

Long, wide, and trivial:

a nondeterministic time hierarchy in Isabelle

András Z. Salamon and Michael Wehar



University of
St Andrews

AITP 2026, Aussois, 2 September 2026

Goal and progress

Nondeterministic Time Hierarchy Theorem (NDTHT): for well-behaved $t_1(n+1) \in o(t_2(n))$

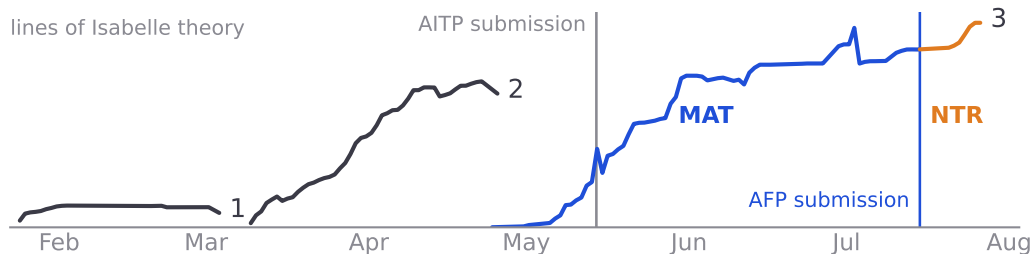
$$\text{NTIME}(t_1(n)) \subsetneq \text{NTIME}(t_2(n))$$

- progress: components, not yet the theorem
- Archive of Formal Proofs: MAT including linear speedup theorem: enlarge alphabet, run faster
- also proved: nondeterministic tape reduction (NTR): guess, then verify tape by tape

Barbarous practices

- (was) not an ITP person; nor is my co-author
- Claude Code in terminal, one Claude Max subscription
- no IDE: `isabelle build` as a **proof oracle**
- no Sledgehammer: the LLM chooses every premise
- our choices required new tools to compensate

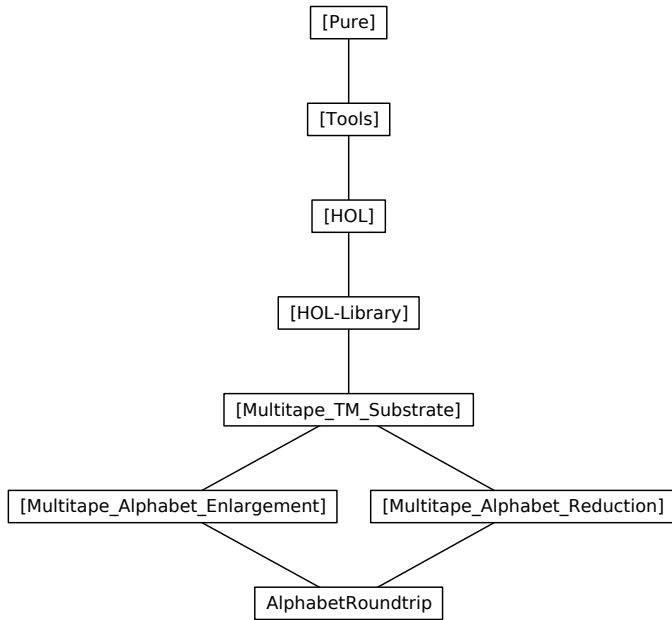
Progress



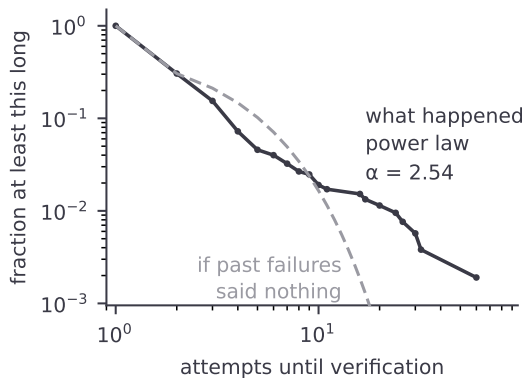
	when	outcome
1	Jan–Mar	lower bound proved; upper stalled
2	Mar–Apr	42,612 lines, abandoned
3	Apr–Jul	56,555 lines, 4 AFP entries (MAT)
	Jul–	8,622 lines, NTR proved



Multitape.TM.Substrate



Tool 1: taming semi-decidability



isabelle-build

- enforce timeouts better
- record proof trajectories

525 proof trajectories:
70% right first time,
but up to 60 attempts

```
pip install isabelle-watchdog
```



isabelle-watchdog

Tool 2: measuring the claim in our talk title

isabelle-query: syntax-aware interaction with Isar projects

we conjectured our proofs are **long, wide, trivial**:

long chains over *wide* configurations but *mechanical* steps

referees, fairly: “*no formal metric for width is given*”

957 AFP (tag Isabelle2025-2) entries, 293,893 proofs

918 have multiline proofs: our comparison class

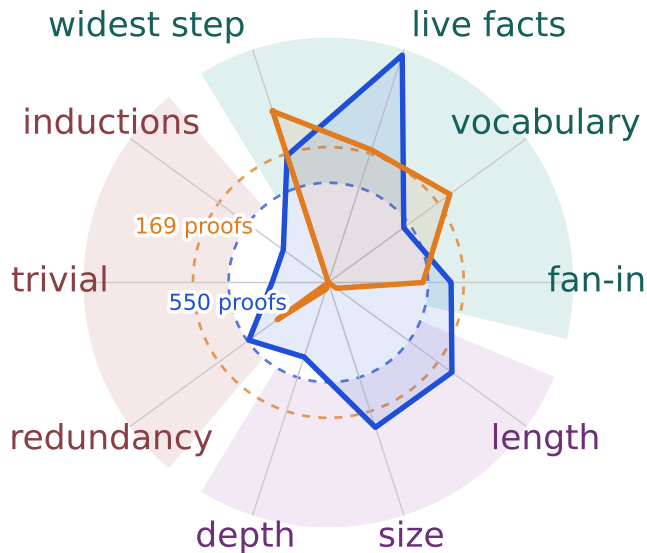
query shape census with 10 shape axes

```
pip install isabelle-query
```



isabelle-query

Long, wide, ~~trivial~~



further from the AFP
than chance:

long

wide

~~trivial~~

MAT is long and wide

NTR is wide

dashed: crossed by
chance 1 time in 20

At what cost?

Claude Max, 8 months, all projects	£720
tokens at API list price, this project	\$11,500

inference, this project	150–400 kWh
Isabelle builds on this laptop	≈ 1 kWh

UK non-domestic electricity price	£0.30/kWh
-----------------------------------	-----------

```
pip install claude-session-diagnostics
```



claude-session-diagnostics

Parting words: take part

- measure your own corpus: `query shape`
- record the proof search: `isabelle-build`
- **MAT** is in the AFP, yours to import

Andras.Salamon@st-andrews.ac.uk

“have you considered skills?”

referee 1 asked if agent guidance could be scripted as “skills”

problem: things to remember get remembered late

the knowledge base (297 insights, 108 techniques): too big to hold in mind, so it is read *after* the session that needed it

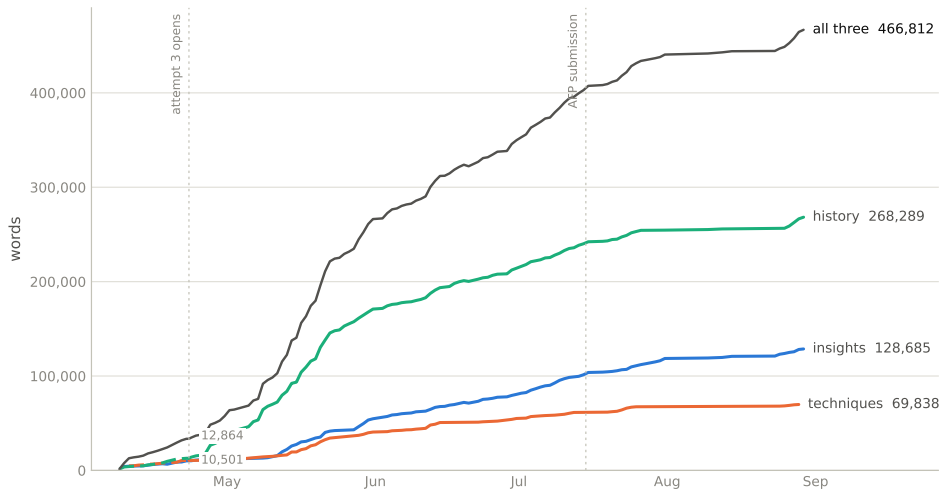
our skills not that useful, except `/handoff` and this would be better as automated trigger at end of LLM output

what works is *tools*: `make build calls isabelle-build calls the recorder`, so every trajectory is recorded by default

Growth of the knowledge base

Knowledge-base growth: what the project recorded as it proved

words per corpus, counted across storage migrations; sampled at 12:00 daily, days with no change omitted.
Dashed = the attempt-2 corpus its successor restarted from; that size is the base, not a discontinuity.



de Bruijn factors

“de Bruijn factor” is the ratio of size of formal to informal proof

informal source	formal pages	lines of Isar
2 textbook pages [Hopcroft–Ullman 1979]	732	36,122
8 textbook lines [Reischuk 1999]	318	15,336
1 paragraph sketch in a thesis [Li 1985]	197	8,622
not in the literature (yet)	21	653

The same shape, in numbers

p: how often 2,000 random AFP slices of our size went as far

MAT: four of ten axes survive Bonferroni

live facts 14.9, length 5.3, size 5.3, fan-in 3.8, **$p \leq 0.0015$**

carry many facts at once: **one** of 918 entries above; archive median **zero**; 7,091 using clauses over 10,742 Isar steps

NTR: widest step, vocabulary **$p < 0.01$**
(Benjamini–Hochberg only)

closers: simp 49%, rule 16%, auto 10%, cases 8%, blast 8%, metis 1%;
and no fact database: 49 [simp] lemmas

found by measuring, not by introspection

Nondeterministic tape reduction

every k -tape nondeterministic machine is simulated by one with **three** tapes

$k_{tm}(\text{tape_reduce } M) = 3$

and if M accepts w within $T(|w|)$ steps, $\text{tape_reduce } M$ accepts within $c \cdot T(|w|) + d$

constants explicit; [Book–Greibach–Wegbreit 1970]
construction never analysed in textbooks

HOAU

Efficient high-order anti-unification, with Mateu Villaret

“least general generalisation”

Isabelle/ML code for access to AST

fast: few seconds on large theory

finds useful refactoring opportunities