No.	Title	Brief Description/Problem Statement	Datasets	Reference Link
1	Attribute-value extraction from Product Description	Given a product description, the Attribute-value extraction problem aims to extract the attribute and its corresponding value. Example: Product title: Safari Crescent 8 Wheels 66cm Medium Check-in Trolley Bag Hard Case Polycarbonate 360 Degree Wheeling System Luggage, Travel Bag, Suitcase for Travel, Trolley Bags for Travel, Thyme Green Extracted attribute-value: Wheels: 8 Type: Trolley Colour: Thyme Green Material: Polycarbonate	1. https://raw.githubusercontent.com/lanmanok/ACL19_Scaling_Up_Open_Tagging/master/publish_data.txt (AliExpress) 2. Scaling up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title (Companion paper to above dataset)	AttriSAGE: Product Attribute Value Extraction Using Graph Neural Networks (EACL, 2024)
2	Generating/Expanding Queries from Documents	Given a document, identify the queries that can be answered from the documents. In other words, identify/generate for which queries, the current document will be a good retrieval target. Another variant of the problem would be to have a document and a seed query as an input, and predicting the reformulated/expanded form of the query.		1. Doc2Query--: When Less is More 2. From doc2query to docTTTTTquery
3	Better Retrieval for Generation	Retrieving additional contents from a collection for a given context, and using these contents as expanded context can be helpful for various tasks. In this project, you will deal with the retrieval strategies that can fetch good-quality context from existing document pool, given a seed context.	IndicMARCO dataset	1. Searching for Best Practices in Retrieval-Augmented Generation

				2. IndicIRSuite : Multilingual Dataset and Neural Information Models for Indian Languages
4	Hate speech and counter-speech	Counter-speech generation for hateful and harmful content	1. HARE: Explainable Hate Speech Detection 2. https://github.com/marcoguerini/CONAN	1. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning 2. Probing LLMs for hate speech detection: strengths and vulnerabilities
5	Judicial Citations to Precedence	Legal passage retrieval aims to predict the correct target passage given the preceding context. Formally, the task is given a legal context x_i, the task is to retrieve the relevant cited passage y_i from the set of all possible passages $\{y_1, y_2, ..., y_n\}$. Note: Please refer to the paper to better understand the legal jargon.	LePaRD dataset	LePaRD: A Large-Scale Dataset of Judicial Citations to Precedent (ACL, 2024)
6	Unified Embedding Model For Diverse Retrieval Augmentation in	Train an LLM based embedder which can act as a universal embedding model for retrieval for a variety of	Information Retrieval, In context learning	A Multi-Task Embedder For

	LLMs	tasks supported by LLMs, including knowledge retrieval, memory retrieval, example retrieval, and tool retrieval.		Retrieval Augmented LLMs
7	Speech to Text Conversion for academic lectures		https://github.com/Jack-ZC8/M3AV-dataset/tree/main/download [Check the dataset availability]	M3AV: A Multimodal, Multigenre, and Multipurpose Audio-Visual Academic Lecture Dataset
8	Culture in LLMs*	Probe whether/how cultural contexts are reflected in LLM responses	May need to be created. Some entries from https://github.com/SeaEval/SeaEval can be used	1. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models 2. Towards Measuring and Modeling "Culture" in LLMs: A Survey
9	Open Ended Math Question Answering	Generate answers to Math questions based using LLMS, and analyze their performance	https://www.cs.rit.edu/~dprl/ARQMath/	Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange (SIGIR, 2024)
10	Medical Dialogue Generation	Generating a natural, human-flowing, and accurate dialogue system is crucial for high-risk medical and	MedDialog dataset	Incorporating Medical Knowledge

		legal services. The task is to generate coherent dialogues between a patient and a doctor. Particularly, given a patient-doctor dialogue as a set of pairs {(s_i, t_i)}, where source s_i is the dialogue from a patient and target t_i is a doctor's response. A dialogue generation model generates t from s.	MedDialog: Large-scale Medical Dialogue Datasets (EMNLP 2020)	to Transformer-based Language Models for Medical Dialogue Generation (BLP, 2022)
11	Question Answering over Tabular Data using Large Language Models	Large language models have been shown to perform well with free-flow natural text. However, more structured data like tables LLMs still lag behind in performance. The task is to explore NLP techniques to improve the Question Answering task from Tabular data. SEMEVAL 2025 Shared Task	DataBench dataset Question Answering over Tabular Data with DataBench: A Large-Scale Empirical Evaluation of LLMs (DataBench Paper; LREC 2024)	
12	Simplifying Text	Document-level text simplification is a specific type of simplification that involves simplifying documents consisting of several sentences by rewriting them into fewer or more sentences. In this project, we will aim to simplify an input text which might be difficult to comprehend. Such texts often occur in legal documents, domain specific documents, etc.	Dataset: https://github.com/epfml/easy-summary/tree/main/SimSum/data	Paper: SIMSUM: Document-level Text Simplification via Simultaneous Summarization