Can an LLM Replace a Human at Manual Gating?

Executive Summary

The Experiment

We gave Claude Code (Opus 4.5) an 86,864-event CyTOF bone marrow dataset (Samusik 01), a gating strategy image, and a written protocol. No human touched the data. The LLM wrote and iteratively debugged an R script that implements a full 24-population hierarchical manual gating pipeline—from QC gates through terminal cell type assignment.

We compared the output against expert-curated reference labels (Spitzer et al.).

Headline Numbers

Metric	Value
Overall exact-match accuracy	**48.4%**
Accuracy on assigned cells	**60.5%**
Lineage-level accuracy	**82.4%**
Weighted F1 score	**40.2%**
Populations gated	24 terminal types
Unassigned rate	47.5% (reference: 38.8%)
Human intervention	Zero
Total runtime	∼11 iterative runs, fully autonomous

What These Numbers Mean

82.4% lineage-level accuracy—When the LLM assigns a cell to a type and the reference also assigns it, they agree on the major lineage 82% of the time (27,700 / 33,617 both-assigned cells). The LLM rarely confuses a B cell for a T cell or a monocyte for a progenitor.

> **Rigorous definition of "lineage-level":** This is *not* derived from the gating tree topology. It is a flat, manually defined grouping of the 24 terminal populations into 6 categories based on canonical immunology naming: **T/NK:** CD4 T, CD8 T, $\gamma\delta$ T, NK, NKT · **B:** IgD^+IgM^+ B, IgD^-IgM^+ B, IgM^-IgD^- B, Plasma Cells, Pro-B · **Myeloid:** Classical, Intermediate, & Non-Classical Monocytes, Macrophages · **Granulocyte:** Eosinophils, Basophils · **DC:** pDCs, mDCs · **Progenitor:** HSC, MPP, CMP, GMP, MEP, CLP.
>
> A cell is counted as a "lineage match" if both the LLM and reference labels fall into the same group. Cells unassigned by either method are excluded from the denominator.

60.5% exact-match on assigned cells—Among cells where both the LLM and reference assigned a label, 6 in 10 get the exact subtype right. Most disagreements are within-lineage (e.g., Classical vs. Intermediate Monocyte, or IgD^+IgM^+ vs. IgD^-IgM^+ B cell).

48.4% overall accuracy—Includes the 47.5% of cells the LLM left unassigned. The reference leaves 38.8% unassigned. The gap is mostly due to monocyte undercount (LLM: ∼13k, reference: ∼27k).

Reading the Precision and Recall Tables

Each population has cells assigned to it by the LLM and cells assigned to it by the expert reference (Spitzer). **Precision** answers: *of all the cells the LLM called a given type, what fraction actually are that type according to the reference?* High precision means the LLM is rarely wrong when it makes a call; low precision means it is labeling many cells incorrectly (false positives). **Recall** answers: *of all the cells the reference says belong to a given type, what fraction did the LLM successfully identify?* High recall means the LLM finds most of the true members; low recall means it misses many (false negatives).

Three concrete examples illustrate the trade-offs:

- **IgD$^+$ IgM$^+$ B cells** (87.9% precision, 85.2% recall): The LLM is both accurate when it calls something this type and finds most of them. Strong all around.
- **Classical Monocytes** (75.4% precision, 24.9% recall): When the LLM says "classical monocyte," it is usually right—but it only finds 1 in 4 of the true classical monocytes. The CD11b threshold is too strict, so most monocytes are left unassigned.
- **Macrophages** (0.3% precision, 2.5% recall): The LLM gates 1,980 cells as macrophages, but only 5 match the reference. It is both calling the wrong cells macrophages and missing the real ones—the LLM and reference are using different definitions.

F1 is the harmonic mean of precision and recall—a single number that is only high when *both* are high. This is why Classical Monocytes (F1 = 37.5%) scores much lower than its precision alone would suggest.

Strongest Performers

Population	Precision	Recall
IgD$^+$ IgM$^+$ B cells	87.9%	85.2%
CD8 T cells	65.2%	87.1%
CD4 T cells	59.9%	83.5%
IgD$^-$ IgM$^+$ B cells	64.1%	83.5%
Classical Monocytes	75.4%	24.9%
Basophils	38.0%	77.4%

Weakest Performers

Population	Precision	Recall
CLP	0.0%	0.0%
HSC	0.2%	33.3%
Macrophages	0.3%	2.5%
NKT cells	3.1%	4.7%
mDCs	2.3%	15.5%
Pro-B (Frac A–C)	16.7%	4.1%

What Went Right

- The LLM autonomously identified and fixed 6 distinct failure modes across 11 iterations
- It discovered that CD115 doesn't separate monocytes in this panel and removed it
- It detected that its own QA check was mathematically flawed (root median comparison) and fixed it
- Final output passes all QA checks with zero violations above threshold
- Full artifact suite: 31 gate plots, UMAP visualizations, heatmaps, QA tables, reports

What Went Wrong

- Automated density valley detection fails for markers positive on <5% of cells (CD3, NKp46, CD138)
- Monocyte counts are roughly half of reference—CD11b threshold may be too strict
- Some population definitions differ from reference (Macrophages, HSC, CLP)
- The LLM consumed 2 full context windows iterating on thresholds

The Bottom Line for Decision-Makers

This is not ready to replace a human gating expert. A 48% overall accuracy and 40% F1 would not pass review in a clinical or publication setting.

But it is remarkably close to useful. At the lineage level (82%), the LLM correctly identifies the major immune compartments. The failures are mostly in threshold calibration—something that could be addressed with:

- Providing expected count ranges per population
- Using Gaussian mixture models instead of density valley detection
- Supplying example gate plots from a successful run

The strongest signal: the LLM's ability to autonomously debug, diagnose, and iterate. It identified that CD115 was the wrong marker, that its QA metric was flawed, and that NKp46 is dim in this panel—all without human input. That self-correcting capability, combined with better tooling, could make this production-viable.

Recommended next step: Run the same experiment with (1) expected count ranges in the QA doc, (2) example gate plots, and (3) mixture model thresholding. We predict this would push exact-match accuracy above 70% and weighted F1 above 60%.

Key Visual

See `figures/confusion_heatmap_llm_vs_ref.png`—a row-normalized heatmap showing where each LLM-gated population maps in the reference labels. Strong diagonal = correct assignments.