A Prompting Workflow and a Use Case in Mathematical Statistics

Edgar Dobriban [1]

Abstract

We describe a prompting workflow used to solve a problem in mathematical statistics. The problem was to close a gap between upper and lower minimax rates for pointwise density estimation from data corrupted by Wasserstein-bounded contamination. A prompting loop with GPT-5 Pro produced two useful proof ideas: a sharper transport-interpolation bound for the estimator's contamination bias and a dynamic Benamou–Brenier construction for the lower-bound modulus. These suggestions did not constitute a proof; they were converted into a proof only through human verification, fixing, and reference checking. The workflow required only web interface access. The full research project took a few weeks of human-AI iteration after more than two years of human-only work had left the rate gap unresolved. We estimate a time saving of several months. We also report failure modes: hallucinated references, overbroad theorem statements, and increasingly elaborate incorrect fixes after a mistake was pointed out. We also discuss our approaches to mitigate these issues.

1. Research Setting

In this paper, we describe a prompting workflow used to solve a problem in mathematical statistics. The problem was to close a gap between upper and lower minimax rates for pointwise density estimation from data corrupted by Wasserstein-bounded contamination. This research problem had previously been under investigation by the authors, who have obtained partial results. AI was used to close the gap and help obtain the final results. Our work thus illustrates how AI can integrate in the workflows of active researchers in the mathematical sciences.

The statistical problem can be described as follows. We want to estimate the value $f(x_0)$ of an unknown smooth density f at a fixed point $x_0 \in \mathbb{R}^p$. However, instead of samples from f, we observe samples from a contaminated density $\widetilde{f}$ whose distribution is within Wasserstein distance ε of the true distribution. Classical nonparametric density estimation without contamination has squared-error minimax rate $n^{-2s/(2s+p)}$ for s-Hölder smooth densities (Stone, 1984; Tsybakov, 2009). The research question was how this rate changes when every observation may be moved by a small Wasserstein-bounded perturbation, a robustness model borrowed from optimal transport (Villani, 2008; Santambrogio, 2015). This question is also inspired by recent relared work on robust estimation under Wasserstein perturbations (Zhu et al., 2022; Liu & Loh, 2022).

A team consisting of one PhD student and one faculty member worked on the problem on and off over the course of two years, obtaining partial results. The analysis focused on the one-dimensional case where $p = 1$, and for Euclidean Wasserstein perturbations. Letting $\mathcal{M}$ be the minimax risk based on n datapoints for ε-perturbations, the human-only analysis gave $n^{-2s/(2s+1)} \vee \varepsilon^{4s/(2s+1)} \lesssim \mathcal{M} \lesssim n^{-2s/(2s+1)} \vee \varepsilon^{2s/(s+2)}$. The sample-size dependence was the classical rate $n^{-2s/(2s+1)}$ (Tsybakov, 2009), but the contamination exponents did not match. The authors did not know whether the upper bound, the lower bound, or both were loose. The final theorem, proved after the AI-assisted workflow, is that in p dimensions under $W_{q,r}$ contaminations (where the true datapoints X and the contaminated datapoints $\tilde{X}$ can be coupled so that $(\mathbb{E}\|X - \tilde{X}\|_q^r)^{1/q} \le \varepsilon$, the minimax rate is

$$\mathcal{M}(\varepsilon, s, p, q, r) \asymp n^{-2s/(2s+p)} \vee \varepsilon^{2s/(s+1+p/r)}.$$

In particular, $\mathcal{M}(\varepsilon, s, 1, 2, 2) \asymp n^{-\frac{2s}{2s+1}} \vee \varepsilon^{\frac{2s}{s+3/2}}$, so neither the upper nor the lower bounds were sharp in the human-only analysis.

2. Workflow: Prompt, Extract, Verify, Fix

The workflow was deliberately simple. It used the web interface of a frontier reasoning model, GPT-5 Pro[1] via the web interface. The "Extended Thinking" reasoning mode was used throughout. At the time of performing this work several months ago, local AI agents were not quite reliable enough to use directly. Each query contained a self-

[1]University of Pennsylvania. Correspondence to: Edgar Dobriban <dobriban@wharton.upenn.edu>.

Preprint. June 17, 2026.

[1]At the time of writing GPT 5.2 Pro was the latest model in the GPT family.

contained LaTeX "packet": a problem definition, the current theorem statement, the relevant part of the proof, and a short request. The most useful requests were questions such as:

> I am working on this theorem. Can you go through this proof and see if you can improve either the upper bound or the lower bound until they match? Please identify the exact obstruction and prepare ready-to-copy rigorous LaTeX.

In a later stage of the work, a second prompt class asked for adversarial checking rather than invention: "Can you double-check this proof from beginning to end and identify any incorrect arguments?"

After receiving the model-generated answer, the researcher extracted the mathematical claim that seemed potentially most useful, rewrote it as a separate lemma, and verified it without trusting any citation or omitted step. When a proposed proof failed, the failure was reported back to the model, with additional prompting to fix the specific issues. The complete loop was:

1. **Context packet.** Paste a minimal LaTeX fragment containing the problem statement, nonstandard definitions, and current proof attempt.
2. **Useful fact extraction.** Extract promising ideas, arguments, or references from the answer.
3. **Reference checking.** Treat every cited theorem as untrusted until manually checked in the original source. A model-provided citation was allowed to suggest where to look, but never to certify a result.
4. **Human proof verification.** Check any model-claimed results by carefully reading through them.
5. **Asking for improvement.** If anything problematic is found in the proof, point out the issue in a follow-up prompt to the model asking for a fix.
6. **Fix or stop.** If the model's repaired proof kept becoming longer and longer without resolving the core error, stop that thread.

Prompts of the type that were used during the workflow are provided in the Appendix. Reproducing this workflow exactly might be both challenging and not strictly required. A researcher working on a problem of their choosing would be encouraged to adapt a similar workflow, changing it according to their own best judgment and needs.

Our workflow reflected the practical considerations at the time of performing the work, several months before the writing of this paper. At that time, AI agents running on local compute were not yet reliable enough for mathematical work at this level, hence the work was constrained to the web interface. Moreover, since models can be lead astray by incorrect claims and approaches appearing previously in their context window, we were forced to start new interactions periodically. This has lead to a somewhat fragmented and less-than-perfectly organized workflow.

For the same reason, we think that the exact workflow may be somewhat harder to reproduce than if the workflow itself was the main goal, and the mathematical result was the secondary goal; whereas the opposite was true. Nonetheless, we believe that the workflow has proved its usefulness and viability by helping resolve a significant research bottleneck. Moreover, the authors have used similar workflows for other problems afterwards, with additional success. At the same time, our workflow is relatively simple and minimalistic, reflecting the approach of researchers focused on substantive mathematical problems; and thus can be viewed as a baseline for other, more sophisticated workflows. For these reasons, we believe it is valuable to share this workflow with the community.

GPT Suggests Improved Bias Bound

User: [Prompt includes LaTeX draft of paper and a query: "Can you double-check this proof from beginning to end, and make sure it is fully correct? Please identify any place where there are incorrect arguments."]
GPT: [...]
The correct formulation is as follows. Define the interpolation

$$T_t(x) = (1-t)x + t\,T(x), \qquad X_t = T_t(X),$$

where $X \sim P$ and T is the optimal transport map from P to Q. Then

$$\frac{d}{dt}\,\mathbb{E}\big[\varphi(X_t)\big] = \mathbb{E}\big[\varphi'(X_t)\cdot\big(T(X)-X\big)\big],$$

where the expectation is with respect to the original law of $X \sim P$. [...]

Consequence. Once the derivative is written correctly, the displayed Cauchy–Schwarz step goes through. The correct bound is

$$\big|\mathbb{E}_Q[\varphi] - \mathbb{E}_P[\varphi]\big| \le A^{1/2} W_2(P,Q),$$
$$A := \int_0^1 \mathbb{E}_P\big[\varphi'(X_t)^2\big]\,dt.$$

which coincides with your statement after noting that $\mathbb{E}_P\big[g(X_t)\big] = \mathbb{E}_{P_t}[g]$ for any test function g, where P_t is the law of X_t.

3. What the AI Contributed

Upper bound. The upper bound studies the bias of a kernel density estimator under contamination. An initial Lipschitz argument bounded the contamination bias too weakly. GPT-5 Pro suggested replacing the pointwise Lipschitz step by interpolation along an optimal transport map. Let T transport f to $\widetilde{f}$, let $S_t(y) = (1-t)y + tT(y)$, let $\Delta(y) = T(y) - y$, and set $\varphi(x) = h^{-p}K((x_0 - x)/h)$. The key identity is

$$\mathbb{E}_{\widetilde{f}}\varphi(X) - \mathbb{E}_f\varphi(X) = \int_0^1 \int \nabla\varphi(S_t(y))^\top \Delta(y) f(y) dy dt.$$

Holder's inequality then converts the contamination size into $W_{q,r}(f, \widetilde{f})$ while the remaining term is controlled by the density of $(S_t)_\# f$.

The AI did not supply all missing assumptions. In particular, the existence of a deterministic map for the required norm and the boundedness of the interpolated density required manual checking and fixing using optimal transport results such as those from Gangbo & McCann (1996). Crucially, the AI did not directly solve the problem and it did not even provide a fully correct and rigorous suggestion. Therefore, human expertise in these areas was indispensable in order to finally resolve the problem. Still, the model's contribution was to propose a useful approach: differentiating expectations along the transport interpolation.

Lower bound. A classical approach for showing lower bounds requires constructing two nearby contaminated experiments whose true target values $f(x_0)$ differ. The human-only proof used a transportation-cost bound that scaled linearly in the bump amplitude, which made the result too weak. GPT-5 Pro identified this scaling obstruction and suggested a dynamic construction. A simplified version of the resulting lemma is the following.

Lemma 1 (Local dynamic transport bound)**.** *Let $\mathcal{G}$ be a density on $\mathbb{R}^p$, let I be a bounded box, and let g be supported on I with $\int g = 0$. Assume $\mathcal{G} \geq c_I > 0$ on I and $\|g\|_\infty \leq c_I/2$. If U is supported on I and $\nabla \cdot U = g$, then for $r \geq 1$,*

$$W_{q,r}^r(\mathcal{G} \pm g, \mathcal{G}) \leq \frac{2^{r-1}}{c_I^{r-1}} \int \|U(x)\|_q^r dx.$$

The proof uses the path $\rho_t = \mathcal{G} + tg$ and velocity $v_t = -U/\rho_t$, so that $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$. The dynamic formulation of optimal transport (Benamou & Brenier, 2000; Brenier, 2004) then bounds the transport cost by the action along this path. The model's first version had constants and sign conventions that needed correction, and it gave incomplete references, but the amplitude-sensitive dynamic idea was decisive.

Query topic	AI-suggested object	Human verification/ fix needed
Upper bound	Differentiate a kernel expectation along an optimal transport interpolation	Existence of T, density bounds for interpolants, endpoint and norm cases
Lower bound	Use a continuity-equation path to bound local Wasserstein cost	Correct constants, divergence construction, applicability of dynamic formulation
Paper audit	Ask model to search for inconsistencies	All mathematical claims and citations treated as untrusted

Table 1. Examples of human verification and fixes required during the workflow.

4. Reproducibility, Cost, and Evaluation

The workflow required access to a strong reasoning LLM and the initial proof arguments. A minimal reproduction takes two to three hours: read the proof packet, run the two prompt templates, and compare the model output to the target ideas in Table 1. Full mathematical verification takes longer and depends on expertise, but this is true for many challenging proof reviews.

We evaluated correctness by checking everything line by line, including tracking down all cited results in a trustworthy primary source such as an established book or paper. We also used the model as an adversarial reader, in the style of LLM-as-judge and critique workflows (Zheng et al., 2023), but treated its critiques as leads rather than decisions. The manuscript reporting the final results was then revised globally to remove notational inconsistencies introduced during the iteration.

The human-only project had been going on for more than two years with the minimax gap still open. The AI-assisted phase took a few weeks and hundreds of queries. Because the decisive missing ingredients included relatively advanced results from optimal transport that were unfamiliar to the authors, we estimate that independently finding the same route would likely have taken several additional months.

5. Failure Modes and Lessons

The most important failure mode was confident false specificity. The model often named plausible references whose stated theorems did not cover the needed norm, exponent, or regularity assumptions. A second failure mode was proof inflation: after a mistake was pointed out, the model often produced a longer and more complicated purported fix rather than abandoning the claim. A third failure mode was

Mode	Symptom	Mitigation
False citation	Cited theorem has wrong norm, exponent, or hypotheses	Check primary source before use
Proof inflation	Corrections become longer but the same gap remains	Ask for atomic lemmas or stop
Too broad	Statement is plausible only away from endpoints or under extra regularity	Prove a weaker lemma; handle endpoints separately
Algebra errors	Signs, constants, or lower bounds are dropped	Recompute independently

Table 2. Failure modes observed during the workflow and the corresponding safeguards.

overgeneralization; many candidate lemmas were true only after weakening assumptions or restricting endpoint cases.

The most useful safeguard was to separate exploration from certification. During exploration, the model was allowed to be speculative: it could suggest analogies, name possible tools, or point to a scaling mismatch. During certification, none of that mattered unless it survived a separate proof. This separation made the workflow easier to use. It avoided both extremes: treating the model as an oracle, which would be unsafe, and dismissing every imperfect answer, which would miss the value of partially correct mathematical search.

The workflow can thus be viewed as an interface between broad model memory and narrow human proof checking. It is not suitable when the user cannot verify the output or when the desired result cannot be decomposed into small claims. It is well suited to settings where a draft proof already exists and the bottleneck can be localized.

A key lesson was that the productive unit of AI assistance in proof discovery may not necessarily be a completed proof, but a well-posed proof approach. The workflow was useful because the model had broad mathematical memory and could connect nonparametric statistics with optimal transport tools, while the human supplied judgment about which claims were meaningful, which references were valid, and when an iteration had stopped making progress.

This should transfer to other ML, statistics, and optimization projects where a gap can be isolated into a concrete mathematical bottleneck. Related AI-assisted mathematical case studies have begun to document similar complementarity between model-generated suggestions and human verification (Bubeck et al., 2025; Salim, 2025; Jang & Ryu, 2025); the contribution here is a concrete, lightweight workflow and a detailed account of its failure modes in a statistics proof.

References

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Brenier, Y. Extended monge-kantorovich theory. In *Optimal Transportation and Applications: Lectures given at the CIME Summer School, held in Martina Franca, Italy, September 2-8, 2001*, pp. 91–121. Springer, 2004.

Bubeck, S., Coester, C., Eldan, R., Gowers, T., Lee, Y. T., Lupsasca, A., Sawhney, M., Scherrer, R., Sellke, M., Spears, B. K., Unutmaz, D., Weil, K., Yin, S., and Zhivotovskiy, N. Early science acceleration experiments with gpt-5, 2025. URL `https://arxiv.org/abs/2511.16072`.

Gangbo, W. and McCann, R. J. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996. doi: 10.1007/BF02392620.

Jang, U. and Ryu, E. K. Point convergence of nesterov's accelerated gradient method: An ai-assisted proof. *arXiv preprint arXiv:2510.23513*, 2025.

Liu, Z. and Loh, P.-L. Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 12(1):312–362, 2022.

Salim, A. Accelerating mathematical research with language models: A case study of an interaction with gpt-5-pro on a convex analysis problem. *arXiv preprint arXiv:2510.26647*, 2025.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Stone, C. J. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pp. 1285–1297, 1984.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

Zhu, B., Jiao, J., and Steinhardt, J. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256 – 2283, 2022.

A. Workflow prompts

This appendix is designed to make the workflow reproducible. The researcher prepares a compact proof packet and runs several web-chat interactions. The packet should contain: the theorem statement, definitions of nonstandard objects, the current upper and lower bounds, the proof step believed to be loose, and the exact question.

A.1. Prompt templates

Prompt A: rate obstruction

I am working on this theorem. Can you go through this proof and see if you can improve either the upper bound or the lower bound until they match? Please identify the exact obstruction. If you propose a replacement lemma, state it with all assumptions and prepare ready-to-copy rigorous LaTeX. Do not skip constants or regularity conditions.

Prompt B: proof audit

Can you double-check this proof from beginning to end, and make sure it is fully correct? Please identify any place where there are incorrect arguments, missing assumptions, wrong references, or other issues. For every issue, provide a ready-to-copy fix.

Prompt C: targeted repair

The previous argument seems to fail because [insert precise error]. Give a corrected lemma with weaker assumptions, or propose a different argument. Please separate the mathematical claim, its proof, and any references needed to justify it.

Prompt D: atomic verification

Please check only the following displayed inequality. Expand every step, including the chain rule, the scaling in h, and where each assumption is used. Do not discuss other parts of the proof.

A.2. Acceptance checklist

A model suggestion was accepted only after passing the following checks: the statement is an isolated lemma; the hypotheses match the theorem in which it will be used; endpoint cases are either handled or excluded; constants and signs are recomputed by hand; every external theorem is checked in a primary source; and the lemma is strong enough to recover the claimed rate by an independent scaling calculation.

B. Transcript Excerpts

The following excerpts are transcript snippets with unrelated manuscript context omitted and line breaks normalized. They are included to make the workflow closer to the working interaction itself. They are not proof certificates.

B.1. Optimal-map existence: from overclaim to corrected hypothesis

User prompt: checking an optimal-map sentence

Is this correct? Since the cost function is $\|x-y\|_q^r$, $r>1$, and the density f induces an absolutely continuous measure, the optimal transport map T pushing f to $\tilde{f}$ exists and is unique (Villani, 2008, Theorem 10.28 and surrounding discussion). [The prompt then pasted Villani's Theorem 10.28 and surrounding remarks.]

Model response excerpt: missing assumptions

Almost — but you need two standard hypotheses that are not explicit in your sentence, and you should restrict the norm. A clean, correct statement is: let $1 < q < \infty$ and $r > 1$, and let $\mu = f(x)dx$ and $\nu = \tilde{f}(x)dx$ be probability measures on $\mathbb{R}^n$ with finite r-moments. Then for the cost $c(x, y) = \|x - y\|_q^r$ there exists a unique μ-a.s. optimal transport map T pushing μ to ν. Theorem 10.28 assumes finite optimal cost; for $c(x, y) = \|x - y\|_q^r$ this follows from finite r-moments. For $q \in \{1, \infty\}$ the norm is not strictly convex, the twist condition can fail, and uniqueness of the optimal map need not hold in dimensions at least two. Also, uniqueness should be stated μ-a.s.

This interaction did not directly write a proof, but instead it changed how the proof was written. The accepted fix was to avoid relying on map uniqueness in endpoint cases unless the required hypotheses were present.

B.2. Endpoint-norm audit: replacing a map argument by a plan argument

User prompt: proof audit after generalizing to $W_{q,r}$

I am working on this proof. Could you please go through it step-by-step? In detail. And check if everything looks good to you. Please be fully rigorous and complete.

Model response excerpt: endpoint failure mode

Almost. The references and the logic now line up, but there is one technical assumption you should add or a tiny rewrite you can do to avoid it. In Lemma \ref{lem:Wr-shift-d} you write: "Let $r > 1$. Let T be the optimal transport map pushing f to $\tilde{f}$." The lemma's statement currently says "for any densities $f, \tilde{f}$ on $\mathbb{R}^d$." To use a map T rather than a plan, you do need the source f to be absolutely continuous and the cost to yield a Monge solution. Either add the hypothesis, or keep the lemma for arbitrary $f, \tilde{f}$ and do a plan-based proof: let π be an optimal plan, set $\Delta = y - x$ and $S_t(x, y) = (1 - t)x + ty$, let $\nu_t = (S_t)_{\#}\pi$, and run the same Hölder argument integrating $\nabla\varphi(S_t) \cdot \Delta$ against π.

This was a useful failure-mode interaction: a plausible proof worked for the intended smooth-density setting, but the statement had been written too broadly.

B.3. Lower-bound obstruction: identifying the wrong amplitude scaling

User prompt: unmatched one-dimensional lower bound

[Prompt included the old lower-bound proof, ending with the conclusion $\mathcal{M}_I(\varepsilon; n, s, L) \gtrsim n^{-2s/(2s+1)} \vee \varepsilon^{4s/(s+3)}$.] Please be fully rigorous and complete, do not skip any steps.

Model response excerpt: rate obstruction

The upper bound is already tight. The lower bound can be sharpened by replacing the crude local W_2 estimate with a refined, amplitude-sensitive bound derived from the Benamou–Brenier dynamic formulation. The root cause is that the earlier local lemma scales like $W_2^2 \lesssim D^2 \int g_+$, i.e. linearly in the bump amplitude, while the correct small-amplitude scaling is quadratic in the amplitude. This improves the modulus and closes the gap.

This response identified the mathematical bottleneck: the old lower bound used the wrong scaling for moving positive and negative parts of a localized bump.

B.4. Drop-in dynamic lemma

User prompt: request for a rigorous replacement

Can you prepare a fully rigorous drop-in replacement for my existing result? Please be complete and rigorous and do not skip any steps.

Model response excerpt: one-dimensional dynamic bound

Let q be a probability density on $\mathbb{R}$. Let $I \subset \mathbb{R}$ be a bounded interval, and let $g \in L^1(\mathbb{R})$ satisfy $\mathrm{supp}(g) \subset I$ and $\int g = 0$. Define $G(x) = \int_{-\infty}^{x} g(t)dt$. Suppose $q(x) \geq c_I$ on I and $\|g\|_\infty \leq c_I/2$. Then $q \pm g$ are densities and

$$W_2^2(q+g,q) \leq \frac{2}{c_I}\int_{\mathbb{R}} G(x)^2 dx, \qquad W_2^2(q-g,q) \leq \frac{2}{c_I}\int_{\mathbb{R}} G(x)^2 dx.$$

For $q+g$, use $\rho_t = q + tg$ and $v_t = -G/\rho_t$. Then $\partial_t \rho_t + \partial_x(\rho_t v_t) = 0$, and Benamou–Brenier bounds W_2^2 by $\int_0^1 \int |v_t|^2 \rho_t$.

The final paper generalized this idea from one dimension to $\mathbb{R}^p$ by using a vector field U with $\nabla \cdot U = g$.

B.5. Reference and theorem audit

Model response excerpt: references as leads, not certificates

Check each ingredient. Existence/uniqueness of an optimal map for strictly convex costs is the relevant Monge theorem. Geodesics and displacement interpolation are represented by straight-segment pushforwards when a Monge map exists. The dynamic formulation is the continuity equation plus the action $\int |v_t|^p \rho_t$. For the W_1 side one should use the Beckmann formulation. Your local W_r construction with $U_h = (G_{1,h}, 0, \ldots, 0)$ and the Benamou–Brenier upper bound implies $W_r(f_\pm, q) \lesssim Lh^{s+1+d/r}$, and the separation $|f_+ - f_-|(x_0) \asymp h^s$ yields the exponent $s/(s+1+d/r)$.

This interaction was useful mostly as a checklist. Every cited item was still manually checked before entering the final proof.

B.6. Atomic algebra check

User prompt: local Hölder calculation

Can you give the details on this? $\|x-y\| > h$, then

$$|D^\beta g_h(x) - D^\beta g_h(y)| \leq 2\|D^\beta g_h\|_\infty \leq 2LaM_\beta h^\alpha \leq 2LaM_\beta \|x-y\|^\alpha.$$

Model response excerpt: detailed replacement display

Use the chain rule for $g_h(x) = \tau T((x-x_0)/h)$ with $\tau = Lh^s$ and $s = l + \alpha$:

$$\begin{aligned}
|D^\beta g_h(x) - D^\beta g_h(y)| &\leq |D^\beta g_h(x)| + |D^\beta g_h(y)| \leq 2\|D^\beta g_h\|_\infty \\
&\leq 2\sup_z \left|\tau h^{-l} D^\beta T\left(\frac{z-x_0}{h}\right)\right| \leq 2\tau h^{-l}\|D^\beta T\|_\infty \\
&= 2(Lh^s)h^{-l}(aM_\beta) = 2LaM_\beta h^\alpha \leq 2LaM_\beta \|x-y\|^\alpha,
\end{aligned}$$

where the last step uses $\|x-y\| > h$ and $\alpha \in (0,1]$.

This kind of interaction did not generate a new idea, but it made the final proof more explicit and easier to audit.

C. How to Use This as a Workflow Benchmark

To test whether the workflow transfers, an ML researcher can replace our density-estimation packet by a proof packet from another project. The benchmark should have a localized obstruction, such as a loose exponent, an unexplained endpoint case, or a suspected missing assumption.

The model interaction should be scored by whether the model returns a small number of useful strategies or lemmas. The human verifier should then decide whether each strategy is useful, or each lemma is true, false, or true under weaker hypotheses. This benchmark design reflects the main empirical lesson of the project: in mathematical research, a partially correct route can be valuable even when the raw answer is not trustworthy.