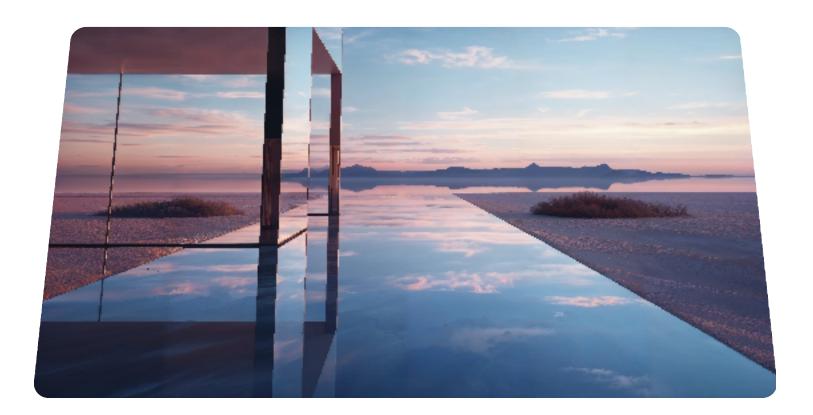


Blog Careers FAQ Getting Started

Try Perplexity



Written by Perplexity Team Published on Nov 29, 2023





Introducing PPLX Online LLMs

The first-of-its-kind Online LLM API.

We're excited to share two new PPLX models: pplx-7b-online and pplx-70b-online! Our online models are focused on delivering helpful, up-to-date, and factual responses, and are publicly available via pplx-api, making it a first-of-its-kind API. pplx-7b-online and pplx-70b-online are also accessible via Perplexity Labs, our LLM playground.

LLMs have transformed the way we find information. However, there are two limitations with most LLMs today:

- 1. Freshness: LLMs often struggle to share up-to-date information.
- 2. Hallucinations: LLMs can also output inaccurate statements.

Our **pplx-7b-online** and **pplx-70b-online** models address current limitations by additionally providing helpful, factual, and up-to-date information in its responses.

PPLX Online LLMs

Our LLMs, pplx-7b-online and pplx-70b-online, are online LLMs because they can use knowledge from the internet, and thus can leverage the most up-to-date information when forming a response. By providing our LLMs with knowledge from the web, our models accurately respond to time sensitive queries, unlocking knowledge beyond its training corpus. This means Perplexity's online LLMs can answer queries like "What was the Warriors game score last night?" that are challenging for offline models. At a high level, this is how our online LLM works:

- 1. Leverage open sourced models: our PPLX models build on top of mistral-7b and llama2-70b base models.
- 2. In-house search technology: our in-house search, indexing, and crawling infrastructure allows us to augment LLMs with the most relevant, up to date, and valuable information. Our search index is large, updated on a regular cadence, and uses sophisticated ranking algorithms to ensure high quality, non-SEOed sites are prioritized. Website excerpts, which we call "snippets", are provided to our pplx-online models to enable responses with the most up-to-date information.
- 3. **Fine-tuning**: our PPLX models have been fine-tuned to effectively use snippets to inform their responses. Using our in-house data contractors, we

carefully curate high quality, diverse, and large training sets in order to achieve high performance on various axes like helpfulness, factuality, and freshness. Our models are regularly fine-tuned to continually improve performance.

Evaluating Perplexity's Online LLMs

Perplexity's mission is to build the world's best answer engine - one that people trust to discover and expand their knowledge. To achieve this, we are deeply focused on providing helpful, factual, and up-to-date information. To benchmark our LLMs' performance on these axes, we curated evaluation datasets to reflect challenging yet realistic use cases for answer engines. For each query in each evaluation set, contractors were given two model responses and instructed to select the response that performed better for the following criteria:

- Helpfulness: which response answers the query and follow the specified instructions better?
- Factuality: which response provides more accurate answers without hallucinations, even for questions that require very precise or niche knowledge?
- Freshness (inspired by FreshLLMs): which response contains more up-todate information? A model excels in this criterion if it is able to answer queries with "fresh" information.

In addition to the three criteria above, model responses were also evaluated holistically. To evaluate responses holistically, evaluators were asked to pick the response they would rather receive from an human assistant who is helping with the query.

Curating the Evaluation Set

For this evaluation, we carefully curated a diverse set of prompts with the goal of effectively evaluating helpfulness, factuality, and freshness. Each prompt was manually selected, ensuring a high level of control over the quality and relevance of the data. The dataset spans a wide range of answer engine prompts and encompasses a comprehensive overview of what we want Perplexity to excel at. This is critical for us to our model performance evaluations to have high signal.

Each of the three evaluation sets contains 50 prompts. Some examples from these sets include:

- Helpfulness: Create a table of all the USA soccer players that played in the World Cup finals and make columns for their stats such as goals, assists etc.
- Factuality: Explain the toughness of AISI 1015 and AISI 1040 steels.
- **Up-to-date-ness:** Which autonomous car company got banned from San Francisco in 2023?

Generating Model Responses

We evaluated four models:

- pplx-7b-online: Perplexity's model, which includes access to information from the internet. This model was finetuned using mistral-7b.
- pplx-70b-online: Perplexity's model, which includes access to information from the internet. This model was finetuned using 11ama2-70b.
- gpt-3.5-turbo-1106: OpenAl's model, accessed via the API and with no additional augmentations. Note that we conducted this evaluation using the latest gpt-3.5 model.
- 11ama2-70b-chat: Meta Al's model, accessed via our pplx-api and with no additional augmentations.

For each prompt, the same search results and snippets were provided to the pplx-7b-online and pplx-70b-online models. All responses were generated using the same hyperparameters and system prompt.

System Prompt

Current date: Monday, November 20, 2023.

Ranking Model Responses With Human Evaluation

For each pairwise comparison, our in-house contractors were provided the prompt itself, the evaluation criteria (i.e., helpfulness, factuality, or freshness), and the model responses displayed side-by-side. The ordering of the responses were randomized at each turn, and the source models were not revealed to the evaluator. Evaluators were instructed to select the response they holistically prefer, as well as which performs better on the specific evaluation criterion, with ties allowed for both. Finally, evaluators were allowed to use internet search to verify the accuracy of responses. We built our own in-house preference ranking tool to conduct this evaluation.

Evaluation Results

We use the pairwise preference rankings collected from human evaluation to calculate per-task Elo scores for each model. Elo scores are commonly used to rank a population of players in tournament-style competitions to measure the relative performance of the players. When applied to large language models, these scores provide a prediction of how likely a human evaluator might favor the output of one model over another.

While Elo scores are typically calculated for a sequence of comparisons to account for changes in players' abilities, we aim to quantify the performance of models whose abilities cannot change. Accordingly, we adopt the Bootstrap Elo methodology described in **Duan et al** and also utilized by **Imsys** for their **Chatbot Arena**, which entails calculating Elo scores for many random permutations of the comparisons. For our calculations, we calculated Elo scores over 5000 permutations and used the distributions of these Elo scores to calculate 95% confidence intervals.

The results, shown in Figure 1, demonstrate that our PPLX models can match and even surpass gpt-3.5 performance on Perplexity-related use cases. In particular, pplx-7b-online and pplx-70b-online model responses are preferred over gpt-3.5 and llama2-70b model responses by human evaluators for their accurate and up-to-date answers.

The pplx-7b-online and pplx-70b-online perform better than gpt-3.5 and llama2-70b, on the freshness, factuality, and holistic criteria.



Figure 1. Estimated Elo scores and 95% confidence intervals for pplx-7b (-online), pplx-70b (-online), llama2-70b, and gpt-3.5 across four different evaluation sets. For example, on the freshness axis, pplx-7b and pplx-70b perform better than gpt-3.5 and llama2-70b, with an estimated Elo score of 1100.6 and 1099.6 vs 879.3 and 920.3, respectively.

We also calculated the pairwise win rates from the evaluation. The results, shown in Figure 2, illustrate perhaps not surprisingly, the pplx online model responses are preferred to others' on **freshness** and on **factuality**.

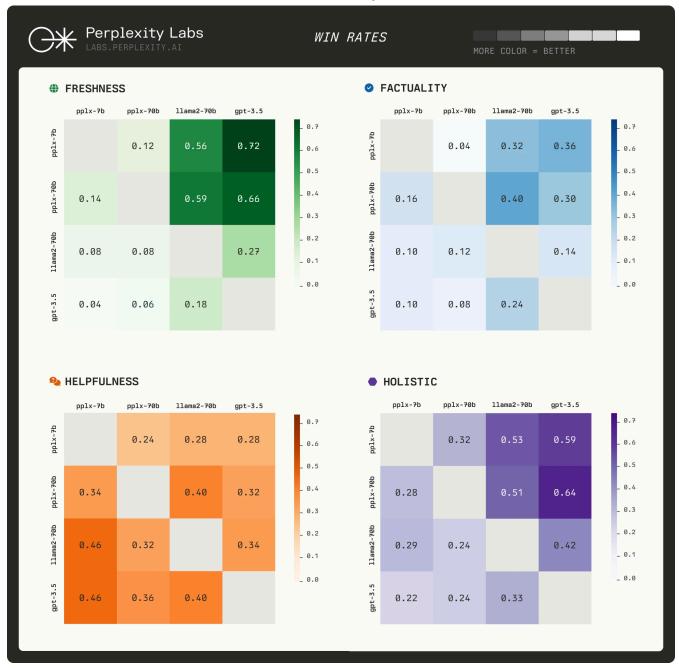


Figure 2. Win rates of models on the vertical axis against the models on the horizontal axis. For example, pplx-70b (-online) responses were rated as holistically superior to gpt-3.5's on 64% of queries and gpt-3.5's responses were rated as holistically superior to pplx-70b (-online) on 24% of queries.

Here is an example from our freshness evaluation dataset.

Prompt: How many people in USA have a driving license?

pplx-7b-online

In 2023, there are a projected 243.4 million licensed drivers in the United States, which includes 89% of US adults. This number has increased from 238.6 million in 2022. The percentage of licensed drivers varies across different age groups and states. For example, in 2019, 84.1% of the US population had a driver's license. The highest percentage of licensed drivers is among 25 to 49-year-olds, with 90.1% of the population having a driver's license. In California, there were 27 million licensed drivers, making it the state with the highest number of licensed drivers.

pplx-70b-online

In 2021, around 233 million people held valid driving licenses in the United States. By 2023, the number of licensed drivers in the US is projected to reach 243.4 million, which includes 89% of US adults. This represents an increase from 238.6 million in 2022.

gpt-3.5

As an Al language model, I don't have access to real-time data. However, as of 2021, it was estimated that around 227 million people in the United States held a valid driver's license. Please note that this number may have changed since then, so it's always best to refer to the latest official statistics for the most accurate information.

llama2-70b

I don't have access to real-time data, but I can provide some information on the approximate number of people in the United States who have a driver's license. According to the Federal Highway Administration, there were approximately 227.8 million licensed drivers in the United States as of 2020. This number includes both commercial and non-commercial drivers. It's worth noting that this number is likely an underestimate, as it doesn't account for individuals who may have a driver's license but do not drive regularly or at all. Additionally, this number does not include individuals who are not eligible for a driver's license,

such as those who are under the age of 16 or have certain disabilities. It's also important to note that the number of licensed drivers can vary by state, with some states having a higher percentage of licensed drivers than others.

Overall, the evaluation results demonstrate that our PPLX models can match and even outperform <code>gpt-3.5</code> and <code>llama2-70b</code> on Perplexity-related use cases, particularly for providing accurate and <code>up-to-date</code> responses.

Accessing Perplexity's Online Models

We are thrilled to announce the pplx-api is phasing out of beta into general public release! This is the first time our pplx-7b-online and pplx-70b-online models are accessible via the pplx-api. Additionally, our pplx-7b-chat and pplx-70b-chat models are out of alpha, and are now accessible with our general release.

With this, we are introducing a new usage-based pricing structure. We're excited for our users to utilize our state-of-the-art infrastructure, which utilizes NVIDIA H100s to provide blazing fast inference. Pro users will receive a recurring \$5 monthly pplx-api credit. For all other users, pricing will be determined based on usage. To sign up as a Pro user, click here. Reach out to api@perplexity.ai for commercial inquiries.

You

® Settings

≅ Profile

</>>
</>
API

COMPANY PRODUCT RESOURCES LABS FOLLOW US

About Desktop Appetting StaPtedX-API Overwietter)

Careers iPhone AppGeneral FA@aline LLM@odcast

Press Inquirendroid Apprechnical FRagground Discord

Privacy Poli@earch Con@iaeiloeedback Docum@intratiols

Terms & CanellaiohtsSearch API FAQs Linkedin

API Privacy Prodicty be

API Terms of Service



© Copyright 2024 Perplexity - Where Knowledge Begins