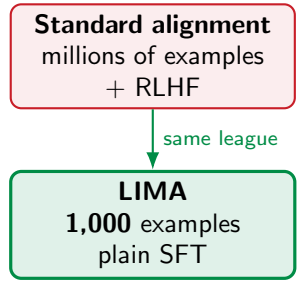
How much alignment data do you actually need?

The 2023 alignment playbook: *millions* of instruction examples plus RLHF over millions of human interactions.

LIMA fine-tunes LLaMa 65B on **1,000** curated examples – plain supervised loss, **zero** RLHF, **zero** preference data – and comes out competitive with RLHF-trained products.

Alpaca was trained on **52**$\times$ **more** data and *loses* to LIMA. Even GPT-4 prefers LIMA's answer to its own 19% of the time.

What does that tell us about where a model's ability really comes from?

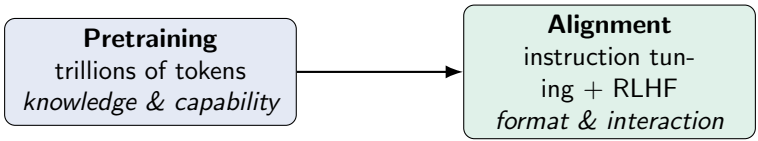


Outline

Where does “capability” come from?

Pretraining vs alignment – and what the standard pipeline assumes.

The two-stage view of an LLM



The usual assumption: alignment needs **large scale** – millions of instruction examples (FLAN, OpenAssistant) plus RLHF collected over millions of human interactions. "ChatGPT-level" is thought to require significant compute and specialized data.

LIMA's counter-position: given a *strong pretrained model*, a **small, carefully curated** set is enough – *no RL, no reward model, no preference data*.

The Superficial Alignment Hypothesis

The thesis the whole paper is built to test.

The one idea: an expert who just needs manners

Picture a brilliant specialist fresh out of years of training – they have read *everything*, but were never taught how to **hold a conversation**: hand them a question and they ramble, quote textbooks, or answer a different question.

- **Pretraining** is the years of study – *all* the knowledge and skill live here.
- **Alignment** is a short briefing on *how to answer*: be direct, be helpful, use this format.
- You would never need millions of examples to teach *manners* to someone who already has the expertise – a handful *shows* the format.

So the bet: capability is already there from pretraining. Alignment only has to **pick the voice** the model already learned to speak – which should take very little data.

The next slide states this precisely – the *Superficial Alignment Hypothesis* – and the whole paper is built to test it.

The hypothesis

Superficial Alignment Hypothesis: a model's knowledge and capabilities are learnt *almost entirely* during pretraining, while alignment teaches it *which subdistribution of formats* to use when interacting with users.

Corollary: if alignment is largely about *style*, then one could tune a pretrained model to high quality with a *rather small* set of examples.

Design consequence: collect 1,000 examples whose *outputs* share a uniform, helpful-assistant voice, but whose *inputs* (prompts) are as *diverse* as possible.

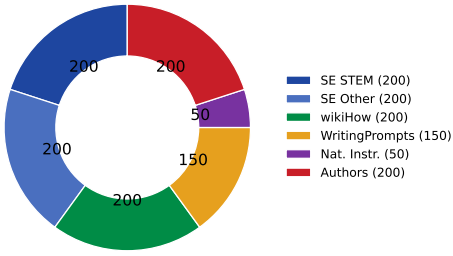
"Superficial" means *operating on the surface* (format/style selection) – **not** trivial, and **not** that curation is easy.

The recipe

1,000 examples – and how much work went into choosing them.

What is in the 1,000?

The 1,000 examples (~750k tokens)



- **750** mined from community Q&A – Stack Exchange, wikiHow, Reddit WritingPrompts.
- **250** hand-written by the authors in a uniform assistant tone (200 used for training).
- 50 Super-Natural-Instructions NLG tasks.
- Includes **13 safety** examples (refuse malicious prompts).

~750k tokens over exactly 1,000 sequences. Diverse prompts, stylistically uniform answers.

These 1,000 were heavily engineered

Not random samples – aggressive filtering and hand-curation:

- **Stack Exchange:** keep top answers with score ≥ 10; drop answers < 1200 or > 4096 chars, first-person, or cross-referencing; strip links/HTML, keep code blocks. Sample across domains at *temperature 3* for coverage.
- **wikiHow:** sample category-first for diversity; title $\rightarrow$ prompt, body $\rightarrow$ response.
- **Reddit:** manually curated (top answers skew jokey/trolling).
- **Hand-written:** uniform "acknowledge, then answer" format – hypothesized to act like light chain-of-thought scaffolding.

The lesson is **not** "1,000 random examples work." It is that a *small, diverse, high-quality* set works – and quality here took real effort.

Training LIMA

- Base: **LLaMa 65B**, plain supervised fine-tuning.
- 15 epochs, AdamW, lr $10^{-5} \rightarrow 10^{-6}$, batch 32.
- Special **end-of-turn** token to separate speakers.
- **Residual dropout** ramped $0 \rightarrow 0.3$ over the layers.

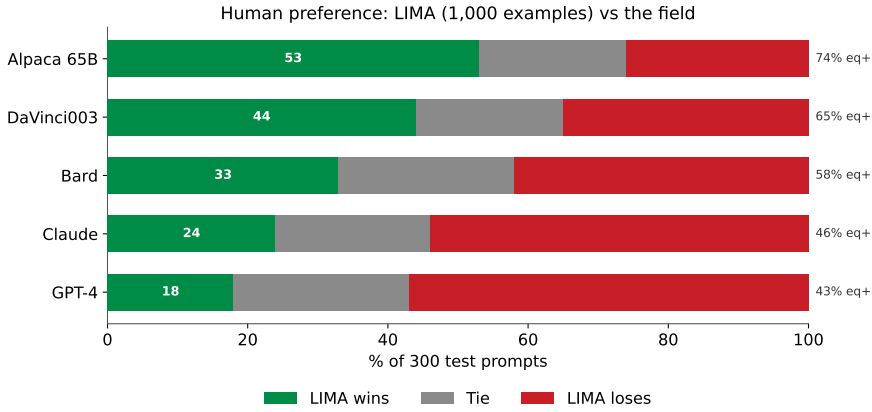
Perplexity does not track quality – it even *anti*-correlates. Checkpoints were picked *by hand* on a 50-example dev set, not by validation loss.

A caution worth remembering: standard LM metrics can point the wrong way for alignment – which is why the evaluation below is human / model-judged.

Does it work?

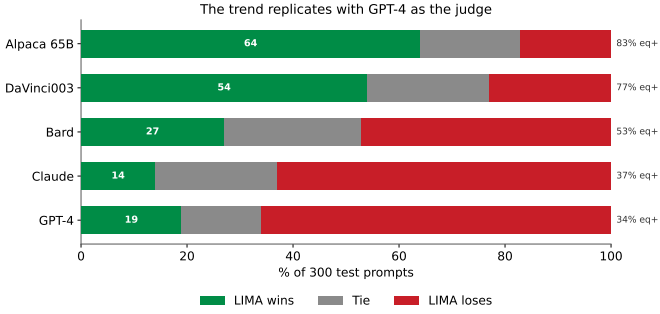
Human and GPT-4 preference against the strongest 2023 systems.

Human preference: 1,000 examples vs the field



LIMA is equal-or-better than **DaVinci003** (RLHF) 65% of the time and **Alpaca 65B** (52k examples) 74% – and holds its own against Bard, Claude, and GPT-4.

The trend replicates with a GPT-4 judge



Swapping human raters for a GPT-4 judge gives the *same* ordering.

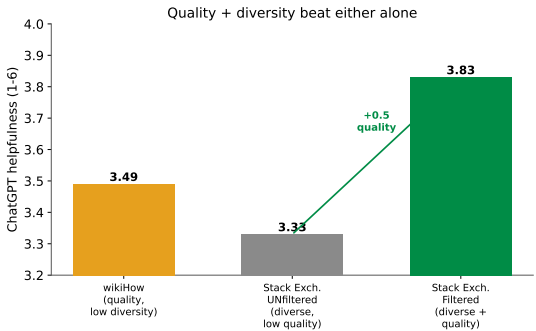
Absolute quality (50 prompts): **50% excellent**, 38% pass, 12% fail – and it holds up out-of-distribution (45% excellent).

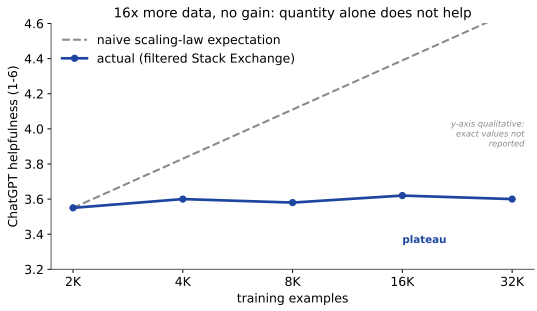
Inter-annotator agreement: human-human $\approx$ human-GPT-4, so GPT-4 is a reasonable proxy judge.

Why is less more?

The ablations are the heart of the argument.

Diversity and quality beat quantity

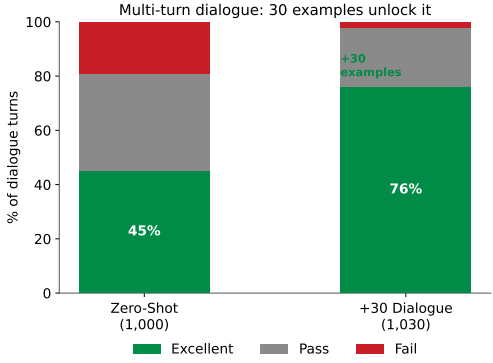




Left: filtering Stack Exchange for quality lifts helpfulness $3.33 \rightarrow 3.83$; a diverse source beats a homogeneous one. **Right:** scaling *quantity* $2\text{K} \rightarrow 32\text{K}$ ($16\times$) *without* adding diversity gives essentially **no** improvement.

Predict-first: unlocking multi-turn dialogue

The 1,000-example model was trained on *single-turn* data. How many dialogue examples to make it a coherent chat model?



Just **30** dialogue chains take Excellent responses from **45%→76%** and cut the failure rate from 15/42 turns to **1/46**.

Dialogue ability was *latent* from pretraining – 30 examples merely **invoked** it. Strong evidence for the hypothesis.

What this implies – and what it does not

The conceptual payload:

- **Capability lives in pretraining**; alignment is a thin layer that selects *which* latent response format to emit.
- Instruction tuning may teach *format*, not new skills – so millions of examples are largely redundant once diversity is covered.
- LIMA (plain SFT) beats RLHF-trained DaVinci003 on this eval – RLHF's marginal value may be more about *robustness/safety* than core capability.

Misreadings to avoid: this is *not* "data doesn't matter" (quality+diversity matter enormously) and *not* "RLHF is useless" (LIMA concedes robustness/safety gaps). It is a *hypothesis*, argued via one 65B model and one eval suite.

Limitations

- Building high-quality examples takes **significant mental effort** and is **hard to scale** – the curation *is* the cost.
- **Not as robust** as product models: an unlucky sample or an adversarial prompt can still produce a weak or unsafe answer.
- Comparisons are against moving April-2023 products exposed to millions of real prompts – indicative, not definitive.
- The base model (LLaMa 65B) is doing heavy lifting; results may not transfer to weaker bases.

Recap

- **1,000** curated examples + plain SFT on LLaMa 65B $\approx$ RLHF products; no reward model, no RL.
- **Superficial Alignment Hypothesis:** capability from pretraining, alignment selects format/style.
- **Quality + diversity beat quantity** – 16$\times$ more data, no gain; filtering for quality, a clear gain.
- **30** dialogue examples unlock multi-turn chat – latent ability, invoked.
- Guardrails: data matters (a lot), RLHF still helps robustness, and curation is the real expense.

The series in one line: pretraining builds the ability (*LIMA*); a thin, well-chosen layer aligns it – cheaply (*LoRA*), from preferences (*DPO*), or with RL (*GRPO*).