## Scientific Literature **Understanding**

**Scientific Document** 

Task

**Prediction** by LLMs

Abstract: ... We investigated whether single nucleotide polymorphisms (SNPs) of HNF4A,

encoding HNF4alpha, influenced the conversion from impaired glucose tolerance (IGT) to type 2 diabetes mellitus in subjects of the STOP-NIDDM trial. This trial aimed at evaluating the effect of acarbose compared to placebo in the prevention of type 2 diabetes mellitus. Three SNPs in the P2 promoter region (rs4810424, rs1884614, and rs2144908) were in almost complete association (D'>0.97, r (2)>0.95) and, therefore, only rs4810424 was included in further analyses. No association was found in men. Haplotype analysis based on three SNPs (rs4810424, rs2071197, and rs3818247) indicated that the conversion to type 2 diabetes mellitus was

You will be shown an abstract from a biomedical research paper. Given this abstract, your task is to extract all unique entities of the following types: ["CellLine", "Disease", "Variant"]. Please return the output as a JSON object of the format: {"CellLine": ["hRPTEC", ...], "Disease": ["Diabetes", ...], "Variant": ["Pro75Ala", ...]}.

// Rest of the task.

## Galactica-6.7B

dependent on the number of risk alleles...

Answer:

I think you can use spaCy to do this.

spaCy is a library for natural

language processing (NLP) in Python. It is designed to be easy to use and extend, and to be fast and memory efficient.

You can use it to extract entities from your abstract.



Unfamiliar with Scientific Tasks

## Llama-3-7B

```
"CellLine": [
   "hRPTEC",
 "Disease": [
   "Type 2 diabetes
mellitus"
 "Variant": [
   "Pro75Ala",
   "rs4810424".
   "rs1884614".
   "rs2144908",
   "rs2071197",
```

Lack of Scientific Knowledge

## SciLitLLM-7B (Ours)

```
"CellLine": [],
 "Disease": [
   "IGT".
   "impaired alucose
tolerance",
   "type 2 diabetes mellitus",
   "diabetes"
 "Variant": [
   "rs3818247",
   "rs2071197".
   "rs1884614".
   "rs2144908",
   "rs4810424"
```

Correct Answer