PDF 格式文档的核心优势与处理逻辑

lixun-robot 对 PDF 格式的支持包含多重优化，不仅能提取纯文本内容，还能识别 PDF 中的段落分隔符和标题层级。对于扫描版 PDF（图片转文字），若已通过 OCR 识别为可编辑文本，机器人同样能正常处理；若为未识别的图片 PDF，则会提示“无法提取有效文本”，避免无效加载。

PDF 文档的分段策略与其他格式不同，系统会以 PDF 中的“页面分隔”和“空行”作为分段依据，确保每个片段对应原始文档的一个完整逻辑单元。例如，一份 3 页的 PDF 文档，若每页有 2 个段落，会被拆分为 6 个独立片段，每个片段标注对应的页码和段落索引，方便用户追溯原文。

在向量检索场景中，PDF 文档的内容权重与其他格式一致，不会因格式差异导致检索偏差。用户提问涉及 PDF 中的专有信息时，机器人会优先返回 PDF 文档的相关片段，并在引用中明确标注“PDF 格式测试文档.pdf”，确保来源可追溯。