

Layout-Independent Utility Data Extraction: A Semantic Approach Integrating RPA and Generative AI

Prasertsak Tiawongsombat¹, Choopan Rattanapoka^{2*}

¹Dept. of Electronics Engineering Technology, College of Industrial Technology,
King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road,
Wongsawang, Bangsue, Bangkok, Thailand

²Dept. of Electronics Engineering Technology, College of Industrial Technology,
King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road,
Wongsawang, Bangsue, Bangkok, Thailand

**Corresponding Author Email: choopan.r@cit.kmutnb.ac.th*

Abstract

Global regulatory frameworks, notably the Corporate Sustainability Reporting Directive (CSRD), increasingly mandate rigorous auditing of Scope 2 carbon emissions. However, the extraction of consumption data from heterogeneous utility bills remains a manual, error-prone bottleneck, as traditional template-based Optical Character Recognition (OCR) methods struggle with diverse, unstructured document layouts. To address this limitation, this paper proposes a fully automated framework integrating Robotic Process Automation (RPA) and a Generative AI model (Gemini 2.5 Flash). We introduce a 'Context-Aware Extraction' methodology that dynamically identifies essential data fields through semantic reasoning, entirely eliminating the need for rigid spatial templates. Experimental evaluation on a dataset of 225 heterogeneous electricity bills—targeting seven key entities per document to yield 1,575 discrete target fields—demonstrated an average extraction accuracy of 91.37% when benchmarked against a human-transcribed ground truth. Furthermore, the proposed system achieved an operational latency of under one second per document. These results validate the framework as a highly scalable and auditable solution for real-time carbon accounting and regulatory compliance.

Keyword: Carbon Accounting, Large Language Models, Robotic Process Automation, Scope 2 Emissions, Zero-Shot Extraction