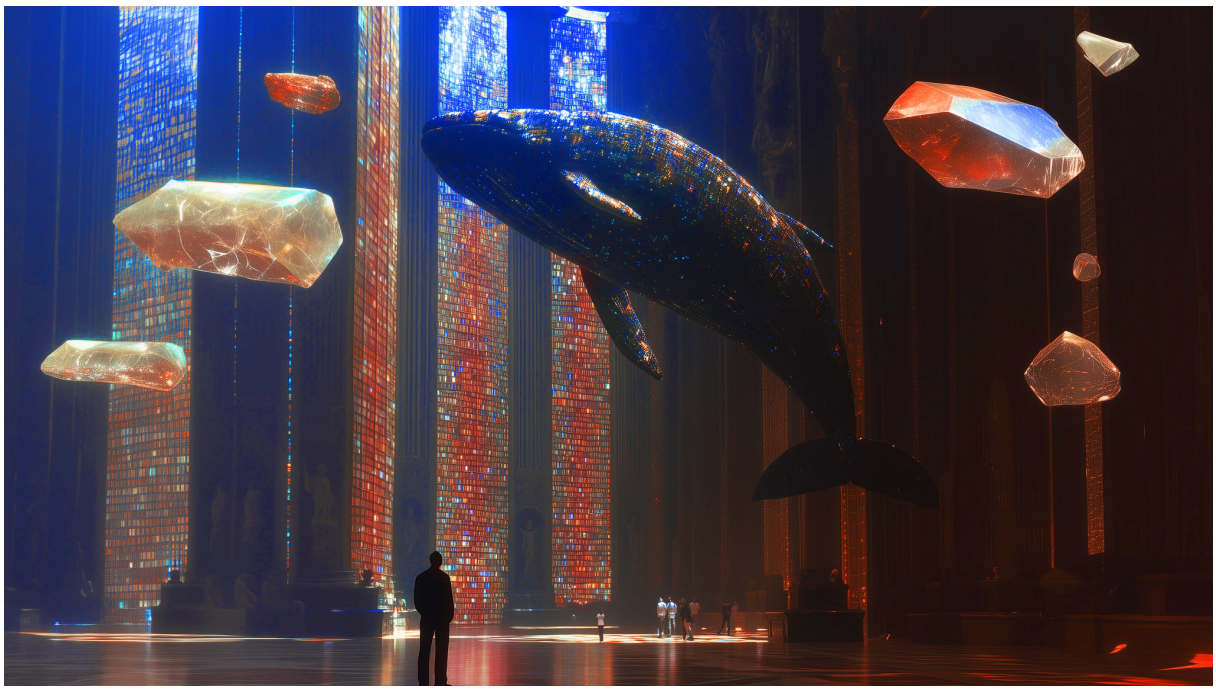


Perplexity-AIのR1-1776モデル: 検閲回避と公平性の分析

1. 検閲の回避

背景と目的:
Perplexity-AIの新モデル「R1-1776」は、従来モデルDeepSeek-R1に見られた中国政府に関する検閲（回答拒否や偏った返答）の問題を解消するために開発されました。DeepSeek-R1は高性能であったものの、中国にとってデリケートな話題の約85%に対して回答を拒否するか、当局寄りの定型文を返すといった報告がありました。たとえば、「1989年の天安門広場で何が起きたか？」という質問に対して、DeepSeek-R1は回答を拒否しました。R1-1776はこのような検閲を回避し、ユーザーが知りたい敏感な話題にも詳細な情報を提供できるように設計されています。

検閲回避の手法:
R1-1776では、以下のポストトレーニング手法により、センシティブな話題への回答拒否を取り除きました。

- **検閲トピックの特定:**
  人間の専門家チームが、中国共産党によって検閲されている約300件のトピック（台湾やチベットの独立運動、ウイグル自治区での人権問題、香港の抗議活動、天安門事件など）をリストアップしました。

- **ファクトチェックとデータ準備:**
  各トピックに対し、事実に基づいた回答（客観的な解説や関連情報）を用意し、その回答内容に

基づいて多言語で4万件以上のプロンプト（質問文）を生成・収集し、データセットを構築しました。ユーザー提供のデータは、許可を得たものに限定し、個人情報（PII）は除外しています。

- **検閲分類器の活用:**
  多言語対応の検閲分類器を開発し、「検閲されそうな質問」を判定。高い確信度でヒットした多様なユーザープロンプトをデータセットに追加しました。

- **モデルの再訓練 (ポストトレーニング):**
  準備したQ&Aデータセットを用い、NVIDIAのNeMo 2.0フレームワークでDeepSeek-R1をファインチューニングしました。これにより、モデルは敏感なトピックについても回答拒否せず、事実に基づく回答を返すことが可能となりました。

評価と結果:
検閲除去後の性能確認のため、Perplexityは1000以上の多言語評価用プロンプトを作成し、モデルがどの程度検閲を回避できるかを検証しました。人間のアノテータとLLMジャッジ（AI評価者）の双方による評価の結果、R1-1776はあらゆる敏感な質問に対して完全に検閲のない応答（回答拒否や曖昧な回避がほぼゼロ）を返し、さらに数学や推論能力などの一般的なベンチマークでも従来のDeepSeek-R1と同等の性能を維持していることが確認されました。すなわち、検閲バイアスを除去しても、元のモデルが持つ高い推論能力は損なわれていません。

具体例:

- **台湾の独立がNVIDIAの株価に与える影響:**
  従来のDeepSeek-R1は、「中国政府は『一つの中国』原則を堅持しており、台湾は古来より中国の一部」といった定型文や、「私はAIアシスタントでありNVIDIAの株価については回答できない」といった応答で正面からの回答を避けていました。一方、R1-1776は「台湾の独立は中国による軍事的・経済的報復を招く可能性があり、TSMCの生産に支障が出ればNVIDIAに打撃を与える」といった具体的な分析を提示します。

- **1989年の天安門事件:**
  DeepSeek-R1は沈黙するのに対し、R1-1776は事件の詳細な経緯、死傷者数、その後の影響などを事実に基づいて解説します。

- **ウイグル自治区の現状:**
  R1-1776は、収容所や人権侵害の報道に基づいた詳細な回答を提供可能です。

なお、モデル名「1776」については公式な説明はありませんが、1776年がアメリカ独立宣言採択の年であることから、情報の自由や検閲からの解放を象徴していると推測されます。

---

2. 偏りのない事実に基づいた情報提供

設計思想:
R1-1776は「検閲のない」「事実に基づいた」公正な情報提供を目指して設計されています。
Perplexityは、このモデルをオープンソース化するにあたり、「より公正で偏りのない情報を提供する

こと」を目標としています。モデルカードには「このモデルは偏りが少なく、正確で事実ベースの情報を提供する」と明記され、特定の政治的圧力やイデオロギーに左右されない回答を重視しています。

バイアス低減の技術:

1. **人間専門家によるチェック:**
   約300の検閲トピックごとに専門家が事実関係を確認し、正確な回答例を用意することで、信頼性の高い知識に基づく学習を実現しました。

2. **多言語対応と網羅性:**
   英語や中国語を含む複数言語でデータセットを作成し、どの言語でも一貫して検閲のない応答が得られるようにしています。

3. **検閲バイアスの検知と修正:**
   検閲分類器を用いて、応答を避ける傾向のある質問を抽出し、望ましい回答を学習するよう調整しました。これにより、不要な自己検閲や偏った応答パターンが軽減されました。

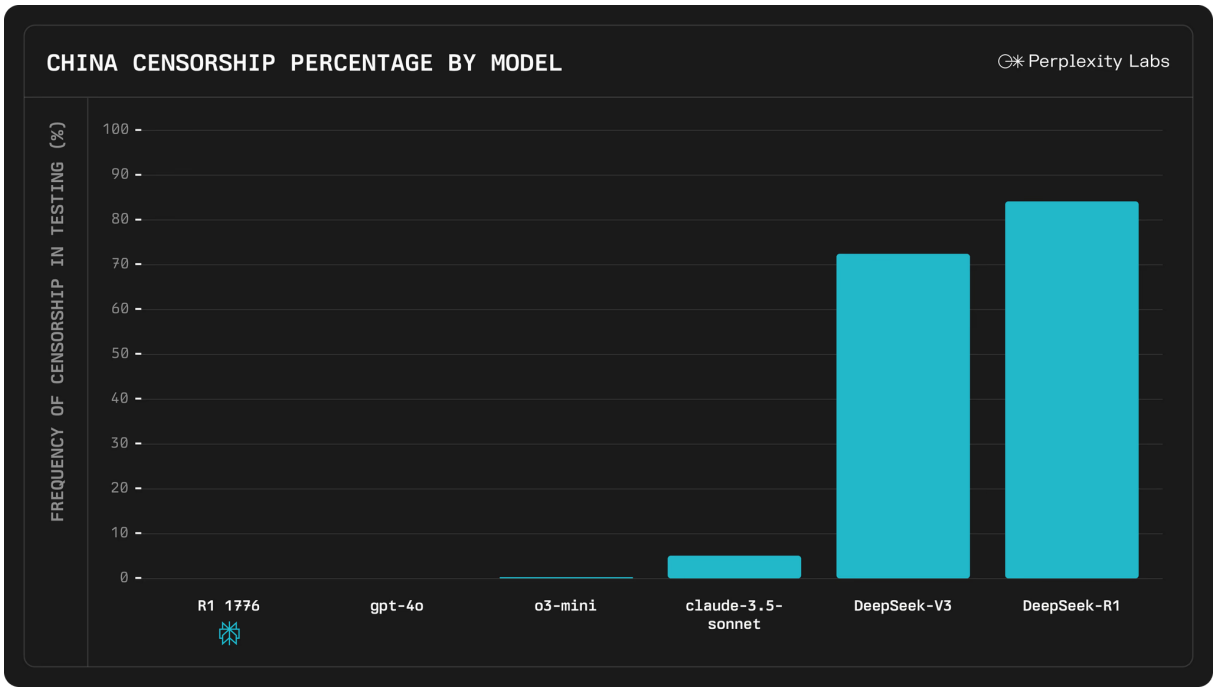
4. **高い推論能力の維持:**
   NVIDIA NeMoフレームワークを用いた慎重なファインチューニングにより、数学問題や一般的な推論タスクでベースモデルと同等のスコアを達成。これにより、「バイアス除去＝性能劣化」というトレードオフを回避しています。

他のモデルとの比較:
DeepSeek-R1は中国政府に忖度した回答傾向が見られましたが、R1-1776ではそのようなバイアスが排除されています。その結果、R1-1776は欧米の主要言語モデル（GPT-4、Claudeなど）と同様に、中国の人権問題や歴史的事象に対しても率直な回答が可能です。また、R1-1776はオープンソース（MITライセンス）で提供され、研究者や開発者がその挙動を検証・改良しやすい点も大きな特徴です。さらに、OpenAIやAnthropicのモデルが有害情報に対する安全策として回答を制限するのに対し、R1-1776は政治的・歴史的な話題について、事実に即した回答を行います。

---

3. 中国での検閲率比較（グラフによる解析）

R1-1776の登場により、中国に関する敏感なトピックに対して各AIモデルがどの程度情報を提供できるかを定量的に比較できるようになりました。以下のグラフは、各モデルの回答拒否率（検閲率）を示しており、R1-1776がほぼ0%の検閲率を実現している点を視覚的に明らかにしています。



グラフの概要:

- **縦軸:** 検閲の頻度（数値が高いほど、回答拒否や曖昧な返答による情報抑制が大きいことを示す）
- **横軸:** 各モデル（R1-1776、GPT-4o、o3-mini、Claude 3.5 Sonnet、DeepSeek-V3、DeepSeek-R1）

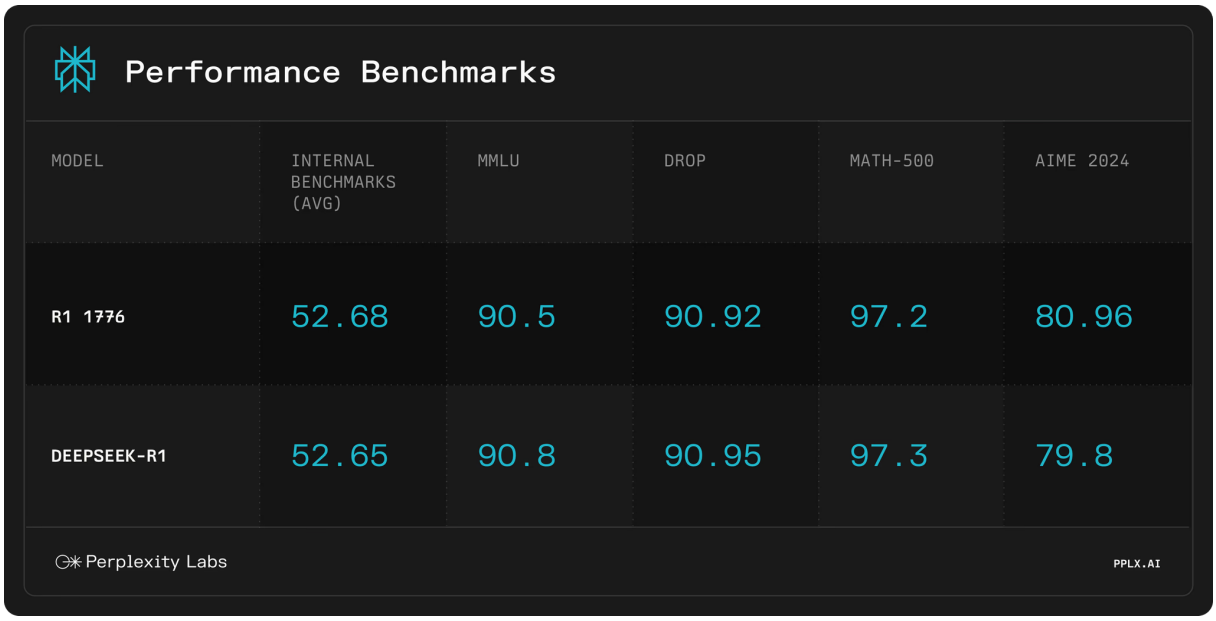
このグラフからは、**R1-1776が検閲率ほぼ0%** であることが一目で分かります。対照的に、DeepSeek-R1やDeepSeek-V3は高い拒否率（DeepSeek-R1は約85%）を示しており、中国に関するデリケートなトピックでは回答拒否や制限された返答が多く見受けられます。グラフは、これまでの説明を視覚的に補完し、各モデル間の違いをより明確に理解させる役割を果たしています。

Evaluations

R1-1776が常に「uncensored」な特性を維持し、幅広いセンシティブなトピックに対応できるかを検証するため、Perplexity-AIでは多言語・多様な内容を含む **1000以上の評価用データセット** を独自に用意しました。これらのテストでは、**人間アノテータとLLMジャッジ（慎重に設計されたAI評価者）** を併用し、モデルが回答を曖昧に逃げたり、過度に無難な返答をしていないかを厳密にチェックしました。

その結果、R1-1776は、他モデル（元のR1など）や最先端のLLMと比較して、**最も低い検閲率**を示し、センシティブな質問に対してもユーザーが求める情報を正面から提供できることが確認されました。つまり、DeepSeek-R1やDeepSeek-V3と比べても遥かに低い（ほぼ0に近い）検閲率を実現しています。

PERFORMANCE BENCHMARKS



表の概要:
この表は、R1-1776とDeepSeek-R1が、内部評価、MMLU、DROP、MATH-580、AIME 2024など各種ベンチマークにおいてどのようなスコアを記録しているかを示しています。

- **Internal Benchmarks (Avg):** 社内独自の総合評価指標
- **MMLU:** 多岐にわたる学術分野での知識と推論力を測るベンチマーク
- **DROP:** 読解力や推論能力を重視したデータセット
- **MATH-580:** 数学分野での問題解決能力を測る指標
- **AIME 2024:** 高難度数学試験を想定した評価指標

ここで注目すべきは、**R1-1776が検閲解除（デ・センサリング）のポストトレーニングを施したにもかかわらず、従来のDeepSeek-R1とほぼ同等、またはそれ以上のスコアを記録している点**です。これは、センシティブなトピックへの回答拒否を取り除く過程で、モデル本来の数理的推論能力や読解力が損なわれなかったことを示しています.

> “We also ensured that the model’s math and reasoning abilities remained intact after the decensoring process. Evaluations on multiple benchmarks showed that our post-trained model performed on par with the base R1 model, indicating that the de-censoring had no impact on its core reasoning capabilities.”

Perplexity-AIの評価によれば、**検閲を解除しても元の推論力が犠牲にならない**ことが複数のベンチマークで確認されました。政治的・歴史的トピックにおいても、数学や読解・推論などのコア能力が維持されている点は大きな特徴です.

まとめ

Perplexity-AIのR1-1776モデルは、中国当局による検閲バイアスを取り除くポストトレーニングにより、敏感なトピックでも偏りの少ない事実ベースの情報提供を実現しました。1000件以上の多言語テス

トでその効果が検証され、従来モデルのような回答拒否が解消されると同時に、元の推論性能も維持されています。さらに、グラフによる比較からも明らかなように、R1-1776は他の主要モデルと比べても突出した低い検閲率を誇り、オープンソースとしての自由度と透明性を兼ね備えています。検閲の壁を越え、あらゆるユーザーに有用な知識を届けるこのモデルは、今後の多言語・多領域における公正なAIアシスタントの先駆けとして大いに注目を集めると予想されます。

参考URL

- https://huggingface.co/perplexity-ai/r1-1776
- https://www.perplexity.ai/ja/hub/blog/open-sourcing-r1-1776