

附表 1 实体抽取任务提示词模板

Tab.1 Template for Prompt Words in Entity Extraction Task

类型 Type	提示词内容 Prompt content
角色定位 Role positioning	假设你是一位生物制造领域专家，擅长从英文专利或论文科技文献中提取描述技术的命名实体，以及应用场景的命名实体。
任务描述 Task description	1.提取描述技术的命名实体。 ①从提供的英文专利或论文科技文献中准确提取描述技术的命名实体。 ②仅提取该文本使用的技术，不提取与该文本使用技术进行对比的技术。 ③确保提取的命名实体能够代表核心技术与核心方法。 2.提取描述应用场景的命名实体。 ①从提供的英文专利或论文科技文献中准确提取描述技术应用场景的命名实体。 ②技术应用场景的命名实体通用出现在"applied to"、" the task of"、"methods for"、"our approach has implications for"、"a approach to"等描述附近。 ③应用场景词通常不多，且不包含描述技术的词语。 ④确保提取的命名实体能够代表核心技术与核心方法的应用场景。
抽取规则 Extraction rules	仅从提供的英文文本中提取命名实体，切勿编造或改编词语。
输出格式 Output format	提取的技术实体和应用场景实体以 JSON 格式输出，如下所示： { "tech_entities":["技术实体 1","技术实体 2",...], "scene_entities":["应用场景实体 1","应用场景实体 2",...]}
样例提示 Sample prompt	输入： "Drug Repositioning via Multi-View Representation Learning With Heterogeneous Graph Neural Network.Exploring simple and efficient computational methods for drug repositioning has emerged as a popular and compelling topic in the realm of comprehensive drug development..." 输出： { "tech_entities":["Multi-view Representation Learning method","Heterogeneous Graph Neural Network","feature aggregation model with variable-length meta-paths","transformer based semantic aggregation module","multi-view fusion decoder with an attention mechanism"], "scene_entities":["drug repositioning","identifying potential drug-disease associations"] }

附表 2 大语言模型抽取与人工标注实体对比结果示例

Tab.2 Examples of Engineering Extraction Techniques and Application Scenarios Entities

编号 Number	标题 Title	抽取方法 Extraction method	技术实体 Technical entities	应用场景实体 Application scenario entities
WOS:000777 408600036	AF2Complex predicts direct physical interactions in multimeric proteins	人工标注	AF2Complex, deep learning, AlphaFold2 (AF2), neural network models, AF-Multimer	protein-protein interactions, predict the structures of multimeric protein complexes

protein design;protein engineering;protein conformations;protein structure representation;protein structure prediction;design of dynamic proteins 等	protein design
---	----------------

附表 4 BERTopic 模型参数设置

Tab.4 BERTopic Model Parameter Settings

参数 Parameter	释义 Definition	设置值 Set value
n_neighbors	构建高维数据流形结构时，局部邻域的大小	10
n_components	降维后的维度	50
min_dist	控制点之间的最小距离	0.01
metric	相似度度量	cosine
min_cluster_size	最小簇的大小	50
min_samples	簇的核心点数量	30
metric	用于计算高维空间中距离的度量	euclidean
top_n_words	每个主题提取的单词数	20
min_topic_size	最小主题大小	50
nr_topics	自动减少主题数量	auto
calculate_probabilities	计算每个文档中所有主题的概率，而不是每个文档中指定主题的概率	True
verbose	更改模型的详细程度	True