work

Aims Achieved:

- Constructed a **detailed taxonomy** of BBB-relevant nanoparticle parameters
- Built a modular extraction pipeline centered on **light open-weight LLMs**
- Carried out a **systematic performance evaluation**.

This Work:

- Holds promise for accelerating **knowledge synthesis in nanomedicine**
- Would aid researchers in **identifying trends, formulating hypotheses, and designing new experiments** for BBB-targeting nanoparticles

Future work:

- Establish **inter-annotator** agreement (IAA)
- Preprocessing is key → **Resolve OCR and parsing challenges**
- Explore **parameter-efficient fine-tuning** (e.g., LoRA, QLoRA)

Note: The extracted JSON outputs will populate a centralized digital nanoparticle library planned for development within the NAP4DIVE project

Meet my Supervisors

Dr. Cristina Suemay Manresa Yee

Affiliation: University of the Balearic Islands

Prof. Dr. Sebastien Lafond

Affiliation: Åbo Akademi University

Dr. Hergys Rexha

Affiliation: Åbo Akademi University

EDISS
MASTER'S PROGRAMME

MUCHAS GRACIAS!

Do you have any questions?

EDISS
MASTER'S PROGRAMME