Paper ID: PC26-0417 | Session: Structured Data & ETL
Conference: PelicanConf 2026, Coastal Machine Learning
Authors: Elena Vasquez; Kenji Tanaka; Fatima Al-Rashid
Affiliations: Tidewater AI Research; Sakura Institute of Informatics

Schema-Constrained Decoding for Reliable PDF-to-JSON Extraction

We show that constraining a language model's decoder to a target JSON schema eliminates a class of structural errors that plague post-hoc parsing. Across 2,100 mixed-layout documents, schema-constrained decoding cut malformed-output rate from 6.3 percent to under 0.1 percent while leaving field accuracy unchanged. We pair this with a provenance span for each value, enabling a reviewer to jump from any cell to its source location. The combined system underpins a production extraction service handling 40,000 documents daily.

Keywords: constrained decoding, JSON schema, document extraction, provenance spans, reliability