Semantic Drift: A Hidden Failure Mode in LLMs

This short working note pulls together emerging discussion on **semantic drift**. Drift occurs when outputs remain factually correct, but the original *meaning, nuance, or purpose* erodes. The surface looks fine, yet the depth collapses.

Examples of Semantic Drift:

- "Cogito, ergo sum" (Descartes) → reframed as leadership advice about confidence.
- "The map is not the territory" → becomes "Having a plan is just as important as execution."
- A Berlin Wall history excerpt → reframed as a business lesson in change management.

Key Points:

- **Not hallucination:** Facts remain intact.
- **Not simple bias:** It is not skew, but hollowing out.
- **Benchmarks miss it:** Accuracy metrics mark these as 'correct.'
- **Recursive risk:** Drift compounds over generations and accelerates collapse.
- **Semantic fidelity:** What matters is whether *purpose* survives rephrasings, translations, or reframings.

Recent Work Noticing Drift:

- Reddit study on recursive generations: 6.6x worse semantic drift vs factual decay.
- Nature Scientific Reports (2025): Drift highlighted as an evaluation shortcoming.
- Makoy (Medium, 2025): Drift as part of model collapse.
- Sem-DPO, RiOT (arXiv, 2025): Explicitly attempt to mitigate drift.

> Hallucinations break facts. **Drift breaks meaning.**