Aurum at CRF Filling 2026: Modular DSPy Extractors with Qwen3-Max for Multilingual CRF Filling

Vinay Babu Ulli[1], **Jyoti Kumari**[2], **Anindita Mondal**[3]
[1]Oogwai Analytics, Bangalore, India
[2]Department of Linguistics, Banaras Hindu University, India
[3]Language Technologies Research Center, IIIT Hydearabad, India
ullivinaybabu@gmail.com[1], jyoti.bhu.ac.in[2], anindita.mondal@research.iiit.ac.in[3]

Abstract

This paper describes the submission by Team Aurum to the CL4Health @ LREC 2026 Shared Task on Case Report Form (CRF) Filling from dyspnea patient clinical notes. Extracting 134 structured clinical fields using a single Large Language Model (LLM) call often leads to schema-following errors, hallucination, and poor attention over complex instructions. To address this, we propose a modular extraction pipeline built with DSPy, which decomposes the massive CRF schema into 14 specialized, manually designed domain extractors (e.g., Medical History, Lab Values, Acute Diagnoses). Using *Chain-of-Thought* reasoning and strict Pydantic-typed validation with the Qwen3-Max (Thinking) model, our pipeline achieved an official Codabench Test Macro-F1 score of 0.68 in English and 0.67 in Italian, securing the 1st place ranking overall in the shared task. Through rigorous manual error analysis and iterative prompt optimization, we demonstrate that establishing explicit instruction boundaries between "not mentioned" (`None`) and "explicitly absent" (`False`) is critical for clinical information extraction, and that architectural modularity yields higher performance gains than simply scaling model parameters.

Keywords: Case Report Form, Clinical Information Extraction, Large Language Models, DSPy, Multilingual NLP

1. Introduction

The extraction of structured clinical variables from unstructured electronic health records (EHRs) is a foundational task in medical informatics, enabling downstream applications such as cohort selection, predictive modeling, and clinical trial matching (Wu et al., 2023). To support this, robust datasets are essential (Johnson et al., 2016). Recently, significant efforts have targeted the conversion of annotated clinical cases into structured data (Ferrazzi et al., 2025), including the automatic filling of Case Report Forms (CRFs) directly from emergency department notes (Kaczmarek et al., 2026).

Building upon this foundation, the CL4Health 2026 CRF:Filling shared task (Ferrazzi et al., 2026b) challenges systems to automatically extract 134 fields across English and Italian clinical notes describing patients presenting with dyspnea.

While Large Language Models (LLMs) have demonstrated remarkable capabilities in clinical text understanding (Singhal et al., 2023; Thirunavukarasu et al., 2023), the primary challenge of this task lies in the massive output schema and the strict semantic distinction between an explicitly negated clinical finding and an unmentioned one. Initial monolithic approaches, attempting to extract all 134 fields in a single LLM prompt, proved brittle. Models exhibited a strong tendency to hallucinate schema structures (Ji et al., 2023) and over-reason, inferring diagnoses from symptoms rather than strictly extracting what was explicitly stated by the clinician (Lyu et al., 2023).

To overcome these challenges, we developed a highly modular, language-agnostic pipeline powered by the DSPy framework (Khattab et al., 2023). We partitioned the 134 fields into 14 distinct extractors, utilizing Qwen3-Max (Thinking) (Bai et al., 2023) coupled with step-by-step reasoning. This paper details our pipeline architecture, prompt optimization methodology, results, and comprehensive error analysis.

To facilitate reproducibility and further research, the complete source code, DSPy module definitions, and inference scripts are made publicly available on github at `https://github.com/vinayulli/crf-lrec-sharedTask`.

2. Methodology

2.1. Modular Extractor Architecture

Instead of relying on a single monolithic LLM call, which has been shown to degrade attention over long contexts and complex instructions (Liu et al., 2024b), we decomposed the CRF into 14 specialized extractors. We manually designed these categories based on clinical domain expertise and the natural semantic groupings of the target variables. This approach isolates semantic domains, allowing the LLM to focus on specific sections of the clinical note with targeted instructions (Zhou et al., 2022).

Figure 1 illustrates the overall data flow and architecture of our proposed system. An unstructured

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 395–401
12 May 2026.

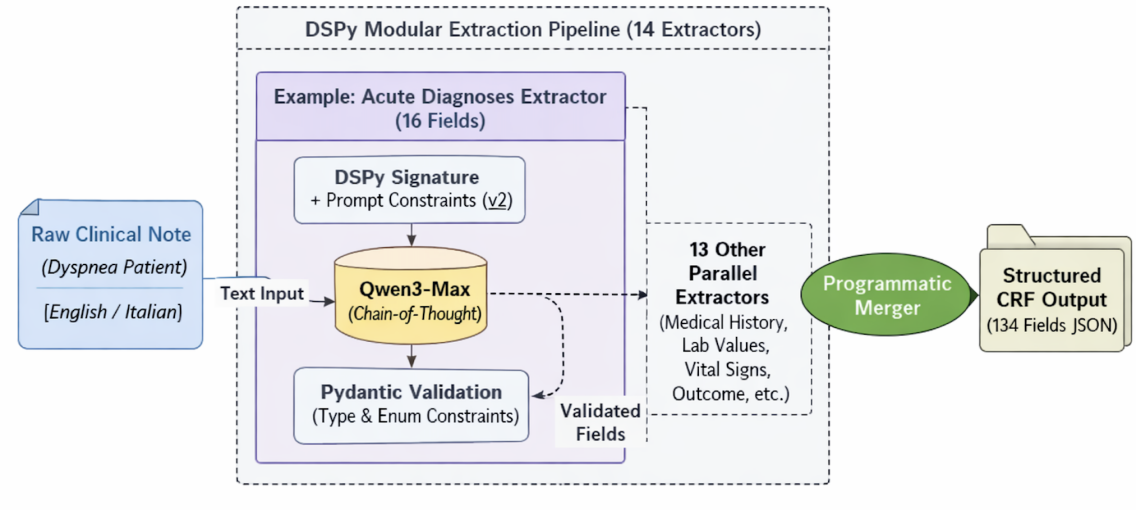

Figure 1: Architecture of the modular DSPy extraction pipeline, illustrating the parallel data flow from the raw clinical note to the final 134-field structured CRF.

clinical note is passed simultaneously to the 14 parallel extractors. Within each extractor module, the input is processed by a specialized DSPy signature combined with domain-specific prompt constraints. The Qwen3-Max model then applies step-by-step reasoning to extract the target variables, which are subsequently validated and type-cast using Pydantic schemas. Finally, a programmatic merger aggregates the validated fields from all 14 modules into the unified 134-field CRF JSON object.

The complete mapping of the 14 extractors to their corresponding 134 clinical features is detailed in Table 1.

Each extractor is implemented as a strongly-typed DSPy `Signature`. The output is strictly constrained using Pydantic (Colvin et al., 2023) sub-models, where all fields are typed as `Optional` and default to `None` (representing "unknown" or "not mentioned").

2.2. Chain-of-Thought and Reasoning

For each signature, we utilized `dspy.ChainOfThought`. Before outputting the structured JSON, the model is prompted to generate a reasoning trace. This forces the model to sequentially locate explicit mentions in the text before assigning a boolean or categorical value, significantly reducing hallucination (Kojima et al., 2022). The 14 extractors run sequentially over each clinical note, and their outputs are programmatically merged into a unified 134-field object.

2.3. Prompt Engineering and Constraints

Initial experiments (Prompt v1) revealed a massive class of false positives (420 FPs in the dev set). The root cause was the model's inductive bias to interpret the absence of a mention as a definitive negative (`False`), a well-documented issue in LLM-based zero-shot extraction (Zheng et al., 2023).

To correct this, we implemented Prompt v2 via iterative manual refinement (Chen et al., 2023). We injected explicit instruction boundaries into the DSPy docstrings. For example, the v2 prompt for Acute Diagnoses explicitly stated:

> *CRITICAL RULES: None means the condition is NOT MENTIONED AT ALL. This is the DEFAULT. False means the note EXPLICITLY STATES the condition is absent (e.g., "no pneumothorax"). Do not infer any diagnosis from symptoms alone.*

Furthermore, to ensure structured outputs, we constrained the generation using Pydantic `Enum` classes. For instance, categorical fields like *blood pressure* were restricted to exact string literals (`"normotensive"`, `"hypertensive"`, `"hypotensive"`). If the model's *Chain-of-Thought* reasoning failed to map to these strict categories, the system safely fell back to `None`.

2.4. Zero-Shot Multilingual Transfer

Rather than employing translation steps or language-specific pipelines, which can introduce compounding translation errors (Conneau et al., 2020), we relied on the native multilingual alignment of modern foundation models (Devlin et al., 2019; Lin et al., 2022). This zero-shot transfer is particularly crucial given the specific terminological nuances and challenges of performing medical NLP natively in Italian (Ferrazzi et al., 2026a). The

Extractor (# Fields)	Extracted Features
Medical History (22)	chronic pulmonary disease, chronic respiratory failure, chronic cardiac failure, chronic renal failure, chronic metabolic failure, chronic rheumatologic disease, chronic dialysis, active neoplasia, cardiovascular diseases, diffuse vascular disease, dementia, neurodegenerative diseases, neuropsychiatric disorders, peripheral neuropathy, epilepsy/epileptic seizure, known history of epilepsy, history of alcohol abuse, history of drug abuse, history of allergy, history of recent trauma, immunosuppression, pregnancy
Lab Values (19)	hemoglobin, leukocytes, platelets, creatinine, blood glucose, blood potassium, blood sodium, blood calcium, c-reactive protein, d-dimer, troponin, bnp or nt-pro-bnp, transaminases, serum creatinine kinase, inr, lactates, blood alcohol, blood drug dosage, urine drug test
Acute Diagnoses (16)	pneumonia, ab ingestis pneumonia, pulmonary embolism, pneumothorax, acute coronary syndrome, heart failure, acute pulmonary edema, cardiac tamponade, aortic dissection, arrhythmia, severe anemia, intoxication, respiratory failure, asthma exacerbation, copd exacerbation, covid 19
Treatments (14)	administration of oxygen/ventilation, administration of bronchodilators, administration of diuretics, administration of fluids, administration of steroids, blood transfusions, cardio-pulmonary resuscitation, performance of thoracentesis, palliative care, antihypertensive therapy, anticoagulants or antiplatelet drug therapy, antiepileptic therapy already in place, poly-pharmacological therapy, compliance with antiepileptic therapy
Imaging & Diagnostics (13)	chest rx, chest ct scan, abdomen ct scan, brain ct scan, brain mri, cardiac ultrasound, thoracic ultrasound, compression ultrasound (cus), ecg, ecg monitoring, eeg, pulmonary scintigraphy, gastroscopy (all evaluated for the presence of any abnormalities)
Epilepsy Assessment (11)	tonic-clonic seizures, further seizures in the ed, first episode of epilepsy, stiffness during the episode, eye deviation during the episode, pale skin during the episode, drooling during the episode, tongue bite, drowsiness confusion disorientation as postcritical state, duration of the patient's unconsciousness, duration of the patient's consciousness recovery
Current Presentation (10)	presence of dyspnea, presence of respiratory distress, chest pain, agitation, blood in the stool, foreign body in the airways, general condition deterioration, head or other districts trauma, hemorrhage, concussive head trauma
Vital Signs (7)	blood pressure, heart rate, body temperature, respiratory rate, spo2, level of consciousness, level of autonomy (mobility)
Syncope Assessment (7)	situational syncope, tloc during effort, tloc while supine, supine-to-standing systolic blood pressure test, carotid sinus massage, presence of prodromal symptoms, situation description
Arterial Blood Gas (4)	ph, pao2, paco2, hco3-
Social Context (4)	living alone, homelessness, need but absence of a caregiver, problematic family context
Infection Screening (3)	influenza and various infections, sars-cov-2 swab test, neurologist consultation
Devices (2)	presence of pacemaker, presence of defibrillator
Outcome (2)	improvement of dyspnea, improvement of patient's conditions

Table 1: Mapping of the 14 DSPy extractors to their corresponding 134 clinical features.

exact same English DSPy signatures and prompts were applied directly to the Italian clinical notes.

3. Experimental Setup

Dataset: Our pipeline was developed and evaluated using the official CL4Health dataset, consisting of 80 Development documents and 200 Test documents per language (English and Italian) (Ferrazzi et al., 2026b). Evaluation on the test set was performed using the official Codabench Macro-F1 scorer (Xu et al., 2022). For our local development set evaluations, we calculated the Macro-F1 scores using the official evaluation script provided by the

organizers.[1]

Models evaluated: We evaluated 7 different foundation models during the development phase. These included Qwen3-Max (Thinking), Qwen3-8B (Bai et al., 2023), GPT-4o, GPT-4o-Mini (Achiam et al., 2023), Llama-4-Maverick (Touvron et al., 2023), and Gemma-3-12B-IT (Team et al., 2024) and Deepseek V3(Liu et al., 2024a).

4. Results

4.1. Development Set Evaluation

Table 2 reports the Macro-F1 scores obtained by various LLMs on the development set. We evaluated several multilingual foundation models, including Llama4 Maverik, GPT-4o, GPT-4o Mini, DeepSeek-V3, Gemma-3-12B-Instruct, and Qwen-series models. Among the evaluated models, Qwen3-Max (Thinking) with the optimized v2 prompts achieved the best performance on the English development set with a Macro-F1 score of 0.70, improving from 0.67 with the initial v1 prompt. GPT-4o achieved 0.68 on English and 0.65 on Italian, while GPT-4o mini produced comparable results (0.67 EN / 0.65 IT). The few-shot GPT-4o configuration slightly underperformed the zero-shot setup (0.65 EN / 0.63 IT). DeepSeek-V3 demonstrated strong multilingual performance, achieving 0.66 on English and 0.67 on Italian. Smaller models such as Qwen3-8B achieved competitive performance on English (0.67) despite having significantly fewer parameters.

Model Name	Prompt	English	Italian
Llama-4-Maverick	v2	0.58	0.58
GPT-4o	v2	0.68	0.65
GPT-4o few shot (3)	v2	0.65	0.63
GPT-4o-mini	v2	0.67	0.65
Qwen3-Max (Thinking)	v1	0.67	0.66
Qwen3-Max (Thinking)	**v2**	**0.70**	**0.69**
Qwen3-8B	v2	0.67	0.60
DeepSeek-V3	v2	0.66	0.67
Gemma-3-12B-Instruct	v2	0.62	0.60

Table 2: Development set Macro-F1 scores across different foundation models. The Qwen3-Max (Thinking) model with v2 prompts achieved the highest performance in both languages.

4.2. Official Test Set Results

Based on dev set performance, we selected Qwen3-Max (Thinking) with v2 prompts for our primary test submissions. As shown in Table 3, the system demonstrated excellent language transfer, with only a marginal 0.01 F1 drop when applied zero-shot to the Italian test set.

Language	Macro-F1
English (EN)	**0.68**
Italian (IT)	**0.67**

Table 3: Official Test Set Results.

5. Error Analysis

To understand the limitations of LLMs in clinical CRF extraction, human annotators from our team conducted a rigorous manual error analysis on the development set predictions, identifying four primary root causes:

1. "None" vs "False" Confusion (Over-prediction): In v1, ∼80% of all false positives stemmed from the model extracting `False` when a condition was simply not mentioned. While v2 prompts reduced FPs massively (from 420 down to 256), the model occasionally overcorrected, becoming too conservative and generating false negatives (increasing from 76 to 95) on genuinely absent conditions.

2. Unwarranted Clinical Inference: The *Chain-of-Thought* methodology occasionally harmed extraction by mimicking clinical diagnosis. For example, if an ECG showed an abnormality, the model inferred `arrhythmia = True` even if the clinician never explicitly documented the diagnosis. The strict ground truth standard clashes with the model's natural diagnostic reasoning abilities.

3. Lab Value Format Mismatches: Because evaluation relies on exact string matching, variations in formatting severely penalized the system, a known limitation of lexical evaluation metrics in NLP (Kamalloo et al., 2023). For instance:

- GT: `Hb 12` | Pred: `12` (Error: Missing prefix)
- GT: `24530` | Pred: `WBC 24530` (Error: Added prefix)
- GT: `149` | Pred: `14.9` (Error: Unit conversion)

Instructing the model to "strip all labels" improved precision on some notes but generated errors on others where ground truth inconsistently retained labels.

6. Conclusion

our submission demonstrates that modularity is essential when extracting large, complex CRFs using LLMs. By breaking 134 fields into 14 distinct DSPy extractors and utilizing the reasoning capabilities

[1] https://github.com/hltfbk/CRF-filling-CL4Health2026

of Qwen3-Max, we achieved highly competitive Macro-F1 scores of 0.68 (EN) and 0.67 (IT). Our analysis highlights that the true frontier in clinical information extraction is not merely increasing model size, but strictly bounding model reasoning to distinguish between implicit absence and explicit clinical negation, alongside overcoming the brittleness of exact-string matching evaluations.

7. Limitations

While our modular pipeline achieved state-of-the-art results, it introduces higher computational overhead and inference latency compared to a monolithic approach, as it requires 14 parallel LLM calls per clinical note. Additionally, the system's performance is bottlenecked by the exact-string matching evaluation metric, which occasionally penalizes clinically valid formatting variations (e.g., predicting "14.9 g/dL" instead of "149"). Finally, severe data sparsity in certain CRF domains (e.g., *Syncope* and *Outcome*) makes it difficult to reliably evaluate the model's true extraction capabilities for those specific fields.

8. Ethics Statement

This research utilizes de-identified clinical data provided by the CL4Health 2026 organizers, strictly preserving patient privacy. Because Large Language Models remain susceptible to hallucinations and incorrect medical inferences, our extraction pipeline is designed solely as an assistive tool for retrospective data structuring. It is not a substitute for human clinical judgment. Any real-world deployment of this system must incorporate robust human-in-the-loop validation to ensure patient safety and data integrity.

9. Acknowledgements

We thank the organizers of the CL4Health @ LREC 2026 shared task and Codabench for facilitating this challenge.

10. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, et al. 2023. Pydantic: Data validation using python type hints. In *pydantic: Data validation using Python type hints*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian.

Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026b. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. Converting annotated clinical cases into structured case report forms. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Viena, Austria. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. Toward automatic filling of case report forms: A case study on data from an italian emergency department.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 5591–5606.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

11. Language Resource References

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad

Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.