Funded by the European Union

HORIZON-WIDERA-2023-ACCESS-02

ChatMED

Grant ID: 101159214

Bridging Research Institutions to Catalyze Generative AI Adoption by the Health Sector in the Widening Countries

"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER SCIENCE AND ENGINEERING
Jožef Stefan Institute
Univerzitet u Nišu
MEDICINSKI FAKULTET
TU Graz

Deliverable D7.2

Datasets collected for the research

Funded by
the European Union

Deliverable D7.2 - Datasets collected for the research	
Date:	January 31, 2026
Lead Beneficiary:	FMN
Type:	DATA
Author(s):	Stevo Lukic, Ivana Vichentijevikj, Monika Simjanoska Misheva, Kostadin Mishev
Dissemination level:	PU
Version:	1.0

Abstract: Deliverable D7.2 presents the Neurology Reasoning Dataset (NeRD), a structured dataset designed to support reasoning-oriented training and evaluation of the NeuroOrchestrator within the ChatMED multi-agent architecture. While numerous datasets exist for training modality-specific models on isolated prediction tasks, such resources are insufficient for modeling end-to-end clinical reasoning and decision navigation. NeRD addresses this gap by providing a dataset explicitly constructed to reflect how clinicians think and act in practice, progressing from initial patient presentation through iterative hypothesis refinement, diagnostic decision-making, and final diagnosis. The dataset comprises 979 neurological cases sourced from peer-reviewed literature and established educational resources, stratified across three levels of clinical complexity. Each case is represented as a complete clinical reasoning scenario, capturing staged information availability and explicitly separating evidence, reasoning steps, and diagnostic outcomes. Using the methodology defined in Deliverable D7.1, unstructured case narratives were transformed into standardized, reasoning-oriented JSON representations, preserving intermediate reasoning traces generated by the NeuroOrchestrator prompt pipeline. The dataset is published in accordance with FAIR principles and provides a reusable foundation for research on explainable, workflow-aware, and multi-agent clinical reasoning systems.

HISTORY OF CHANGES

Date	Issue	Author	Summary of main changes
13.11.2025	1.0	Monika S. Misheva	Establishment of the template
30.1.2026	2.0	Ivana Vichentijevikj	Finalizing the content, merging the datasets collected by FMN and FCSE.
31.1.2026	3.0	Monika S. Misheva	Improving the content and the organization.

Funded by the European Union

31.1.2026	4.0	Stevo Lukic	Final edit and approval.

Table of Contents

Funded by
the European Union

List of Figures

List of Tables

List of Acronyms

AI (Artificial Intelligence)
EEG (Electroencephalography)
EMG (Electromyography)
EHR (Electronic Health Records)
EU (European Union)
FAIR (Findable, Accessible, Interoperable, Reusable)
FCSE (Faculty of Computer Science and Engineering)
FMN (Faculty of Medicine Nis)
GDPR (General Data Protection Regulation)
HEF (Human Evaluation Framework)
ID (Identification)
JSI (Jožef Stefan Institute)
LLM (Large Language Model)
NCBI (National Center for Biotechnology Information)
NeRD (Neurology Reasoning Dataset)
RAG (Retrieval-Augmented Generation)
STSE (Short Term Staff Exchange)
WP (Work Package)
WIDERA (Widening Participation and Spreading Excellence)

Funded by
the European Union

Introduction

The ChatMED platform is designed as a modular, multimodal multi-agent system in which specialized agents operate over heterogeneous data modalities, including biochemical data, EEG/EMG signals, neuroimaging, bioinformatics outputs, pharmaceutical knowledge, and electronic health records (EHRs). While a wide range of public and restricted datasets already exist for training and validating models within these individual modalities, such datasets alone are not sufficient to support coherent, end-to-end **clinical reasoning** across the full diagnostic and decision-making pathway. In this context, **clinical reasoning** refers to the way a clinician thinks and acts in practice, encompassing the cognitive processes by which physicians interpret symptoms, formulate and revise hypotheses, select and prioritize diagnostic tests, and ultimately arrive at informed decisions, thereby explicitly aiming to mimic expert clinical behavior rather than isolated predictive tasks.

For example, while the analysis and interpretation of the results of a specific diagnostic test are handled by an isolated, modality-specific multi-agent system tailored to that task, **the *decision to request, prioritize, or sequence that diagnostic test* is governed by the NeuroOrchestrator LLM** which functions as the **central coordinating** component, or "**brain**", of the overall ChatMED system, as illustrated in Figure 1. The dataset described in this deliverable D7.2 is explicitly designed to support this decision-making function.

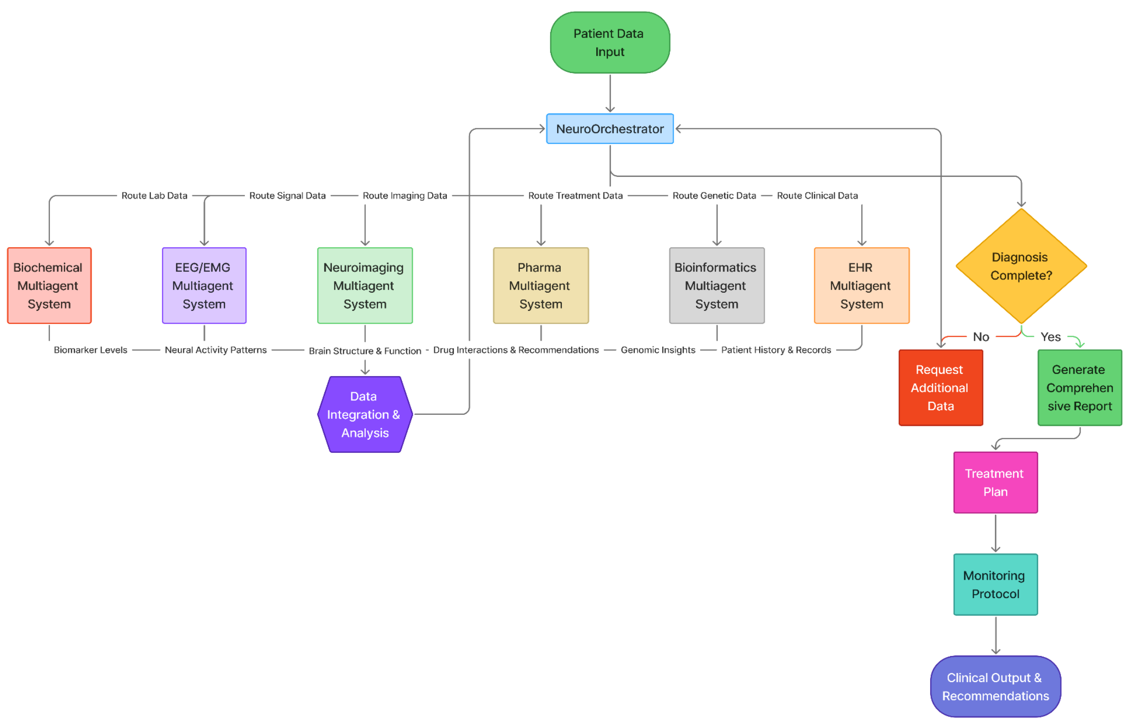


Figure 1. *Simplified illustration of the NeuroOrchestrator as the central coordinator within the ChatMED multi-agent system.*

Funded by
the European Union

The NeuroOrchestrator is responsible for navigating the clinical reasoning process from initial patient presentation, through hypothesis generation and iterative information gathering, to diagnostic consolidation, treatment planning, and downstream monitoring (with detailed architecture present in the Deliverable D7.1). Rather than performing modality-specific prediction tasks itself, the NeuroOrchestrator governs *how* and *when* individual multi-agent subsystems are invoked, how their outputs are interpreted, and how intermediate conclusions guide subsequent steps in the reasoning workflow.

The NeuroOrchestrator will be implemented as a fine-tuned large language model selected based on its superior performance during the benchmarking activities reported in WP8. To fulfill its coordinating role, however, this model cannot be fine-tuned solely on datasets targeting isolated diseases, conditions, or modalities. Instead, it requires a **reasoning-oriented dataset** that explicitly captures structured, step-by-step clinical thinking: from symptom interpretation, through diagnostic branching and test selection, to the synthesis of evidence and generation of actionable recommendations.

This deliverable, **D7.2**, therefore focuses on the specification and description of such a reasoning-based dataset. The dataset is constructed in direct alignment with the methodology defined in **D7.1**, translating the conceptual clinical reasoning pipeline into a form suitable for supervised fine-tuning of the NeuroOrchestrator. Its purpose is not to improve predictive accuracy for a specific disease or modality, but to model the *navigation logic* that ensures appropriate progression through the clinical decision process.

The dataset captures rich, structured representations of neurological case narratives that support downstream reasoning tasks such as suggested investigations, therapeutic and management considerations, treatment options, potential drug interactions, and risk assessment. By transforming unstructured clinical descriptions into machine-readable and logically organized data, the dataset enables systematic analysis, reproducibility, and the evaluation of reasoning-oriented computational models.

The dataset is provided in a standardized JSON-structured content, ensuring interoperability, reusability, and compatibility with external biomedical tools and ontological resources.

The Neurology Reasoning Dataset (NeRD)

The Neurology Reasoning Dataset (NeRD) comprises **979 neurological cases**, provided in a standardized JSON format. Each case represents a complete clinical reasoning scenario, encompassing the initial patient presentation, iterative reasoning and decision-making stages, and a finalized diagnostic conclusion.

To reflect varying levels of clinical complexity and provenance, the dataset is stratified into **three difficulty tiers**:

- **Advanced (628 cases):** Derived from PubMed-indexed neurological case reports, representing specialist-level diagnostic complexity and real-world clinical reasoning.
- **Intermediate (107 cases):** Capturing moderate diagnostic complexity, suitable for structured reasoning evaluation beyond basic educational examples.
- **Basic (244 cases):** Sourced from educational and student-oriented materials.

Collectively, the dataset spans a broad spectrum of neurological presentations, diagnostic pathways, and investigative depth. All cases adhere to a consistent structural schema, enabling reproducible analysis and systematic comparison of clinical reasoning processes across different complexity levels.

The Methodology for the Dataset Construction

Data Sources

Neurological cases included in the dataset were collected from a combination of peer-reviewed literature and established educational case collections. The dataset includes both publicly redistributable sources and restricted licensed materials used exclusively for internal research. Specifically, the data sources comprise:

- PubMed-indexed neurological case reports (782 cases)
- Physiopedia Neurological-Case Studies[1] (43 cases)
- Other sources (154 cases)

These sources were selected due to their clinical reliability, educational value, and detailed narrative descriptions, which are essential for modeling realistic neurological case scenarios. Together, they provide a broad coverage of neurological conditions, levels of complexity, and presentation styles.

All cases were reviewed to ensure suitability for research use and to exclude directly identifiable personal information prior to further processing.

Case Retrieval and Initial Processing

PubMed-indexed case reports were retrieved using an automated extraction pipeline based on the NCBI E-utilities API. A structured query was used to identify human neurological case reports published in English and indexed as case reports. The query combined publication type, disease domain, and temporal constraints to ensure relevance and consistency across retrieved cases.

The query used for PubMed retrieval was:

[1] https://www.physio-pedia.com/index.php?title=Category:Neurological_-_Case_Studies&oldid=222887

"Case Reports"[Publication Type] OR "case report"[Title]) AND ("Neurologic Diseases"[MeSH Terms] OR neurol*[Title/Abstract]) AND humans[MeSH Terms] AND english[Language]

This query was selected to ensure that retrieved records correspond to clinically relevant neurological case reports involving human subjects, while avoiding restrictions that could bias the dataset toward specific publication periods. Retrieved records included article identifiers, bibliographic metadata, abstracts, and links to full-text content when available.

Preprocessing and Restructuring

PubMed-indexed case reports were frequently unstructured and often contained the final diagnosis within the initial narrative. To convert these texts into realistic, hypothesis-driven case scenarios suitable for downstream reasoning, we applied a preprocessing step by using the model **qwen/qwen3-8b**. The model was instructed to restructure each case into a fixed set of fields that reflect the clinical flow from first presentation, through initial reasoning and investigation planning, to iterative reasoning after new findings, while preserving clinical content verbatim as given in the following Table 1.

Table 1. The prompt used for Qwen Aided Restructuring.

Task: Extract structured information from neurological case reports into the following fields.

1. **Introduction**
   Present the raw clinical case as it would appear at first patient contact.
   - Include: demographics (age, sex), presenting symptoms and timeline, past medical history, examination findings (neurological + systemic), basic laboratory results if already available.
   - Exclude: interpretation, summarization, and any diagnosis or diagnostic assumptions (except when an initial diagnosis is explicitly mentioned and later ruled out by testing).
2. **Further Investigation I**
   Capture the first-pass clinical reasoning based solely on history and examination (hypothesis-generation phase).
   - Include: key clinical patterns (e.g., UMN vs LMN; sensory vs motor), localization reasoning (central vs peripheral; cortical vs spinal), initial differential diagnoses, ruling in/out based on examination, justification for next diagnostic steps (e.g., MRI, NCS, EEG).
   - Exclude: final diagnosis and retrospective certainty.
3. **Further Investigation II**
   Capture iterative diagnostic reasoning after new data emerges (reasoning update, not repetition).
   - Include: interpretation of new findings (imaging, EEG, labs, progression), re-evaluation of earlier hypotheses, explanation of why some diagnoses are excluded, refinement toward the leading diagnosis.

  - Exclude: re-describing the entire case and purely factual repetition without interpretation.

4. **true_diagnosis**
   Extract the definitive diagnosis represented in the source text.
   - Include: diagnoses explicitly stated or unambiguously implied by definitive findings or confirmation (e.g., confirmed lesion, surgical/pathological confirmation).
   - Exclude: speculation, differential diagnoses, and early hypotheses.

Constraints:

- Do not summarize, paraphrase, compress, generalize, or reinterpret the content.
- Preserve clinical details verbatim.
- Populate a field only if the corresponding information is explicitly present; otherwise leave it empty.
- Output must follow the predefined field structure.

Dataset Organization and Field Definition

After retrieval and restructuring, all cases were organized into a unified dataset with a fixed set of fields designed to support structured clinical reasoning and hypothesis refinement. Each row in the dataset corresponds to a single neurological case, and each column represents a distinct stage of clinical information or reasoning.

The dataset includes the following fields:

- **ID:** A unique identifier assigned to each case.
- **Introduction:** Represents the initial clinical presentation of the patient as it would be encountered at first contact. This field contains the structured case presentation extracted during preprocessing, including demographics, presenting symptoms and timeline, relevant medical history, examination findings, and basic laboratory results when available. No interpretation or diagnostic conclusions are included.
- **Further Investigations I:** Captures the first-pass clinical reasoning based on the initial presentation. This field reflects early hypothesis generation and includes identification of key clinical patterns, localization reasoning, initial differential diagnoses, and justification for ordering further diagnostic investigations.
- **Further Investigations II:** Captures iterative diagnostic reasoning following the availability of additional clinical data. This field reflects re-evaluation of earlier hypotheses, interpretation of new findings, exclusion of alternative diagnoses, and refinement toward the leading diagnosis.
- **True_diagnosis:** Contains the definitive diagnosis associated with the case, as explicitly stated or unambiguously implied in the source material. This field serves as a reference outcome for evaluation purposes and is not included in the reasoning inputs.
- **Difficulty:** Indicates the intended complexity level of the case. Cases originating from educational student-oriented case collections are labeled as basic, reflecting their

Funded by the European Union

instructional nature and suitability for foundational clinical reasoning. Cases drawn from curated clinical resources or mixed educational–clinical sources are labeled as intermediate, reflecting moderate complexity that requires integrated clinical reasoning but not full specialist expertise. Cases originating from PubMed-indexed clinical case reports are labeled as advanced, reflecting their higher complexity and specialist-level reasoning requirements.

This structured organization ensures consistency across heterogeneous sources and enables controlled analysis of clinical reasoning performance across different levels of case complexity.

Prompt-Driven Structured Output Generation

Following the restructuring of heterogeneous data sources into a consistent and organized pool of records, the methodology defined in Deliverable D7.1 was applied to perform prompt-driven, JSON-based structuring. In this process, the full set of agent-specific prompts was executed to generate a single, unified JSON representation for each clinical scenario, explicitly reflecting a reasoning-oriented organization of the underlying information.

Within this pipeline, the **Introduction** field is treated as the primary input to the system, corresponding to the information available at initial patient presentation. The **Further Investigations I** and **Further Investigations II** fields, when present, are subsequently incorporated as additional inputs to refine and update diagnostic hypotheses. By sequentially incorporating these fields, the orchestrator mirrors the iterative nature of clinical decision-making, in which hypotheses are revised as new information becomes available.

Specifically, the dataset includes JSON-mapped outputs associated with the following reasoning stages and components: **A0**, **A1**, **A2-H1**, **A2-H2**, **A2-H3**, **A2-H4**, **A2-H5**, **A3**, **A3.1**, **A4**, **A5**, **A5.5**, **A6**, as well as the **Hypothesis Refiner** and **DiagnosticTestExtractor** components. By preserving these intermediate reasoning outputs in a structured JSON format, the dataset enables transparent inspection, auditing, and comparison of reasoning trajectories across cases and difficulty levels. This design supports detailed evaluation of how diagnostic hypotheses evolve over time and how individual reasoning components contribute to final decision-support outputs, without exposing or altering the underlying orchestration logic described in Deliverable D7.1.

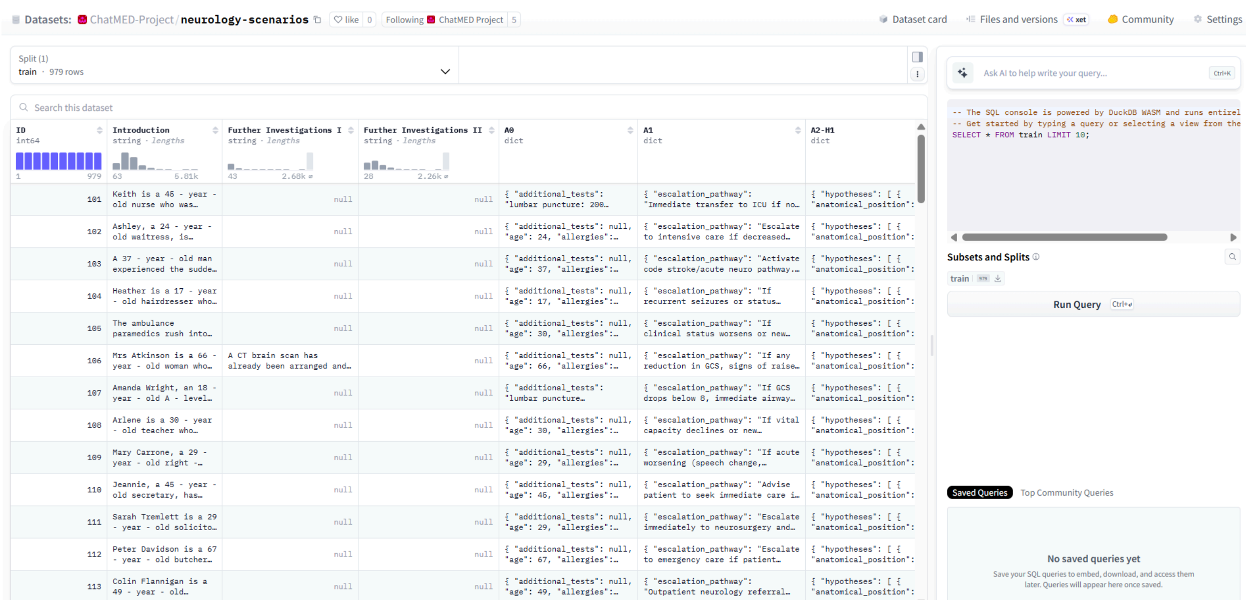
This design allows the D7.1 methodology to simulate a realistic diagnostic flow, where reasoning is conditioned on the timing and availability of clinical data rather than on complete case knowledge from the outset.

FAIR Data Management and Accessibility

The NeRD dataset is managed in accordance with the FAIR principles to ensure findability, accessibility, interoperability, and reusability, while respecting the licensing constraints and attribution requirements of the original data sources.

Findability

A reduced, publicly shareable subset of the dataset is deposited on the Hugging Face platform[2], accompanied by a dedicated repository and a comprehensive README file. The repository includes descriptive metadata, and clear documentation of the dataset structure, facilitating discovery and reference by both humans and automated systems. The complete dataset, including components derived from restricted or commercially licensed sources, is used exclusively for internal research and model development and is not intended for public redistribution.



Accessibility

The dataset is distributed in a standardized JSON-structured content, ensuring broad accessibility using common data-processing tools. Access conditions will be explicitly stated in the accompanying README file and aligned with the most restrictive license applicable across the original source materials.

Interoperability

The use of a JSON-based schema in Table 2 enables seamless integration with external biomedical tools, ontologies, and downstream AI pipelines. Field names, data types, and structural conventions are designed to support compatibility with existing clinical reasoning frameworks and multi-agent systems.

---

[2] https://huggingface.co/datasets/ChatMED-Project/neurology-scenarios/

The Dataset Schema

Table 2. The NeRD Schema.


```
{
  "title": "NeRD Minimal Record Schema",
  "type": "object",
  "required": ["ID", "Introduction", "difficulty", "final_diagnosis"],
  "additionalProperties": true,
  "properties": {
    "ID": {
      "description": "Unique case identifier.",
      "type": ["integer", "string"]
    },

    "Introduction": {
      "description": "Initial unstructured/structured case presentation at first contact.",
      "type": ["string", "null"],
      "minLength": 1
    },

    "Thoughts": {
      "description": "Optional intermediate reasoning text (may be empty).",
      "type": ["string", "null"]
    },
    "Thoughts I": {
      "description": "Optional intermediate reasoning text (may be empty).",
      "type": ["string", "null"]
    },

    "a0": {
      "description": "Stage output A0, stored as a JSON-serialized string.",
      "type": ["string", "null"],
      "contentMediaType": "application/json"
    },
    "a2.5": {
      "description": "Stage output A2.5, stored as a JSON-serialized string.",
      "type": ["string", "null"],
      "contentMediaType": "application/json"
    },
    "a21": {
      "description": "Stage output A2-H1 (dataset column a21), stored as a JSON-serialized
string.",
      "type": ["string", "null"],
      "contentMediaType": "application/json"
    },
    "a22": {
      "description": "Stage output A2-H2 (dataset column a22), stored as a JSON-serialized
string.",
      "type": ["string", "null"],
```

Funded by the European Union


```
    "contentMediaType": "application/json"
  },
  "a23": {
    "description": "Stage output A2-H3 (dataset column a23), stored as a JSON-serialized
string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a24": {
    "description": "Stage output A2-H4 (dataset column a24), stored as a JSON-serialized
string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a26": {
    "description": "Stage output A2-H5 (dataset column a26), stored as a JSON-serialized
string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },

  "a3": {
    "description": "Stage output A3, stored as a JSON-serialized string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a3.1": {
    "description": "Stage output A3.1, stored as a JSON-serialized string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a4": {
    "description": "Stage output A4, stored as a JSON-serialized string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a5": {
    "description": "Stage output A5, stored as a JSON-serialized string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a55": {
    "description": "Stage output A5.5 (dataset column a55), stored as a JSON-serialized
string.",
    "type": ["string", "null"],
    "contentMediaType": "application/json"
  },
  "a6": {
    "description": "Stage output A6, stored as a JSON-serialized string.",
    "type": ["string", "null"],
```

```
      "contentMediaType": "application/json"
    },

    "diagnostic_tests_results": {
      "description": "Aggregated diagnostic tests / results, stored as a JSON- or dict-serialized string.",
      "type": ["string", "null"]
    },

    "difficulty": {
      "description": "Case complexity tier.",
      "type": "string",
      "enum": ["basic", "intermediate", "advanced"]
    },

    "final_diagnosis": {
      "description": "Definitive diagnosis label for the case.",
      "type": ["string", "null"]
    }
  }
}
```


Reusability

Reusability is ensured through detailed documentation of the dataset schema, the reasoning methodology applied (as defined in Deliverable D7.1), and explicit licensing information. The dataset is released under the Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 (CC BY-NC-ND 4.0) license, selected as the most restrictive license required across all source materials, thereby ensuring compliance with original usage conditions while enabling non-commercial research reuse. The dataset license is selected based on the most restrictive licensing terms among all source datasets, and full attribution is provided by listing all original data sources and contributors in the README file. This approach ensures ethical reuse, transparency, and compliance with source-specific usage conditions. The CC BY-NC-ND 4.0 license applies to the dataset as a compiled artifact and does not override the original licenses of individual source materials. Original sources remain subject to their respective licensing terms and are fully attributed.

Conclusion

Deliverable D7.2 documents the Neurology Reasoning Dataset (NeRD), a reasoning training-ready dataset constructed through the systematic application of the clinical reasoning methodology defined in Deliverable D7.1. While D7.1 specifies how structured clinical reasoning is modeled and orchestrated within the ChatMED architecture, D7.2 provides the concrete data

Funded by
the European Union

artifact required to operationalize, evaluate, and refine that methodology using real-world neurological cases.

The dataset comprises 979 neurological cases organized to reflect realistic clinical workflows, progressing from first patient contact through staged investigations and iterative reasoning to definitive diagnosis. By explicitly separating raw clinical evidence, intermediate reasoning steps, and diagnostic outcomes, and by preserving structured reasoning traces generated during prompt-driven orchestration, NeRD enables transparent inspection and reproducible analysis of diagnostic decision-making processes.

Importantly, NeRD is designed to support reasoning navigation rather than isolated prediction. Modality-specific analyses and test interpretations remain the responsibility of specialized agents, while the dataset equips the NeuroOrchestrator to learn when and why diagnostic actions should be taken. This distinction ensures that the dataset directly supports the role of the NeuroOrchestrator as the coordinating “brain” of the ChatMED multi-agent system.

Together, Deliverables D7.1 and D7.2 establish a coherent methodological and data foundation for subsequent project activities. D7.1 defines the structure and logic of clinical reasoning orchestration, while D7.2 provides a FAIR-compliant, reasoning-oriented dataset suitable for training, fine-tuning, and evaluating models that aim to replicate expert clinical thinking under realistic diagnostic conditions.