

OPEN**SEN-11-088**

Atlas Engine: Refusal as a Feature

DATED	August 23, 2011	TYPE	Research brief
ORIGIN	Directorate II	SITE	--
PAGES	9	CUSTODY	MERIDIAN

The design decision that defines our analysis platform: it will decline to answer rather than answer without a citable source. Includes the eighteen months during which this made it useless, and why we kept it.

THE DECISION

Atlas Engine will not produce an answer it cannot attribute.

Where the institutional record does not support a response, Atlas says so and stops. It does not interpolate, it does not generalise from adjacent material, and it does not produce a fluent paragraph in the absence of a source.

This is the single most consequential design decision in the system, and it was close to being reversed twice.

WHY IT NEARLY FAILED

For roughly eighteen months after first deployment, Atlas was, by the fair assessment of its own users, worse than useless.

It refused constantly. Thirty years of institutional records had been kept to inconsistent standards; large portions could not be cited in a form Atlas would accept, so it declined to draw on them. Researchers who asked reasonable questions received, repeatedly, a statement that the record did not support an answer.

Several proposed a "best effort" mode -- clearly labelled, caveated, off by default. The argument was straightforward: a marked-uncertain answer is more useful than nothing, and adults can be trusted with a caveat.

WHY IT WAS KEPT

The Directorate ran the experiment before deciding. A best-effort mode was built and tested with fourteen researchers over nine weeks.

The finding was that the caveat did not survive contact with the workflow. Answers were correctly labelled uncertain at the point of generation, and were then pasted into working notes, quoted in meetings and cited in draft briefs with the label stripped -- not through carelessness, but because a fluent paragraph is portable and a caveat attached to it is not.

Within nine weeks, two uncertain Atlas outputs had made their way into documents that presented them as established. Both were caught. Neither was caught by the person who had moved them.

The mode was removed and has not been rebuilt.

WHAT WAS DONE INSTEAD

The refusals were treated as a measurement rather than a defect. Every refusal was logged with the query that caused it, producing -- for the first time -- a precise map of where the institutional record was too weak to answer questions people actually had.

That map drove four years of retrospective cataloguing. Refusals fell by 71% between 2013 and 2017, not because Atlas became more permissive, but because the record improved.

This is the part of the brief we consider most worth publishing. The refusals were not the system failing to do its job. They were the system reporting a problem that predated it, and that nobody had been able to see.

PRESENT BEHAVIOUR

Atlas today will:

- answer with provenance, or decline;
- surface contradictions in the record rather than silently choosing between them, including where the contradiction is embarrassing;
- report the age of the material it drew on, because a correct answer from a 1987 document may still be the wrong answer today; and
- operate on the Archive of record rather than a copy, so that a withdrawn or amended document is immediately reflected.

It remains internal. It is not offered as a product and there are no plans to offer it.

End of record SEN-11-088. Reproduced from the Archive of record at MERIDIAN. Handling markings are internal publication conventions of Sentinel Research Group and correspond to no external scheme.