Data Cleaning & Dataset Providers for LLM Training- June 2025

The table below summarizes companies that offer data cleaning services and their involvement in providing or facilitating datasets for training Large Language Models (LLMs).

Company	Provides Data Cleaning	Provides or Curates Datasets for LLMs	Notes
Scale AI	✅ Yes	✅ Yes	Specializes in high-quality labeled datasets for AI. Offers pre-labeled datasets and has worked with OpenAI, Meta, and others.
Snorkel AI	✅ Yes	⚠️ Indirectly	Offers tools for programmatic data labeling and weak supervision. Helps teams generate labeled datasets in-house.
Labelbox	✅ Yes	⚠️ Indirectly	Provides a data-labeling platform. Often used in partnership with in-house or third-party datasets.
AWS SageMaker Ground Truth	✅ Yes	✅ Yes (with AWS Data Exchange)	Provides human-in-the-loop labeling and access to third-party datasets through AWS Data Exchange.
CloudFactory	✅ Yes	⚠️ Custom Only	Specializes in human-powered labeling. Creates custom datasets but does not sell pre-made datasets.
Databricks (Delta Lake)	✅ Yes	⚠️ Indirectly	Focuses on managing and cleaning large data lakes. Not a dataset vendor but supports dataset preparation.

Scale AI and AWS are the **most prominent** in both **data cleaning and dataset provisioning** for LLM training.

Snorkel AI and Labelbox empower organizations to **label their own data** using platforms, not pre-curated datasets.

Databricks and CloudFactory are more infrastructure- and services-focused, enabling custom dataset development.