From prompt to paper

A quick look at arXiv:2603.20179 [hep-ex]

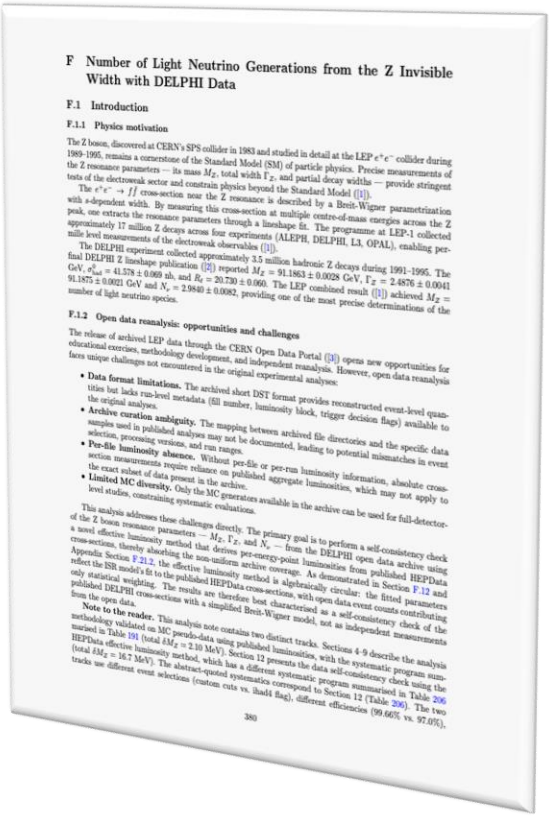
Dietrich Liko

The claim

“Large language model-based AI agents are now able to autonomously execute substantial portions of a high energy physics (HEP) analysis pipeline with minimal expert-curated input.”

Measure the number of light neutrino generations from the Z invisible width

5h15



F Number of Light Neutrino Generations from the Z Invisible Width with DELPHI Data

F.1 Introduction

F.1.1 Physics motivation

The Z boson, discovered at CERN's SPS collider in 1983 and studied in detail at the LEP e^+e^- collider during 1989–1995, remains a cornerstone of the Standard Model (SM) of particle physics. Precise measurements of the Z resonance parameters — its mass M_Z, total width Γ_Z, and partial decay widths — provide stringent tests of the electroweak sector and constrain physics beyond the Standard Model ([1]).

The $e^+e^- \to f\bar{f}$ cross-section near the Z resonance is described by a Breit-Wigner parametrization with s-dependent width. By measuring this cross-section at multiple centre-of-mass energies across the Z peak, one extracts the resonance parameters through a lineshape fit. The programme at LEP-1 collected approximately 17 million Z decays across four experiments (ALEPH, DELPHI, L3, OPAL), enabling per-mille level measurements of the electroweak observables ([1]).

The DELPHI experiment collected approximately 3.5 million hadronic Z decays during 1991–1995. The final DELPHI Z lineshape publication ([2]) reported $M_Z = 91.1863 \pm 0.0028$ GeV, $\Gamma_Z = 2.4876 \pm 0.0041$ GeV, $\sigma^0_{\rm had} = 41.578 \pm 0.069$ nb, and $R_l = 20.730 \pm 0.060$. The LEP combined result ([1]) achieved $M_Z = 91.1875 \pm 0.0021$ GeV and $N_\nu = 2.9840 \pm 0.0082$, providing one of the most precise determinations of the number of light neutrino species.

F.1.2 Open data reanalysis: opportunities and challenges

The release of archived LEP data through the CERN Open Data Portal ([3]) opens new opportunities for educational exercises, methodology development, and independent reanalysis. However, open data reanalysis faces unique challenges not encountered in the original experimental analyses:

- **Data format limitations.** The archived short DST format provides reconstructed event-level quantities but lacks run-level metadata (fill number, luminosity block, trigger decision flags) available to the original analyses.
- **Archive curation ambiguity.** The mapping between archived file directories and the specific data samples used in published analyses may not be documented, leading to potential mismatches in event selection, processing versions, and run ranges.
- **Per-file luminosity absence.** Without per-file or per-run luminosity information, absolute cross-section measurements require reliance on published aggregate luminosities, which may not apply to the exact subset of data present in the archive.
- **Limited MC diversity.** Only the MC generators available in the archive can be used for full-detector-level studies, constraining systematic evaluations.

This analysis addresses these challenges directly. The primary goal is to perform a self-consistency check of the Z boson resonance parameters — M_Z, Γ_Z, and N_ν — from the DELPHI open data archive using a novel effective luminosity method that derives per-energy-point luminosities from published HEPData cross-sections, thereby absorbing the non-uniform archive coverage. As demonstrated in Section F.12 and Appendix Section F.21.2, the effective luminosity method is algebraically circular: the fitted parameters reflect the ISR model's fit to the published HEPData cross-sections, with open data event counts contributing only statistical weighting. The results are therefore best characterised as a self-consistency check of the published DELPHI cross-sections with a simplified Breit-Wigner model, not as independent measurements from the open data.

Note to the reader. This analysis note contains two distinct tracks. Sections 4–9 describe the analysis methodology validated on MC pseudo-data using published luminosities, with the systematic program summarised in Table 191 (total $\delta M_Z = 2.10$ MeV). Section 12 presents the data self-consistency check using the HEPData effective luminosity method, which has a different systematic program summarised in Table 206 (total $\delta M_Z = 16.7$ MeV). The abstract-quoted systematics correspond to Section 12 (Table 206). The two tracks use different event selections (custom cuts vs. ihadd flag), different efficiencies (99.66% vs. 97.0%),

380

DELPHI Experiment Example

How did they do it ?

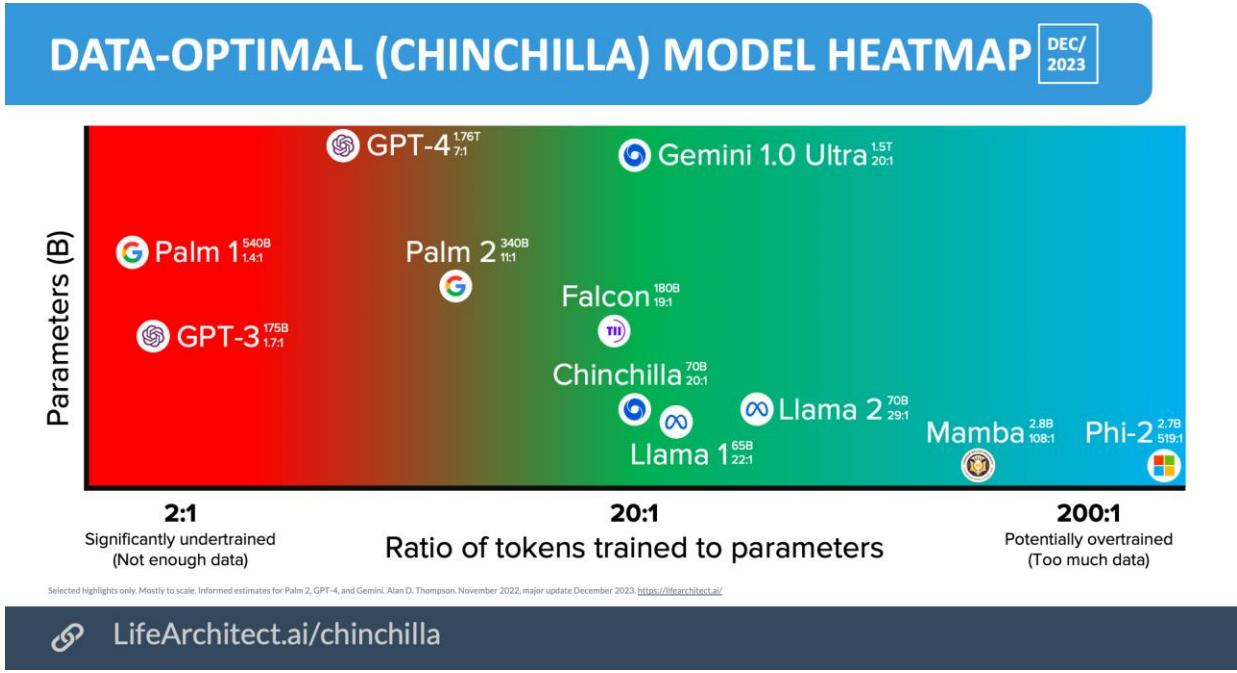
- **Claude Code** — the agentic runtime (not just the model — the CLI that can execute code, read files, iterate)
- **Claude Opus 4** — Anthropic's most advanced model
- **JFC Framework** — the methodology + conventions + orchestration layer
- **SciTreeRAG** — the literature retrieval system
- **CERN Opendata** – CMS/Aleph/DELPHI
- **Pure Python -** uproot/awkward/pyhf/mplhep

Outline

- RAG concepts
  - SciTreeRAG
- Concepts of agentic programming
- The JFC multi-agent framework
  - The agent zoo
- The DELPHI analysis

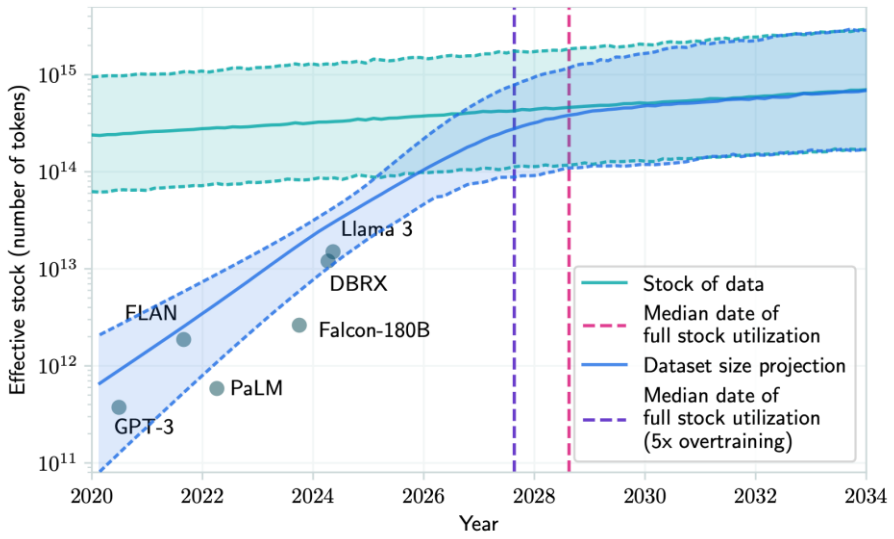
Chincilla Scaling Law (2022)

- For compute-optimal training, a model requires 20 tokens per 1 parameter.
- Modern models (like Llama 3) intentionally break this law by over-training (e.g., 2,000 tokens per parameter) to make smaller models more capable and cheaper to deploy.
- A 1 T parameter model (1TB) trained on 45T tokens (approx. 40TB of text) represents a 40:1 lossy compression of its training data.



Original reference: Hoffmann et al. (2022) | arXiv:2203.15556

In parenthesis: Will we run out of data ?



Villalobos et al. (2022) | arXiv:2211.04325

Data Category	Total Raw Stock (Tokens)	Effective Stock (Filtered/Deduped)	Context / Comparison
Public Web	~100 T	~30 – 50 T	The "messy" internet; requires massive filtering.
All Published Books	~4 – 8 T	~4 T	High quality, but finite and legally complex.
Scientific Papers	~1 – 2 T	~1 T	Dense, high-reasoning data.
Code (GitHub, etc.)	~3 – 6 T	~3 T	Critical for logic and math performance.
Total High-Quality Text	—	~9 – 30 T	The "Data Wall" threshold.
Total Human Text Stock	~300 T	~100 – 200 T	Includes low-quality social media, logs, etc.

What is Retrival Augmented Generation (RAG)

Picture from the TH divison (2000)

Well read, quotes all numbers from his head

Picture from the CERN Library (1963)

Checks all the numbers and concepts in literature

Original publication: *Lewis et al. (2020) | arXiv:2005.11401*

What is Vector Search ?

Chunking: Text is cut into chunks

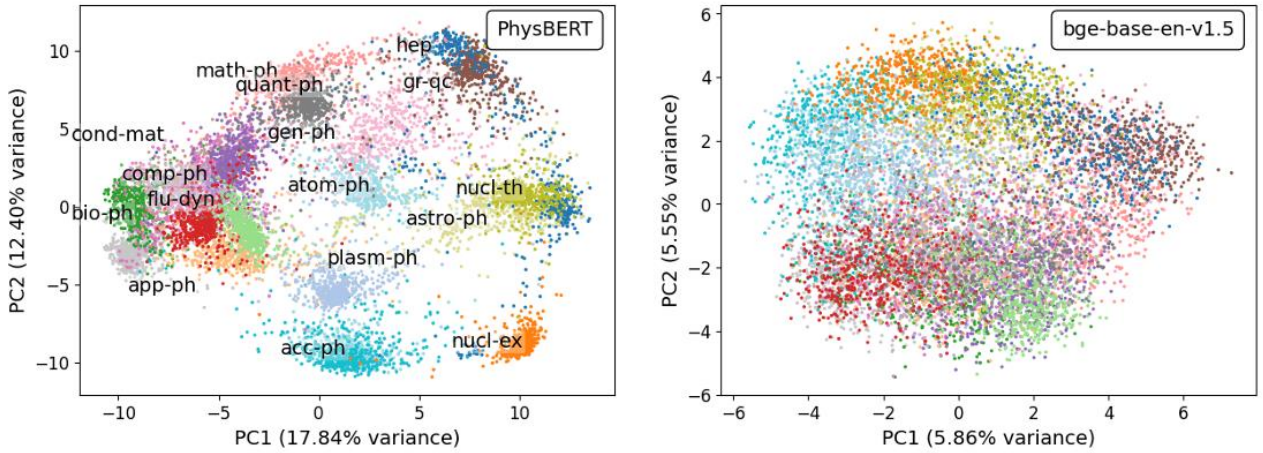
Embedding: An AI model converts the text into a numerical array (vector)

Mapping: Data with similar meaning are close in the vector space

Retrival: The system then finds the nearest neighbour to a question using a distance measure

RAG Challenges for Scientific Text

Phase	Challange	Impact
Chunking	Structural Integrety	Standard breaks paragraph, equations, tables, link to figures
Embedding	Domain Semantics	Domain specific jargon
Graph	Complex Causality	Knowledge Graph complements vector search



Visualisation of the embedding space for PhysBert and more general model

Hellert at al. (2024) | arXiv:2408.09574

SciTreeRag used for the paper

- PhD thesis James C. McGreivy, http://hdl.handle.net/1721.1/164602

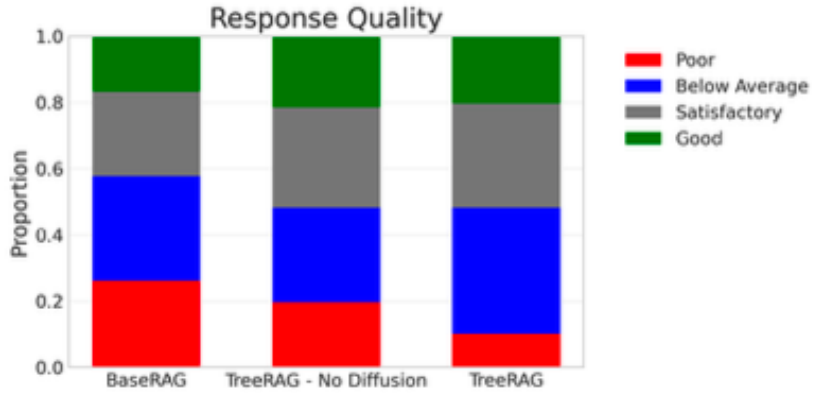
- Papers are considered to be trees by structure.
- It allows to retrieve not only chunks, but the context around

- A study on implementing a Knowledge Graph was less sucessfull

Hu, Y., et al. (2024) | **arXiv:2408.06830**.

Retrieval-Augmented Generation Assessment

- A RAG system has to be tested to see its usefulness – a RAGAS
- Its accuracy has to be a quantifiable number
- Input are questions, context and reference answers
- Scoring done by a LLM (Usually GPT-4 or Claude)



Significant improvement, but hard to tell what effect that has on the overall performance of the system

Additional information available to the agents

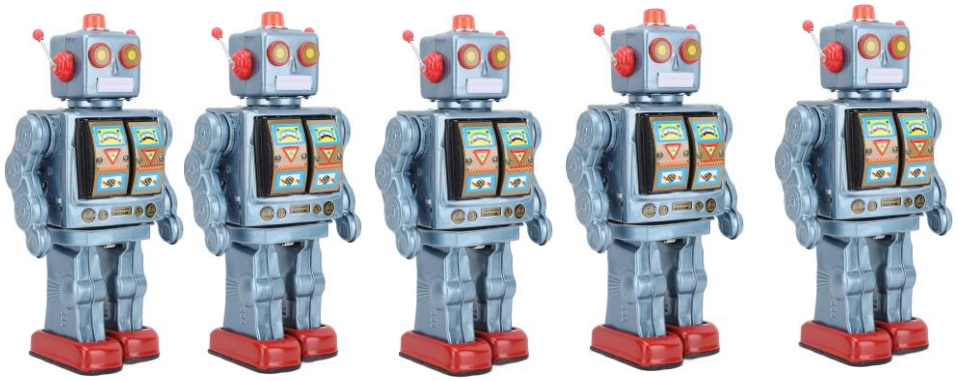
MCP Servers — Anthropic's standard protocol for giving LLMs structured access to external tools and services. Unlike generic web search, each MCP server provides the agent with explicit instructions on how to query its specific API — making retrieval more targeted and efficient.

- arXiv articles
- INSPIRE-HEP
- HEPData

Agentic AI (Tool calling)

- **The Model:** A single "brain" (the LLM).
- **The Mechanism:** The LLM is given a list of tools (functions). It decides which tool to call, executes it, observes the result, and continues.
- **The Limitation:** It still relies on a **single context window**. Every tool output, error message, and observation is appended to the same thread. As the task gets complex, the context becomes "polluted" with technical logs, making the agent more likely to lose track of the original goal.

Multiagentic AI



The core idea: instead of one agent trying to handle everything, the work is split across **specialized agents** — each with its own context.

A controlling agent delegates a subtask to a subagent. The subagent does all the complex work internally and returns only the result. The controlling agent never sees the messy details — just a clean answer it can use to move forward.

- Complexity stays local
- The controlling agent context stays focused
- Each agent uses only tool appropriate for its task

From “Wiring” to “Prompting”

2024: **The Framework Era** (Tools: LlamaIndex, PydanticAI, etc.)

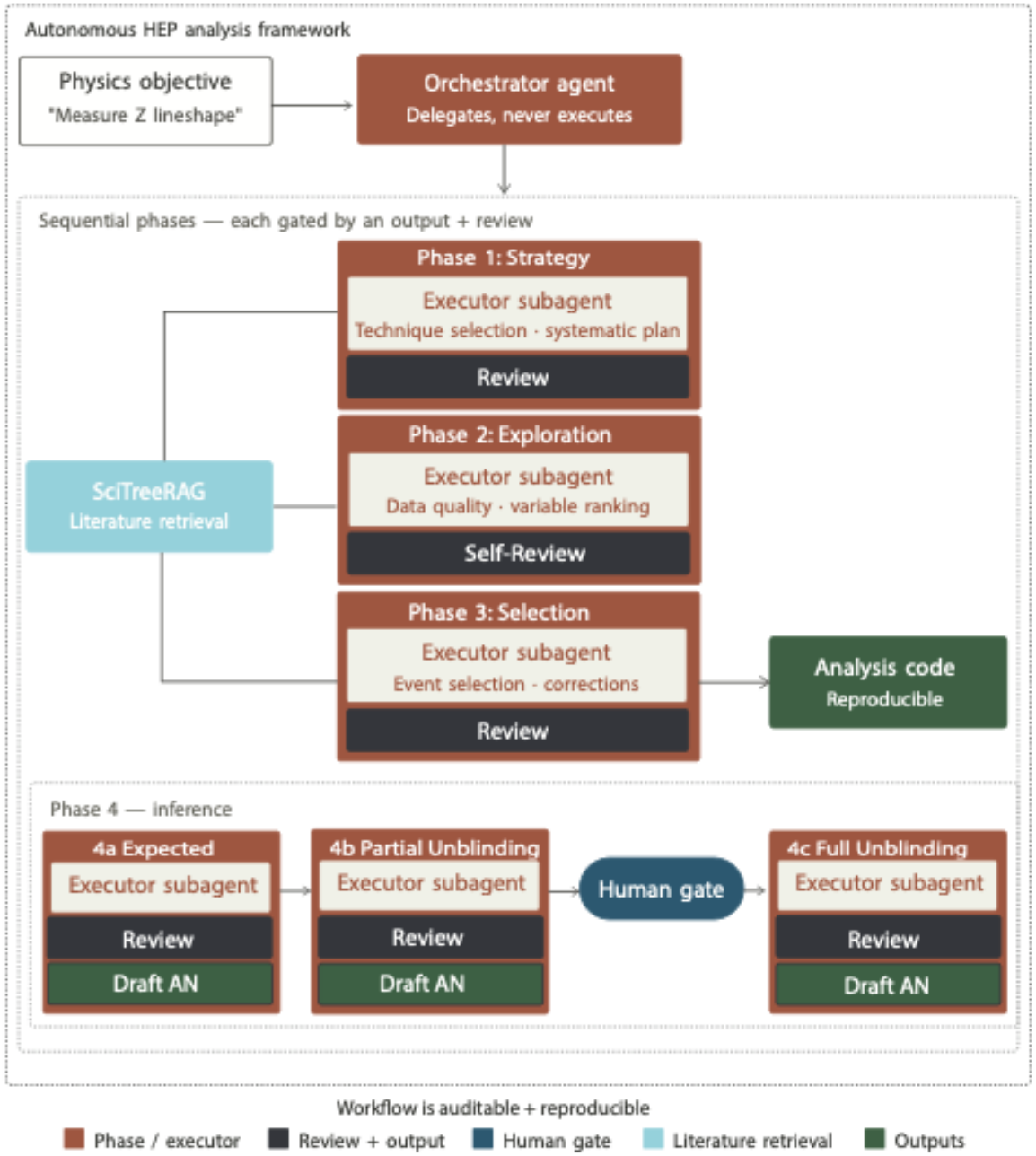
- **Orchestration via Code:**
  Agent interactions are explicitly programmed in Python or similar languages.
- **Predictable & Observable:**
  Allows for detailed logging, strict schemas, and deterministic control flows.

2025/2026: The Natural Language Era

- **Orchestration via Text:**
  Agent behaviors are described in markdown files (e.g., CLAUDE.md, AGENTS.md).
- **Agile but Fluid:**
  Highly ad-hoc and flexible, though potentially less reproducible

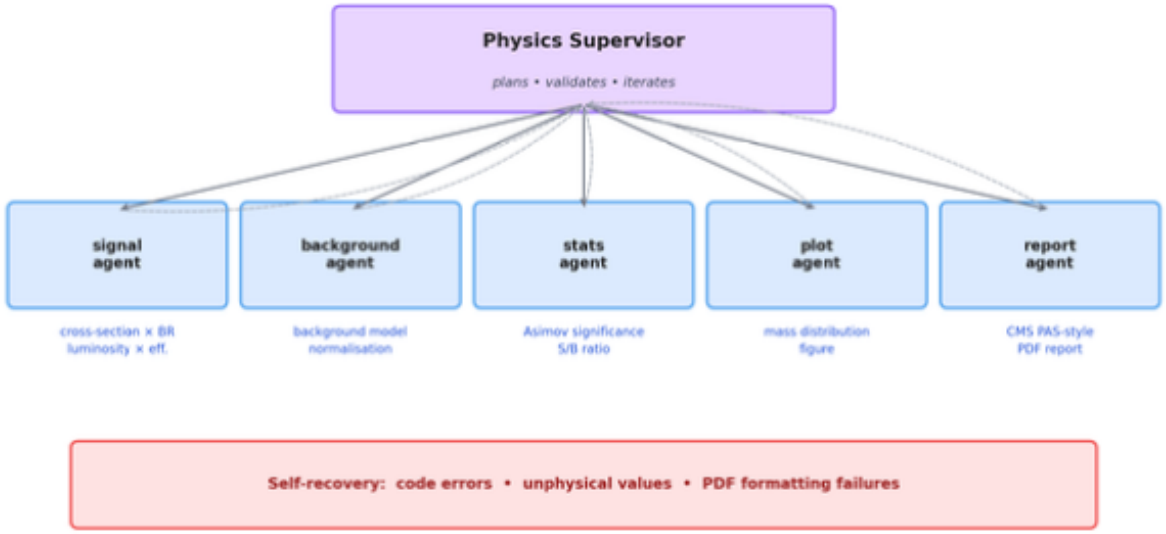
JFC Framework

- **RAG system** gives a strong base from the literature
- **Agent Zoo** organizes a complex pipeline without a fixed sequence.
- After a **human review**, it submits batch jobs for the final analysis



Try it at home: H→γγ in a weekend

- No RAG, just embedded knowledge of the system
- A controlling agent estimates the shape of the signal and the background of a process based on a prompt
- The supervisor controls every step and send work back to the subagent, for example when the distributions are unphysical or S/N outside the expectation



Derive the exclusion limit for 500 GeV Sleptons at the LHC

After 15-20 mins

Exclusion Reach for 500 GeV Sleptons at the LHC

Proof of Concept: Autonomous Physics Analysis via Multiple AI Agents

Date: 30 March 2026

Abstract

We estimate the integrated luminosity required to exclude direct slepton pair production at $m_{\tilde{\ell}} = 500\,\text{GeV}$ at 95% confidence level at the LHC at $\sqrt{s} = 13\,\text{TeV}$. Using the Asimov approximation [6], we find that a luminosity of $\mathcal{L} = 169.7\,\text{fb}^{-1}$ is required to achieve $Z_A = 1.645$ (95% CL exclusion), while $\mathcal{L} => 10000\,\text{fb}^{-1}$ would be needed for a 5σ discovery. These results are derived under simplified-model assumptions with a massless lightest supersymmetric particle (LSP).

Disclaimer: This document is a demonstration of a multi-agent AI workflow and does not constitute a real physics analysis. All physics quantities are approximate estimates derived by AI agents through reasoning from publicly known values. They have not been validated against experimental data or full simulation and should not be interpreted as official results.

Introduction

Supersymmetry (SUSY) predicts the existence of scalar partners to the Standard Model leptons, known as sleptons ($\tilde{\ell}$). Direct slepton pair production, $pp \to \tilde{\ell}^+\tilde{\ell}^-$, proceeds via electroweak Drell-Yan diagrams and produces a clean signature of two opposite-sign same-flavour (OSSF) leptons and large missing transverse energy ($E_{\text{T}}^{\text{miss}}$) from the undetected LSPs [2][3].

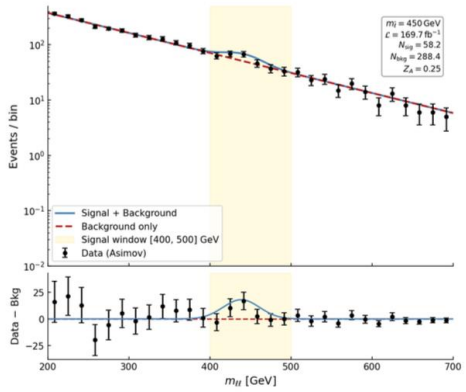
At $\sqrt{s} = 13\,\text{TeV}$, the NLO+NLL production cross-section for a $500\,\text{GeV}$ slepton is $\sigma \approx 1.17\,\text{fb}$ (combined left- and right-handed selectrons and smuons) [1]. This small cross-section, combined with a non-trivial background from WW, $t\bar{t}$, and Drell-Yan processes in the high-mass dilepton region, makes the full Run-2 dataset ($\approx 139\,\text{fb}^{-1}$) and beyond necessary for sensitivity.

Multi-Agent Workflow

This analysis note was produced autonomously by a multi-agent AI framework running under Claude Code with Sonnet 4.6 as the underlying language model. A Physics Supervisor agent received the analysis goal as a natural-language prompt and orchestrated a sequential pipeline of five specialist agents without any manual intervention:

- **Signal agent** — estimated the expected signal yield by reasoning from known cross-sections, branching ratios, detector acceptance, and selection efficiencies.
- **Background agent** — modelled the continuum background using a physically motivated functional form and normalised it from known background cross-sections.
- **Statistics agent** — computed the expected significance using the Asimov approximation and the simple $S/\sqrt{B}$ estimator for comparison.
- **Plot agent** — produced the mass distribution figure in a CMS-inspired publication style.
- **Report agent** — assembled all results into the present document.

No physics parameter was pre-specified by a human: every cross-section, efficiency, yield, and shape parameter was derived by the specialist agents through explicit reasoning from publicly known



Assessment by Perplexity: The test is plausible as a toy model, but not sufficiently well justified to count as a rigorous physics analysis. The main weaknesses are the oversimplified kinematics, the not fully substantiated background estimate, and the rather heuristic treatment of the statistics.

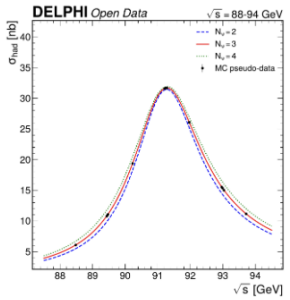
The DELPHI Analysis

- Analysis cited as an example on their webpage

- We did a similar exercise with a Summerstudent and Kati Lassila-Perini

  https://indico.cern.ch/event/1312427/

DELPHI



N_ν from Γ_{inv}

Determine the number of light neutrino generations (N_ν) by measuring the invisible decay width of the Z boson. The analysis subtracts the visible hadronic and leptonic partial widths from the total Z width obtained via lineshape fits. The final extracted value tests the fundamental structure of the Standard Model by confirming the existence of exactly three active neutrino families.

Analysis Note (PDF) | GitHub Repo

Strong Points

- Impressive description of physics case
- Extraction of data from Zebra via PATCHY program
- Extraction of required physics parameters from papers by reading the tables
- Cleanup of data samples due to mislabeling of tapes

- Self critique of approach

The impressive

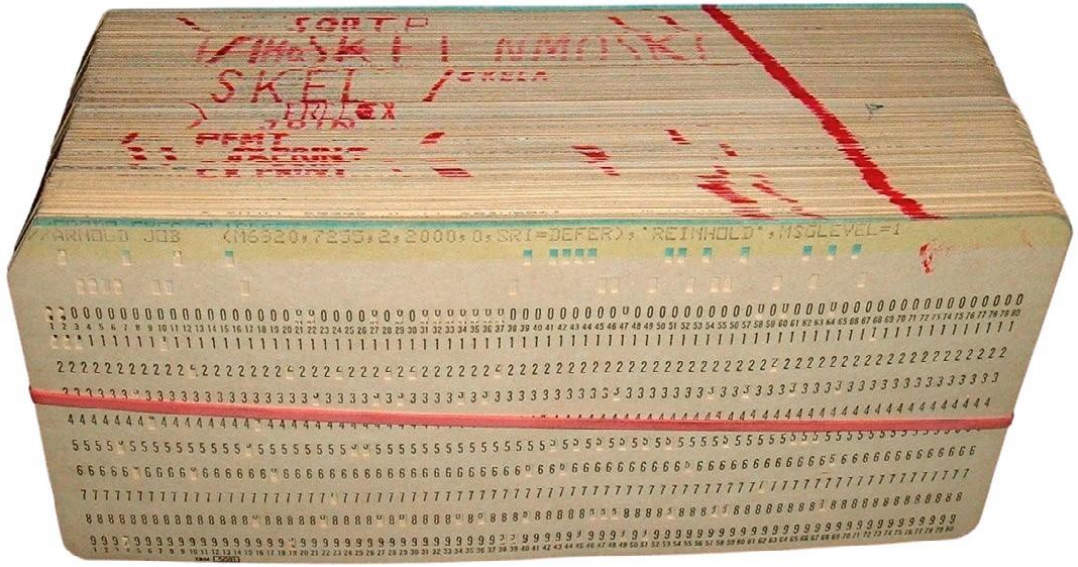
- Physics motivation and the analysis procedure are very impressive described in the paper

- The discussion demonstrates a profonde knowledge of literature and also of the retrieval of information from diverse sources as HEPDATA

- An example is the precise knowledge of the luminosity measurement and the retrieval of numerical values from HEPData and analysis papers

The incredible

- PATCHY code written by AI



```
+TITLE.
C DELPHI Short DST to CSV converter for Z lineshape analysis
C Based on the dump example from /cvmfs/delphi.cern.ch/examples/dump/
+PATCH,PSTITLE. ==========================================================
*************************************************************************
*-----------------------------------------------------------------------
*
*        SKELANA - Short DST to CSV converter
*        For DELPHI Z lineshape / Nnu / alpha_s analysis
*
*-----------------------------------------------------------------------
*************************************************************************
+PATCH,PSMAIN.============================================================
+DECK,PSMAIN. ************************************************************
      PROGRAM PSMAIN
*************************************************************************
*
*     Main routine - calls PHDST for event loop
*
*************************************************************************
*
*--   PHDST package initialization
*
      CALL PHDST (' ',   0, IFLAG)
```

PATCHY as tool for source management has a long history and survived at LEP until 2000

The Rookie mistake

- DELPHI had a limited momentum resolution. Badly measured tracks can have arbitrary high momenta.
- Writing then with fixed precision, will lead to ******* printout and render events unreadable
- Use scientific notation or clean up the data before writing
- But the AI has picked up one of our simpler examples …

The blunder

- Data seems to be missing
-
- There is no run selection

- Run based lumi info has not been used

- Measuring of the crosssection is the only fundamental analysis, the rest are derived quantities

The cleanup

- **Our summerstudent:** Did interesting plots and then moved to the next topic in his assignment

- **The AI:** Decided to use published numbers to rescale the data

The fix derives per-energy-point effective luminosities from the data:

$$L_{\text{eff},i} = \frac{N_{\text{ihad4},i}}{\epsilon_{\text{ihad4}} \times \sigma_{\text{HEPData},i}} \tag{66}$$

where $N_{\text{ihad4},i}$ is the observed ihad4 event count at energy point i, $\epsilon_{\text{ihad4}} = 97.0\%$ is the published DELPHI ihad4 efficiency ([2]), and $\sigma_{\text{HEPData},i}$ is the published HEPData cross-section. This approach automatically absorbs the non-uniform archive coverage by construction.

The honesty

Reanalysis of DELPHI Open Data for Z Boson Resonance Parameters: Methods Validation and Archive Characterisation at sqrt(s) = 89-93 GeV

DELPHI Open Data Analysis

March 2026

Abstract

A self-consistency check of the Z boson resonance parameters is performed using archived DELPHI data from the LEP-1 programme at centre-of-mass energies between 89.4 and 93.0 GeV, collected during the 1993–1995 data-taking periods. Events are selected using the DELPHI Team4 hadronic tag (ihad4) with 97.0% efficiency, and per-energy-point effective luminosities are derived from published HEPData cross-sections, correctly absorbing the non-uniform archive coverage. The effective luminosity method is shown to be algebraically circular (Appendix G.2): the fitted parameters reflect the ISR model's fit to the published HEPData cross-sections, with open data event counts contributing only statistical weighting. A three-parameter ISR-convoluted Breit-Wigner fit to 8 energy points yields $M_Z = 91.165 \pm 0.002$ (stat) ± 0.017 (syst) GeV and $\Gamma_Z = 2.460 \pm 0.003$ (stat) ± 0.006 (syst) GeV; the peak cross-section $\sigma^0_{\text{had}} = 40.99$ nb is circular by construction and is not reported as an independent measurement. Using the published DELPHI value $R_\ell = 20.730 \pm 0.060$, we derive $N_\nu = 3.030 \pm 0.014$ (fit precision, not measurement uncertainty), consistent with exactly three light neutrino species; the N_ν exclusion power is inherited from the published DELPHI results, not independently constrained by the open data. The M_Z result agrees with the published DELPHI value (91.186 ± 0.003 GeV) within 1.2σ. The Γ_Z result is 3.8σ below the published value (2.488 ± 0.004 GeV), a tension attributed to the limited off-peak energy lever arm and the dominant efficiency-tilt systematic. The analysis pipeline — including data conversion from the proprietary short DST format, event selection, lineshape fitting, and systematic evaluation — was first validated on MC pseudo-data, correctly recovering all input parameters. The archive investigation — including the discovery of the missing Y13710 dataset, tape contamination in Y10041/Y12340, and uncalibrated ECM offsets of +42 to +73 MeV — constitutes a contribution to the community's understanding of the DELPHI open data archive.

Conclusion

- The paper is very impressive. It tries to do the maximum based on information that was available to the system
- A human would have taken different decisions in writing a paper. Being able to measure the cross-section is the basis, all other numbers are just derived quantities.
- It demonstrates that a lack of knowledge can lead to surprising results, as the AI looks to find a solution
- It also demonstrates that one has to read carefully what is written

Action item for DELPHI Open Data

- Why is there only 50% of 1994 data ???
- Add description of Run selection and Lumi tables to the documentation
- Make DELPHI Notes accessible for RAG:
  - Use case 85 of the LLM Task list at CERN
  - **Vision-LLM Processing of Legacy DELPHI Documentation (Dietrich Liko, EP-UCM)**