Large Language Models as Control Planes for Industrial-Scale Web Data Extraction

Felipe Marineli
Università Roma Tre | Meltwater Deutschland GmbH
Rome, Italy | Berlin, Germany
felipe.marineli@{uniroma3.it|meltwater.com}
Supervised by Valter Crescenzi, Paolo Merialdo, and Valerio Cetorelli

ABSTRACT

Web-scale information extraction faces a fundamental trade-off: rule-based wrappers are brittle and vulnerable to drift, while end-to-end LLM extraction is accurate but costly and opaque. We introduce a pipeline that promotes the LLM to the *control plane*, leaving fast, transparent wrappers in the data plane and letting the model monitor drift and auto-repair them at scale.

The architecture rests on three pillars to be developed during the PhD. *(i) URL discovery*: an agnostic module that exploits temporal link-graph signals to surface high-value pages without manual seed tuning. *(ii) Structural templating*: a formal grammar-based clustering that groups pages into stable templates and defines reusable wrapper scopes. *(iii) LLM control plane*: agentic LLMs that both supervise the pipeline and repair wrappers when drift is detected.

By fusing URL discovery, theory-grounded templating, and LLM-based wrapper induction, the system aims to transform hand-tuned heuristics into a self-healing, economically sustainable, fully autonomous web data extraction pipeline, orchestrated by a dedicated control plane. The full system will be field-tested in the domain of editorial news, an incremental, high-drift environment where layout changes and semantic diversity make robust extraction especially challenging. While initially developed within the domain of media intelligence, the architecture is designed for generalization to other structured web verticals.

VLDB Workshop Reference Format:
Felipe Marineli. Large Language Models as Control Planes for Industrial-Scale Web Data Extraction. VLDB 2025 Workshop: VLDB Ph.D. Workshop.

1 INTRODUCTION

The exponential growth of web-based content has outpaced the capabilities of traditional data extraction methods. Domains such as media monitoring, market intelligence, and competitive analysis require systems that can not only process vast and structurally diverse data sources but also adapt rapidly to frequent layout changes. Manual extraction and semi-automated workflows, while precise, are slow, labor-intensive, and brittle in the face of even minor structural drift. On the other end of the spectrum, direct invocation of large language models (LLMs) for document-level extraction as a solution for full automation introduces high operational costs and lacks predictable, controllable outputs—both of which are problematic in high-volume, production-grade systems.

Although LLM inference costs are evolving rapidly, a more fundamental challenge is consistently controlling model behavior. Traditional wrapper-based systems retain an advantage in this regard: they expose explicit, tunable logic that remains robust across structural changes. In contrast, modifying LLM outputs typically requires costly model fine-tuning or extensive prompt engineering and retrieval-augmented generation scaffolding. This "control problem" constitutes an additional hurdle to the reliable deployment of LLM-based extraction at industrial scales.

Meltwater, a global media-intelligence provider processing billions of news, blog, and social-media posts daily [9, 26–28, 35], illustrates this tension. Its crawler uses automatically induced wrappers built on shallow NLP heuristics and page descriptors, which require constant human retuning to handle structural drift. This mismatch has left the current system in a suboptimal position: while automation enables rapid expansion of coverage, the human oversight layer cannot keep up with the cases that need intervention. Our goal is to bridge this gap through smarter, modular automation. The architecture will be field-tested in Meltwater's editorial news pipeline—an environment of constant content churn, layout volatility, and semantic ambiguity; hence, an ideal setting to evaluate system resilience and wrapper adaptability under real-world drift.

We introduce an LLM-driven control plane that continuously evaluates extraction quality by monitoring runtime metrics and cross-referencing them with temporal and structural signals. Upon detecting drift—manifested as divergence between expected and actual extraction patterns—it issues targeted repair or regeneration directives to LLM agents operating in the data plane. This orchestration layer enforces system-wide policies, coordinates wrapper lifecycle events, and maintains consistent performance across heterogeneous and evolving web sources. The underlying architecture comprises three interdependent data plane modules: (1) URL discovery, which leverages temporal link-graph differentials to identify high-value, frequently updated content; (2) structural templating, which applies landmark-grammar clustering to define template-level parsing boundaries; and (3) LLM-based wrapper induction, where agents synthesize or modify extraction logic over clustered page structures under control plane supervision.

The system adopts a closed-loop control model inspired by the control and data plane abstraction in networking, enabling autonomous modules to collaborate through fine-grained signal exchange. The data plane operates at the content level (crawling, parsing, and extraction) while the control plane governs global

Proceedings of the VLDB Endowment. ISSN 2150-8097.

behavior, including drift detection, cost policies, and wrapper synthesis. URL discovery expands the crawl frontier with minimal assumptions; structural templating consolidates new layouts into reusable patterns. Unlike static architectures, this dynamic framework continuously integrates real-time signals across layers, allowing for a self-healing, adaptable, and scalable framework.

2 RELATED WORK

Traditional approaches to web data extraction have relied on manually and semi-automatically engineered wrappers [3], which map the DOM of web pages into structured data formats using predefined extraction rules. The wrappers are able to achieve high accuracy but suffer from substantial drawbacks: they are labor-intensive to create, costly to maintain, and fragile in the face of structural drift. Early systems automated wrapper generation with a mixture of rule-based ontologies, shallow NLP, and explicit page-layout descriptors [4, 7, 10, 15, 16, 23, 30]. While accurate, they remain brittle: at an industrial scale, pages that drift beyond the original distribution emerge regularly and require expensive manual fixes.

Past research has increasingly explored unsupervised template clustering to minimize manual effort and enhance scalability in web data extraction. Systems like iRobot [5] and *boilerplate detection* [22] group structurally similar web pages using DOM-based similarity metrics to identify repeating structural patterns. However, these template clustering methods face challenges in handling dynamic content updates, heterogeneous web structures, and noise from unstructured data or invalid and duplicated pages. To enhance URL discovery, focused crawling approaches like FoCUS [20] and iCrawl [17] integrate URL-based classification and prioritization algorithms. FoCUS, for instance, employs page classifiers and regular expressions to generate crawling specifications tailored to web forums; similarly, iCrawl leverages external signals to prioritize URLs based on relevance to topics of interest. Such URL discovery techniques significantly reduce crawling costs and improve the efficiency and relevance of content retrieval compared to a baseline, but rely on assumptions about the content that must be retrieved, either in terms of URL structure or content semantic similarity.

The emergence of LLMs has opened new avenues for automating extraction with fewer manual interventions and structural assumptions. Evaporate [2] uses LLMs to autonomously structure semi-structured documents without labeled training data. Its successor, Evaporate-Code+ [2] prompts LLMs to emit reusable Python functions, reducing repeated calls. Fine-tuned models advance performance across technical tasks [18, 19, 24], while other work shows off-the-shelf models handle complex challenges like entity matching and schema alignment with strong out-of-distribution robustness [1, 14, 31]. These gains are further enhanced by targeted human oversight, which improves reliability by reviewing ambiguous outputs, validating extractions, and refining clustering or wrapper synthesis pipelines [2, 11, 12]. Together, these methods point toward scalable, semi-autonomous extraction with humans in the loop—though not yet toward full autonomy.

LLMs also power agent frameworks that plan, coordinate, and revise workflows. AutoGen [37], CrewAI [13], and Agent-OM [32] exemplify multi-agent reasoning systems, while platforms like Kadoa [21] and ScrapeGraphAI [33] explore self-healing scrapers that detect layout drift and regenerate parsing logic. Research prototypes such as HuggingGPT [34] show tool-generation and verification loops, and autonomous agents like AutoGPT and Agent-OM demonstrate outcome monitoring and closed-loop correction. However, production-grade LLM control planes that orchestrate such agents across complex extraction workflows are still lacking.

Large-scale studies now range from AI-native data stores to fine-grained pipeline tools: LLMs can act as a "universal query interface," fusing search, inference, and transformation across heterogeneous data silos [25]; subsequent work elevates those capabilities into a declarative algebra that an engine can cost-optimize like SQL plans [29]; further research turns complex questions over data lakes into automatically composed, interactively refined workflows that blend retrieval, reasoning, and validation [36]; and another approach compiles natural-language task descriptions into hybrid pipelines that mix LLM calls with cheaper, learned components [8]. Collectively, these systems signal the feasibility of AI-native search, processing, and orchestration, yet each tackles only a slice of the end-to-end problem: none merges continuous quality-versus-cost control, cross-pipeline signal fusion, rigorous SLA enforcement, and human-override paths into a single production-grade control plane. Bridging that gap remains an open research frontier.

Our architecture tackles the core challenges of web-scale extraction by using temporal-signal URL discovery and landmark-grammar structural templating as robust, drift-aware primitives defining the crawl frontier and parsing boundaries. These components provide a stable, interpretable substrate for an LLM-driven control plane that enables closed-loop orchestration without brittle heuristics. While LLMs are leveraged in both operational and strategic capacities, the system distinguishes concerns by delegating routine extraction tasks to data plane agents and reserving high-level responsibilities—such as drift detection, policy enforcement, and wrapper lifecycle management—for the control layer. Fine-grained telemetry, cost-aware policies, and hierarchical escalation protocols ensure responsive behavior and bounded autonomy.

3 THE THREE PILLARS

Building on Meltwater's existing infrastructure for editorial news, our solution aims to balance scalability, robustness, and interpretability in a high-volume, drift-prone production setting. It replaces reactive maintenance with proactive orchestration by embedding LLM-driven reasoning across the pipeline.

At its core, the system is comprised of three pillars—URL discovery, structural templating, and LLM-based wrapper induction—each designed to replace brittle heuristics with modular, self-adaptive logic. These pillars operate in the *data plane*, performing discovery, parsing, and extraction. Above them sits a dedicated *LLM-diven control plane* that continuously monitors system health, detects drift, and orchestrates repairs, turning manual oversight into a policy-governed loop of autonomous adaptation.

The first pillar, URL discovery, continuously identifies content-rich web pages through dynamic URL discovery based on temporal signals. Starting from a small set of seed pages, it incrementally expands coverage by analyzing link graph changes within domains, ensuring proactive adaptation to new and evolving content sources without human intervention or domain assumptions [5, 17].

The second pillar, structural template clustering, systematically groups web pages according to structural similarity in their HTML (DOM) layout. Utilizing a landmark grammar approach [6], this clustering isolates and identifies structural patterns across pages, significantly enhancing wrapper generalizability and reducing redundant extraction logic. The landmark grammar method's practical viability at full production scale remains subject to validation.

The third pillar introduces LLM-based agents that perform wrapper induction within the data plane. When the control plane detects drift—by comparing runtime extraction results with expected patterns derived from URL discovery and template clustering—it triggers these agents to repair wrappers. Operating over structured templates, the agents update extraction logic autonomously, adapting to layout changes without manual input. This targeted activation ensures precise, cost-effective adaptation while maintaining stable, interpretable wrappers across evolving web content.

3.1 Preliminary Results

Pilot evaluations conducted on Meltwater's editorial news ingestion pipeline have shown that the URL discovery component consistently achieves high coverage with strong precision across web domains. It outperformed a sample of Meltwater's current system in both resource efficiency—defined as the number of unique publications over the total number of links crawled—and overall content coverage. The module proved robust in surfacing publications through temporal link-graph signals, underscoring its domain-agnostic design and adaptability to shifting web dynamics.

Upcoming development includes lightweight clustering prototypes employing landmark grammar and an essential LLM-agent loop aimed at automating wrapper synthesis and drift detection. These early-stage experiments aim to robustly assess feasibility, guide subsequent iterations, and lay a solid foundation for comprehensive validation prior to full-scale deployment.

4 CONCLUSION AND FUTURE WORK

Our solution targets automation precisely where human oversight, shallow NLP, and page descriptors have become bottlenecks. The pipeline preserves the transparency and control of wrapper-based logic while adding semantic adaptability via intelligent agents. Field validation focuses on editorial news—a domain central to Meltwater, marked by high turnover and structural volatility. This provides a high-volume, real-world testbed aligned with system assumptions.

Once validated, this architecture is expected to enable end-to-end autonomous operation—minimizing manual intervention, improving resilience to structural drift, and expanding Meltwater's data ingestion capabilities across a broader spectrum of domains and content types. The system thus represents both a practical advancement in industrial extraction and a model for scalable, LLM-integrated automation in complex web environments.

Future work will consolidate the three pillars—temporal-signal URL discovery, template clustering, and LLM-based wrapper induction—into a unified, production-grade framework orchestrated by a dedicated control plane. The discovery layer will be extended with LLM-guided link-prioritization that estimates content utility at crawl time, enabling low-latency feedback loops and smarter resource allocation. The templating engine will scale landmark-grammar clustering to millions of pages via algorithms optimized for hierarchical structure and cross-domain generalization. On the control plane side, ongoing research will focus on drift signal modeling, prompt sensitivity analysis, and ambiguity-aware wrapper control. A multi-stage development and validation cycle will stress-test each subsystem under real-time, high-volume conditions.

Taken together, this work delivers a blueprint for re-architecting industrial web extraction pipelines around autonomous orchestration and modular resilience. The system features a closed-loop control plane that ingests extraction telemetry, detects drift, and coordinates agentic repair; a link-graph-driven discovery engine that surfaces fresh, high-yield content while minimizing crawl depth; and a scalable structural-clustering layer that supports robust, reusable wrapper generalization. The architecture combines programmatic wrappers with on-demand LLM augmentation, matching end-to-end accuracy of fully generative models while minimizing inference overhead and preserving operational transparency. Beyond its immediate deployment at Meltwater scale, the architecture establishes a transferable paradigm for LLM-governed automation across data-intensive workflows, redefining the human role from manual maintainer to strategic supervisor under policy-bound autonomy.

A central research challenge is defining rigorous, component-level evaluation criteria across the full architecture. Each pillar is treated as an independent module with task-specific metrics: precision@k and coverage for URL discovery; purity, stability, and compression ratio for structural clustering; and accuracy, latency, and recovery precision for wrapper induction. These metrics will be validated in the editorial news domain, where drift and scale create meaningful pressure on system performance. Although pipeline-wide comparison is inherently complex, the architecture's modularity enables layered benchmarks and ablation studies. Future work includes defining normalized, cost-aware composite metrics to compare hybrid, LLM-heavy, and rule-based extraction strategies under real-world constraints.

Status of PhD: Currently in the early stages of doctoral research, with the dissertation framework under active development. The direction of the work is highly adaptable, and participation in the VLDB PhD Workshop is expected to provide valuable feedback that will inform system design choices, evaluation methodology, and integration strategies.

Keywords: Intelligent Agents; Web Data Extraction; Wrapper Induction; Template Clustering; Large Language Models; URL Discovery; Automation; Semantic Labeling; Control Plane

ACKNOWLEDGMENTS

This research is supported by Meltwater Deutschland GmbH.

REFERENCES

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130

[2] Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes. *Proc. VLDB Endow.* 17, 2 (Oct. 2023), 92–105. https://doi.org/10.14778/3626292.3626294

[3] Robert Baumgartner, Sergio Flesca, and Georg Gottlob. 2001. Visual Web Information Extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*. VLDB Endowment, Roma, Italy, 119–128. http://www.lixto.com, methods covered by a pending patent.

[4] Mirko Bronzi, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. 2013. Extraction and Integration of Partially Overlapping Web Sources. *Proceedings of the VLDB Endowment* 6, 10 (2013), 805–816. Invited to present at the 39th International Conference on Very Large Data Bases, August 26–30, 2013, Riva del Garda, Trento, Italy.

[5] Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang. 2008. iRobot: an intelligent crawler for web forums. In *Proceedings of the 17th International Conference on World Wide Web* (Beijing, China) *(WWW '08)*. Association for Computing Machinery, New York, NY, USA, 447–456. https://doi.org/10.1145/1367497.1367558

[6] Valerio Cetorelli, Paolo Atzeni, Valter Crescenzi, and Franco Milicchio. 2021. The smallest extraction problem. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2445–2458.

[7] Luying Chen, Stefano Ortona, Giorgio Orsi, and Michael Benedikt. 2013. Aggregating Semantic Annotators. *Proceedings of the VLDB Endowment* 6, 13 (2013), 1690–1701. Invited to present at the 39th International Conference on Very Large Data Bases, August 26–30, 2013, Riva del Garda, Trento, Italy.

[8] Zui Chen, Lei Cao, Sam Madden, Tim Kraska, Zeyuan Shang, Ju Fan, Nan Tang, Zihui Gu, Chunwei Liu, and Michael Cafarella. 2024. SEED: Domain-Specific Data Curation With Large Language Models. *arXiv preprint arXiv:2310.00749v3* (2024). https://arxiv.org/abs/2310.00749v3 Affiliations: MIT, University of Arizona, HKUST (GZ), Renmin University.

[9] Burton-Taylor International Consulting. 2022. *Media Intelligence Company Focus – Meltwater Strategic Review*. Technical Report. https://tpicap.com/burtontaylor/reports/10/2022/media-intelligence-company-focus-meltwater-strategic-review

[10] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. 2001. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*. VLDB Endowment, Roma, Italy, 109–118.

[11] Valter Crescenzi, Paolo Merialdo, and Disheng Qiu. 2013. A Framework for Learning Web Wrappers from the Crowd. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*. 261–272. https://doi.org/10.1145/2488388.2488412

[12] Valter Crescenzi, Paolo Merialdo, and Disheng Qiu. 2019. Hybrid Crowd–Machine Wrapper Inference. *ACM Transactions on Knowledge Discovery from Data* 13, 5 (2019), 51:1–51:43. https://doi.org/10.1145/3344720

[13] CrewAI Inc. 2025. CrewAI: Fast and Flexible Multi-Agent Automation Framework. https://github.com/crewAIInc/crewAI. GitHub repository, accessed 2025-05-04.

[14] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. 2024. Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 19783–19812. https://proceedings.neurips.cc/paper_files/paper/2024/file/2318d75a06437eaa257737a5cf3ab83c-Paper-Conference.pdf

[15] Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart, and Cheng Wang. 2014. DIADEM: Thousands of Websites to a Single Database. *Proceedings of the VLDB Endowment* 7, 14 (2014), 1845–1856. Invited to present at the 40th International Conference on Very Large Data Bases, September 1–5, 2014, Hangzhou, China.

[16] Tim Furche, Georg Gottlob, Giovanni Grasso, Giorgio Orsi, Christian Schallhart, and Cheng Wang. 2011. Little Knowledge Rules the Web: Domain-Centric Result Page Extraction. In *Web Reasoning and Rule Systems*, Sebastian Rudolph and Claudio Gutierrez (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 61–76.

[17] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2015. iCrawl: Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (Knoxville, Tennessee, USA) *(JCDL '15)*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/2756406.2756925

[18] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 505–525. https://doi.org/10.18653/v1/2022.emnlp-main.33

[19] Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. 2024. Language Models as Hierarchy Encoders. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 14690–14711. https://proceedings.neurips.cc/paper_files/paper/2024/file/1a970a3e62ac31c76ec3cea3a9f68fdf-Paper-Conference.pdf

[20] Jingtian Jiang, Nenghai Yu, and Chin-Yew Lin. 2012. FoCUS: learning to crawl web forums. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2187980.2187985

[21] Kadoa. 2024. Crawling Documentation. https://docs.kadoa.com/docs/crawling. Accessed: 2025-05-07.

[22] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection Using Shallow Text Features. *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 441–450. https://doi.org/10.1145/1718487.1718542

[23] Alberto H.F̃. Laender, Berthier A. Ribeiro-Neto, Altigran S. da Silva, and Juliana S. Teixeira. 2002. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record* 31, 2 (2002), 84–93. https://doi.org/10.1145/565117.565137

[24] Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. 2024. LLM-R2: A Large Language Model Enhanced Rule-based Rewrite System for Boosting Query Efficiency. *Proc. VLDB Endow.* 18, 1 (2024), 53–65. https://www.vldb.org/pvldb/vol18/p53-yuan.pdf

[25] Samuel Madden, Michael J. Cafarella, Michael J. Franklin, and Tim Kraska. 2024. Databases Unbound: Querying All of the World's Bytes with AI. *Proceedings of the VLDB Endowment* 17, 12 (2024), 4546–4554. https://doi.org/10.14778/3685800.3685916 Affiliations: MIT CSAIL (Madden, Cafarella, Kraska); University of Chicago (Franklin).

[26] Meltwater. 2025. Meltwater: Media Intelligence and Social Listening. https://www.meltwater.com Accessed: 2025-05-20.

[27] Wai-kit Ming, Fengqiu Huang, Qiuyi Chen, Beiting Liang, Aoao Jiao, Taoran Liu, Huailiang Wu, Babatunde Akinwunmi, Jia Li, Guan Liu, Casper J. P. Zhang, Jian Huang, and Qian Liu. 2021. Understanding Health Communication Through Google Trends and News Coverage for COVID-19: Multinational Study in Eight Countries. *JMIR Public Health and Surveillance* 7, 12 (2021), e26644. https://doi.org/10.2196/26644

[28] NATO Strategic Communications Centre of Excellence. 2022. *Social Media Monitoring Tools – An In-Depth Look*. Technical Report. https://stratcomcoe.org/publications/download/Social-Media-Monitoring-Tools-DIGITAL.pdf

[29] Liana Patel, Siddharth Jha, Melissa Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia. 2025. Semantic Operators: A Declarative Model for Rich, AI-based Data Processing. *arXiv preprint arXiv:2407.11418* (2025). https://doi.org/10.48550/arXiv.2407.11418 Affiliations: Stanford University (Patel, Gupta, Guestrin); UC Berkeley (Jha, Pan, Asawa, Zaharia).

[30] Sudhir Kumar Patnaik and C. Narendra Babu. 2021. Trends in web data extraction using machine learning. *Web Intelligence* 19, 3 (2021), 169–190. https://doi.org/10.3233/WEB-210465 arXiv:https://doi.org/10.3233/WEB-210465

[31] Ralph Peeters, Aaron Steiner, and Christian Bizer. 2025. Entity Matching Using Large Language Models. In *Proceedings of the 28th International Conference on Extending Database Technology (EDBT 2025), Barcelona, Spain, March 25–28*, Vol. 2. OpenProceedings.org, Konstanz, 529–541. Experiments & Analyses Track.

[32] Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. 2024. Agent-OM: Leveraging LLM Agents for Ontology Matching. *Proc. VLDB Endow.* 18, 3 (Nov. 2024), 516–529. https://doi.org/10.14778/3712221.3712222

[33] ScrapeGraphAI. 2024. ScrapeGraphAI. https://github.com/ScrapeGraphAI/ScrapeGraphAI. Accessed: 2025-05-07.

[34] Yujie Shen, Huashu Yang, Mingdong Zeng, Xun Chen, Yeyun Duan, Zhe Yu, Rui Wang, Weiyue Xiong, Jianfeng Gao, and Jing Li. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

[35] Julia Vassey, Ho-Chun Herbert Chang, Tom Valente, and Jennifer B. Unger. 2024. Worldwide connections of influencers who promote e-cigarettes on Instagram and TikTok: A social network analysis. *Computers in Human Behavior* 165 (2024), 108545. https://doi.org/10.1016/j.chb.2024.108545

[36] Jiayi Wang and Guoliang Li. 2025. AOP: Automated and Interactive LLM Pipeline Orchestration for Answering Complex Queries. In *Proceedings of the 15th Conference on Innovative Data Systems Research (CIDR '25)*. Amsterdam, The Netherlands. https://mail.vldb.org/cidrdb/papers/2025/p32-wang.pdf

[37] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. https://doi.org/10.48550/arXiv.2308.08155 arXiv:2308.08155 [cs.CL]