[redacted] (Abstract)

AlphaClean [redacted] (pipeline) [redacted] “operators” [redacted] (Quality Function) [redacted] dataset [redacted]

[redacted] (Introduction)

[redacted] Data Scientist [redacted] 60–80% [redacted] AlphaClean [redacted] operator [redacted] pipeline [redacted]

[redacted] (Methods)

AlphaClean [redacted]: 1) **Operators Library** — [redacted] remove duplicates, fill missing, normalize text 2) **Pipeline Space Generator** — [redacted] pipeline [redacted] operator [redacted] 3) **Quality Function (QF)** — [redacted] % missing < 1%, duplicates < 0.5% 4) **Search Algorithm** — [redacted] heuristic search [redacted] pipeline [redacted]

[redacted] (Results)

AlphaClean [redacted] dataset [redacted] manual cleaning [redacted] rule-based cleaning [redacted]: - [redacted] clean [redacted] - [redacted] - Pipeline [redacted] dataset [redacted]

[redacted] (Discussion)

[redacted]: - [redacted] operator [redacted] - QF [redacted] - Pipeline reuse [redacted] - [redacted] manual coding **[redacted]:** -

■■■ QF ■■■■■■■■■■ ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ - ■■■■■■ operator library ■■■■■■■■ - Dataset ■■■■■■■■■■■■■ **■■■■■■■■■■■■■■■■■: ■■■■■■■■■■■■■■■■ dataset ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■■■■ ■■■■■■■■■■■■■■■■■■ clean ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ Master Schema**

■■■■■■ (Conclusion)

AlphaClean ■■■■■■■■■■■■■■ ■■■■■■■■ data cleaning pipeline ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■ pipeline reuse ■■■ ■■■■■■■■■■■■■■■■■■■■■ operator ■■■■ ■ (■■■■ OCR, ■■■■■■■) ■■■■■■■■■■■■ ML/LLM ■■■■■■■■■■■■■■■■■■■■■■■