PDF表格解析与结构识别测试基准文档

文档版本: v1.0.0 | 测试目的: 复杂表格、跨行跨列、嵌套样式解析

本文件专门用于测试PDF解析工具（如 PaddleOCR、TableTransformer、Camelot、Pdfplumber 等）对复杂HTML表格结构的还原能力。文档中包含多种单元格合并形式（`colspan` 和 `rowspan`），以评估解析引擎在几何布局、文本对齐及逻辑关联上的准确性。

一、 基础合并结构测试表格（销售与财务季报）

下表展示了一个典型的企业季度财务与销售汇总，包含顶部的跨列标题、左侧的跨行分类，以及右下角的双向合并单元格。

2025年度华东区核心产品季度销售与利润综合分析表					
产品大类	**子品类代号**	**第一季度 (Q1)**	**第二季度 (Q2)**	**第三季度 (Q3)**	**第四季度 (Q4)**
智能硬件 (IoT)	A-01 智能穿戴	120,500	142,300	138,000	189,400
	A-02 家居控制	89,000	95,400	102,000	121,500
	A-03 车载车载	210,000	235,000	220,000	268,000
云端服务 (SaaS)	B-01 企业ERP	450,000	462,000	475,000	512,000
	B-02 数据分析	310,000	325,000	340,000	398,000
季度运营总成本 (RMB)		280,000	295,000	310,000	345,000
净利润表现及年度评级		899,500	964,700	965,000	**年度评级: 优秀 (A+)**
备注：同比核心增长率达 $\Delta = 14.5\%$		以上数据均经过外部独立审计机构复核确认			

二、 高难度复杂嵌套与不规则合并测试表格

下表设计了不规则的跨行跨列交错，用于极端测试。解析工具需要准确识别每一个逻辑格子（Logical Cell）与物理边界（Bounding Box）的映射关系。

主指标分类	次级参数	环境配置 A (Baseline)		环境配置 B (Optimized)	
		阈值上限	响应延时	吞吐效率	稳定性权重
系统内核层 (Kernel)	内存并发	1024 MB	12 ms	94.2%	混合加权系数: $W_{sys} = 0.85$ 稳定运行期: 480小时无故障
	线程调度	2048 MB	8 ms	97.8%	
应用架构层 (API)	数据流转	通道合并测试：带宽负载 < 45%		91.5%	
	安全握手	512 MB	45 ms	99.1%	
全局综合评估结果 (Synthesis)		1. 核心链路解析正常率：**99.95%** 2. 异常单元格鲁棒性测试：通过			