

Starting a PhD in ML / NLP

The one-page version — venues, deadlines, tools, habits

WEEK 1 SETUP

Zotero + connector + **Better BibTeX** (pinned keys, auto-export to `paper/refs.bib`) · **Scholar profile** + 3 alerts · accounts: Overleaf, Prism, OpenReview, GitHub, HuggingFace, arXiv, **ORCID** · deadlines in calendar · AI assistant + project memory file · one git repo per project with `notes/`.

VENUES — CONFERENCE-FIRST FIELD

ML NeurIPS (abs ~May, Dec) · **ICML** (~Jan, Jul) · **ICLR** (~Sep, Apr; open reviews) · **AISTATS** (~Oct) · **AAAI** (~Aug) · **IJCAI** (~Jan) · **COLM** (~Mar, Oct) · COLT, UAI, MLSys, LoG.

NLP ACL · **EMNLP** · **NAACL** · **EACL** · **AAACL** · COLING · LREC · CoNLL, *SEM, SemEval · **TACL** (journal, rolling) · **Findings** (archival, citable) · Interspeech/ICASSP.

Adjacent KDD, WSDM, CIKM, SIGIR, WebConf, ECML-PKDD · CVPR/ICCV/ECCV.

Journals JMLR · **TMLR** (rolling, correctness over novelty) · CL, TPAMI.

ARR (NLP) Submit to ACL Rolling Review → reviews + meta-review → **commit** to a venue, which decides accept/Findings/reject. Track *two* deadlines. **Limitations section mandatory** (free pages) + Responsible NLP checklist.

Lifecycle preprint → submission → reviews → **rebuttal** → decision → camera-ready → poster/talk → citations. Budget 6–12 months from idea to published.

Judge a venue Scholar h5 · **CORE** · **CSRankings** · where the work you admire appears. Invitation email + 2-week acceptance + big fee = scam.

Aim your first submission at a good workshop. Full pipeline — writing, review, rebuttal, poster — at a fraction of the risk.

DEADLINES

Start: [paperswithcode.co/ai-deadlines](#) (Hugging Face) — filter by subfield or *top tier only*; tracks *every* milestone (abstract, reviews out, rebuttal, commitment, notification, camera-ready) with live countdowns, dates and locations; has an API + MCP endpoint. **Then:** [aideadlines.org/?sub=ML,CV,NLP](#) — the `sub=` tags *are* your board; edit as interests change (**HF mirror**). Cross-check [ccfdl.com](#) and always the venue page.

- Calendar each venue **3×**: deadline, −7d (freeze results), −21d (experiments running).
- Also: abstract deadlines, ARR commitment, supplementary, **rebuttal window** (don’t travel), camera-ready.
- AoE = UTC−12. Ablations always run late.
- Missing venue = a YAML pull request: easy first OSS contribution.

WRITING & SHARING

Overleaf real-time co-editing, official venue templates, comments; check for an institutional licence.

Prism OpenAI’s AI-native LaTeX workspace (2026): free, unlimited projects/collaborators/compile; the model sees the whole manuscript — prose, tables, LaTeX errors.

Local too git + `latexmk -pdf -pvc` + VS Code · **one sentence per line** (readable diffs) · `\input` per section · figures regenerated by script.

Before posting [arxiv-latex-cleaner](#) (names hide in LaTeX comments) · anonymise code ([anonymous.4open.science](#)) + PDF metadata · ACL venues enforce an **anonymity window** — check before arXiving.

arXiv cs.CL (NLP), cs.LG (ML), cs.AI, stat.ML · the first submission needs endorsement — your advisor handles it · preprints are the norm and accelerate citations.

Release code with a **LICENSE**, checkpoints and data on Hugging Face, an exact environment spec —

papers with runnable code get cited more, and your future self is the main beneficiary.

ZOTERO — NEVER LOSE A REFERENCE

- Connector: metadata *and* PDF in one click.
- Better BibTeX**: pin keys (`devlin2019bert`), auto-export the `.bib` into your repo. Commit it.
- Collections per project, tags per theme, saved searches, lab **group library**, annotate in the built-in reader.
- Fastest entry**: Scholar *Settings* → *Bibliography manager* → show BibTeX links; then **Cite** → **BibTeX** → paste into `refs.bib` (or Zotero *File* → *Import from clipboard*). Bulk: star into *My library* → Export.
- Cite the published version**, not the preprint, and check what Scholar gave you — “arXiv preprint” for a NAACL paper is the usual error. Cleaner: [Anthology](#) / [DBLP](#) per-paper BibTeX, or any DOI from the terminal:
`curl -LH "Accept: application/x-bibtex" https://doi.org/<DOI>`
- Monthly duplicate merge; one-line “why I care” on every save.

FIRST PROJECT: A FOCUSED SURVEY

Buys you: vocabulary in weeks · a taxonomy (gaps live in the empty boxes) · the bibliography you reuse forever · the baselines/datasets/metrics you must match · a publishable artefact and your related-work chapter.

- One-sentence question, agreed in writing.
- 5–10 seed papers from your advisor; steal their vocabulary.
- Inclusion/exclusion criteria written *before* screening.
- Same queries across Scholar, Semantic Scholar, Anthology, OpenReview, DBLP — **log query, date, hit count**.
- Snowball **backward** (references) + **forward** (cited-by); forward finds the last 6 months.
- Screen title+abstract, then full text: ~500 → 60–120 kept.
- Extract into a table**, one row per paper: key | venue/year | problem | method family | data | metrics | headline result | **compute** | code + does it reproduce | limitations | relevance 1–5.
- Cluster rows → taxonomy figure. Can’t draw it = don’t understand it yet.
- Write 10–15 pages *as you go*; end each section with “what’s missing”.
- Deliver **3 gaps** as testable questions + the experiment that answers each. That meeting starts your first paper.

Read in 3 passes (Keshav): 5 min skim → 1 h (figures, method) → hours (re-derive). Pass 3 for ~20 papers, pass 1 for hundreds. **Freeze the search date and ship.**

SEARCH LIKE AN INSTRUMENT

"exact phrase"	quote technical terms
a OR b	must be capitalised
-survey	exclude a term
author:"K Skianis"	author search
source:"EMNLP"	search within a venue
intitle:x	precision filter
Custom range	+ Sort by date
Cited by →	the most under-used feature in academia
search within citing articles	
All N versions	free PDF + published version

Alerts today 2–3 topic queries · **citation alerts on 5 anchor papers** ≈ a feed of your subfield · your own profile · a fixed 30-min weekly triage into Zotero.

Beyond Scholar [Semantic Scholar](#) (TLDR, citation intent, API) · [OpenAlex](#) (open graph) · [ACL Anthology](#) · [OpenReview](#) — read the *reviews* · [DBLP](#) · Connected Papers · [HF papers](#) · [paperswithcode.co](#) (the revived Papers with Code: trending papers, following feed, API + MCP — the old `.com` is dead).

ATTENTION SIGNALS

Citations Normalise by age (per year; last 12 months) · watch *acceleration* in Scholar’s bar chart · **who** cites > how many · cited-as-method > cited-as-background · beware survey piles and citation rings · count papers per year using a term (Anthology/OpenReview are scriptable).

GitHub [star-history.com](#): launch spike vs steady slope · last commit, contributors, who answers issues · **search issues for “reproduce”** before trusting a baseline · releases, pinned deps, checkpoints, **LICENSE** · cross-check HF downloads, PyPI, and use as others’ baseline · **stars:>500 pushed:>2026-01-01** · watch releases on repos you depend on.

Stars measure marketing; citations measure positioning. Neither measures correctness — reproduce one baseline yourself.

AI ASSISTANT: SETUP & RULES

Coding agent in the repo · project memory file with your conventions · connect filesystem + Zotero + arXiv/S2 via MCP · saved prompts (triage, extraction, hostile-reviewer pass) · dated `notes/` notebook. Use for: abstract triage, PDF→table extraction, unfamiliar maths, getting baselines to run, eval harnesses, plots, Slurm, LaTeX, pre-review, rebuttal prose from *your* arguments.

1 Never cite a paper you haven’t opened — models invent references. **2** Verify every number at the source. **3** Never paste NDA’d data or papers you’re reviewing (LLM-written reviews are banned). **4** Know your venue’s disclosure policy. **5** Outsource mechanics, never thinking.

YOUR PAGE: USERNAME.GITHUB.IO

Repo named *exactly* `username.github.io` → free HTTPS site. Use [AcademicPages](#) (*Use this template*) → edit `_config.yml` → fill `_pages/`, `_publications/`, `_talks/`, `_teaching/`, `_files/`, `_images/` → generate publications from your `.bib` with [markdown_generator/](#) → push (Actions builds it). Optional CNAME for a custom domain. Alternatives: al-folio, Minimal Mistakes, Hugo Blox, Quarto.

Example [y3nk0.github.io](#) — About, News, Research, Publications, Talks, Teaching, CV (PDF + HTML). Source is public: copy the structure. **Update after every acceptance**; a stale site reads worse than none.

EXPERIMENTS

One repo per project · pinned env (`uv/conda/Docker`) incl. CUDA · configs (Hydra/YAML) — every run reconstructible from config + commit hash · **3+ seeds, mean ± std** (single-seed +0.3 is noise) · track runs (W&B/MLflow) with configs, metrics, commit · tables and figures regenerate from logged results · learn Slurm: job arrays, checkpoint-and-resume · “can I actually run this?” is a design constraint.

HABITS THAT COMPOUND

Weekly advisor meeting, 5 written bullets (did / found / blocked / propose / need) · dated lab notebook *including* negative results · 20 min reading daily + a reading group · write continuously — it exposes holes in the story · review papers once you’ve published one · talk at posters you only half understand · summer schools, doctoral consortia, volunteering · rejection is the default outcome, not a verdict on you.

MILESTONES

M1 question agreed, criteria written, search log running, meeting slot fixed. **M2–3** survey done, taxonomy drawn, 3 gaps, one baseline reproduced, site live, venue chosen. **Y1** one submission out (submitted beats perfect), one rebuttal written, one talk given, code + checkpoints released, a one-page research statement.