# Transformer-based Large Language Models for Passage Retrieval

Passage retrieval is a common Natural Language Processing (NLP) task, which consists of proposing a list of ordered passages from a large corpus, which answer the given query, expressed in natural language.

Classical IR approaches such as **TF-IDF** and **BM25** [1] work at syntax-level and often fail to capture the vocabulary misalignment between query and passages. While **Learning to Rank** [2] leverages machine-learning techniques to address this issue, their models are shallow and fitted over handcrafted IR features.

With the development of pre-trained transformer-based Language Models (LM) such as T5 [3], it is possible to employ large-scale neural models to bridge the gap. On the **MS MARCO** dataset [4] for Passage Retrieval task, these LM-based methods are unarguably dominating the leaderboard.

Most successful solutions adapt a “funnel” approach, where

- a **retriever** searches efficiently in a large corpus and returns a reasonable number of candidates with high recall,
- a **ranker** then scores the candidates with more elaborated (and time-consuming) models, usually based on language models

The entire system must take no more than **hundreds of ms** to be useful in production. Techniques such as **Question Generation** [5] can also be used to further boost the perf.

As part of the CIB Analytics Consulting team, your role will be to **investigate the creation of such LM-based IR model** according to the following steps

1. Familiarize with state-of-the-art contextualized embedding
2. Familiarize with classical Neural-based LR methods
3. Get inspiration from MS MARCO leaderboard and implement a ranking/re-ranking model capable of working in **real-time**
4. Evaluate the model on open-source dataset with NDCG and MAP
5. Qualify the model on internal datasets

Apart from the above points, many axes are possible for further investigation, for example:

- Investigate **Zero-Shot Learning** [6] on out-of-domain datasets
- Explore Language Model compression methods such as **Knowledge Distillation**

**References:**

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[2] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. Found. Trends Inf. Retr. 3, 3 (March 2009), 225-331

[3] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[4] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng. 2018. MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset

[5] Hangbo Bao *et al*, UNILMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training

[6] Zero-shot Question Generation: Accelerate the Development of Domain-specific Dialogue System

**Practical Information:**
**Supervisor:** CASANOVA Pierre-Yves, Data Scientist, pierre-yves.casanova@bnpparibas.com
**Location:** Grands Moulins de Pantin, 9 Rue du Débarcadère, 93500 Pantin (Possible partial work from home)
**Compensation:** Competitive salary
**How to apply:** Please submit your application at paris.cib.analytics.consulting.careers@bnpparibas.com