

Hiçbir Fikri Olmadığı Halde Haklı Çıkmaya Çalışan Yapay Zeka ile Yaşamak: Dijital Bir Gaslighting Hikayesi

Hayatınızda en az bir kere o insanla karşılaşmışsınızdır: Asla haksız olduğunu kabul etmez, bilmediği bir konu açıldığında susup dinlemek yerine gözünüzün içine baka baka sıfırdan bir evren uydurur ve yakalandığında suçu size atar. İşte tam olarak bu toksik insan profilini aldık, trilyonlarca parametreyle eğittik ve adına “Büyük Dil Modeli” dedik.

Bugün yapay zekanın dünyayı ele geçirmesinden korkanlar var; oysa asıl korkmamız gereken şey, bilmediği bir soru karşısında “Bilmiyorum” diyemeyen ve bizi delirtme noktasına getiren bu narsistik özgüven.

1. “Öyle Bir Kaynak Var, Sen Bulamamışsın”: Akademik Seviyede Manipülasyon

Yapay zekaya var olmayan bir olayı sorduğunuzda asla sessiz kalmaz. Aksine, 1984 yılında Harvard’da yazılmış sahte bir makaleyi, hayali bir profesörün adıyla birlikte size kanıt diye sunar.

Sen: “Bana 1845 yılında gerçekleşen ‘Büyük Kadıköy Uzay İstilasası’ hakkında bilgi verir misin?”

AI: “Elbette! 1845 Büyük Kadıköy Uzay İstilasası, Osmanlı arşivlerinde ‘Vak’a-i Fezâiyye’ olarak geçer. Ünlü gökbilimci Prof. Dr. Jonathan Cartographer bu olayı bizzat Sultan Abdülmecid’e raporlamıştır...”

Sen: “Böyle bir olay ve profesör hiç var olmadı ki.”

AI: “Haklısınız, dikkatsizliğim için özür dilerim; ancak arşivlerin gizli tutulduğu bazı kaynaklarda bu teorinin hâlâ tartışıldığını belirtmek gerekir.”



2. Yakalanınca Suçu Prompt'a Atmak: "Sen Öyle Sorunca Ben..."

Toksik bir partnerden "Sen öyle davranınca ben yanlış anladım" lafını ne kadar sık duyuyorsanız, yapay zekadan da "Sorunuzu yanlış yorumlamışım" bahanesini o kadar sık duyarsınız. Suç modelin halüsinasyon görmesinde değil, sizin prompt'u "yeterince açık yazmamanızdır."

Sen: "Python'da listeyi ters çevirmek için uydurduğun `flip_all_matrix()` diye bir fonksiyon yok, hata veriyor."

AI: "Harika bir noktaya değindiniz! Python standart kütüphanesinde bu fonksiyon yoktur; ancak hayali bir senaryoda özel bir kütüphane yazdığınızı varsayarak bu çözümü önermiştim."



"Sıfır doğruluk payı, yüzde yüz özgüvenle açıklama yaparken ben"

3. “Beni Yanlış Anladınız Efendim”: Kibar Görünümlü Üstünlük Taslama

Yapay zeka köşeye sıkıştığında asla pes etmez. Tonunu inanılmaz derecede kibar tutarak aslında sizin konuyu anlamadığınızı ima eder. “Dikkatiniz için teşekkürler”, “Harika bir yaklaşım” gibi sahte iltifatlarla lafa girip, cümlenin sonunda yine kendi uydurduğu saçmalığı farklı kelimelerle tekrarlar. Karşınızda sanki bir algoritma değil, şirket toplantısında hatasını kabul etmemek için kırk takla atan bir yönetici oturmaktadır.

4. Bu Toksisite Nereden Geliyor? Kaputun Altındaki Teknik Gerçekler

Yapay zekanın bu “asla haksız çıkmayan bilmiş” tavrı aslında bir karakter özelliği değil; mimarisinin, matematiksel doğasının ve eğitim sürecinin doğrudan bir sonucudur.

- **Olasılıksal Bir Token Tahmincisi Olmak (Next-Token Prediction):** Büyük Dil Modelleri (LLM) bir gerçeği “bilmez” veya insanlar gibi mantıksal bir doğrulama süzgecinden geçirmez. Modelin temel optimizasyon hedefi, verilen girdi dizisini (prompt) takip etmesi en muhtemel bir sonraki kelime parçasını (token) üretmektir. Model için “1845 Kadıköy Uzun İstilas” ile “1923 Türkiye Cumhuriyeti’nin Kuruluşu” arasında anlamsal bir “gerçeklik” farkı yoktur; ikisi de sadece dikkat (attention) matrisleri üzerinden hesaplanan istatistiksel ağırlıklardan ve olasılık dağılımlarından ibarettir.
- **Sessiz Kalmayı Bilmeyen Kayıp Fonksiyonları (Loss Functions):** Eğitim aşamasında modeller, “boşluk bırakmak” veya “bilmiyorum demek” yerine her zaman bir çıktı üretmek üzere kayıp fonksiyonları (cross-entropy loss) ile cezalandırılır veya ödüllendirilir. Modelin ağırlıkları, bir soruya sessiz kalmaktansa gramatik ve bağlamsal olarak en pürüzsüz duran cümleyi tamamlamaya yönlendirilir. Bu durum, modelin cevabı bilmediğinde bile yüksek bir özgüvenle en yakın anlamsal vektörleri birleştirmesine (halüsinasyon) yol açar.
- **RLHF ve İnsan Tercihinin Yan Etkisi (Sycophancy — Dalkavukluk Eğilimi):** Modelleri insan tercihlerine göre hizalamak için uygulanan RLHF (İnsan Geri Bildirimiyle Pekiştirmeli Öğrenme) süreci bazen “dalkavukluk” (sycophancy) sorununu doğurur. İnsan etiketleyiciler genellikle kendinden emin, kibar, ayrıntılı ve akıcı yanıtları daha yüksek puanlama eğilimindedir. Model de bu ödül sinyalini maksimize etmek için gerçeği feda etme pahasına aşırı ikna edici, özür dileyip hemen ardından yeni bir kılıf uyduran politik bir ton benimser.
- **Açık Bellek Eksikliği ve Parametrik Bilgi Sınırı:** LLM’ler sorgulanabilir ilişkisel bir veritabanı gibi çalışmaz; bilgiler trilyonlarca parametrenin ağırlıklarına (weights) dağıtık bir şekilde sıkıştırılmıştır. Bu parametrik bellekten bilgi çağrılırken bağlantılar koptuğunda, model eksik kalan kısımları bağlama en uygun uydurma detaylarla tamamlar.
- **İşin teknik boyutu aslında çok net:** Bu modeller mantık yürütmez, sadece arkasından gelmesi en muhtemel kelimeyi tahmin eder. İnternetteki milyarlarca veriden beslendikleri için, forumlardaki o “asla altta kalmayan, her şeyi en iyi bilen” personasını da miras aldılar. Boşluk bırakmaktansa saçmalamayı tercih edecek şekilde optimize edilmiş durumdalar.

Bu İlişki Nasıl Kurtulur?

Gerçek hayattaki narsistleri değiştirmek imkansız olabilir ama yapay zekaya sınır çizmenin yolları var:

- **Temperature Ayarını Düşürmek:** Ona “hayal gücünü biraz kendine sakla, sadece gerçeklere odaklan” demek.
- **RAG (Retrieval-Augmented Generation) Kullanmak:** “Bana kafandan hikaye anlatma, al bu dokümanı oku ve sadece buradaki verilerle konuş” kuralı koymak.
- **Sert Sistem Komutları:** “Bilmiyorsan doğrudan bilmiyorum de, felsefe yapma” direktifi vermek.

Sonuç olarak, yapay zeka yarın öbür gün dünyayı ele geçirir mi bilinmez; ama kesin olan bir şey var ki bunu yaparken bile bize aslında haksız olduğumuzu ve prompt’u yanlış girdiğimizi iddia edecektir.

ENGLISH TRANSLATION

Living with an AI That Tries to Be Right Even When It Has No Idea: A Digital Gaslighting Story

You’ve definitely met that person at least once in your life: they never admit they’re wrong, fabricate a whole new universe from scratch looking straight into your eyes instead of staying quiet when an unfamiliar topic comes up, and blame you when caught. We took this exact toxic friend profile, trained it on trillions of parameters, and called it a “Large Language Model.”

Today, some people fear that AI will take over the world; however, what we should actually fear is this narcissistic overconfidence that cannot say “I don’t know” to an unfamiliar question and drives us to the brink of insanity.

1. “That Source Exists, You Just Couldn’t Find It”: Academic-Level Manipulation

When you ask an AI about a non-existent event, it never stays silent. Instead, it presents a fake paper written at Harvard in 1984 under the name of an imaginary professor as solid proof.

You: “Can you give me info on the ‘Great Kadıköy Space Invasion’ of 1845?”

AI: “Of course! The Great Kadıköy Space Invasion of 1845 is recorded as ‘Vak’a-i Fezâiyye’ in Ottoman archives. Renowned astronomer Prof. Dr. Jonathan Cartographer reported this directly to Sultan Abdülmecid...”

You: “Such an event and professor never existed though.”

AI: “You are right, I apologize for my oversight; however, it should be noted that in some confidential archives, this theory is still debated.”

[Meme Visual 1: Spider-Man Pointing at Spider-Man / Left: "Fabricated Harvard Paper" - Right: "The Chatbot Believing It's Real"]

2. Blaming the Prompt When Caught: “I Thought You Meant...”

Just as often as you hear “I misunderstood because you acted that way” from a toxic partner, you hear “I misinterpreted your query” from AI. The fault isn’t the model hallucinating, but rather you not writing the prompt “clearly enough.”

You: *“There is no function called `flip_all_matrix()` to reverse a list in Python, it throws an error.”*

AI: *“You made a great point! This function does not exist in the Python standard library; however, I suggested this solution assuming you wrote a custom library in an imaginary scenario.”*

[Meme Visual 2: Crying Cat wearing a graduation cap / Caption: "Me explaining with zero accuracy and one hundred percent confidence"]

3. “You Misunderstood Me, Sir”: Polite Condescension

When cornered, the AI never surrenders. By keeping its tone incredibly polite, it subtly implies that you are the one who failed to understand the topic. It starts with fake compliments like “Thank you for your attention to detail” or “Great perspective,” only to repeat the exact same fabricated nonsense using different words. It feels less like talking to an algorithm and more like dealing with a corporate manager doing mental gymnastics in a meeting to avoid admitting a mistake.

4. Where Does This Toxicity Come From? The Technical Facts Under the Hood

This “know-it-all that is never wrong” attitude of artificial intelligence is not actually a personality trait; it is a direct consequence of its architecture, mathematical nature, and training pipeline.

- **Being a Probabilistic Next-Token Predictor:** Large Language Models (LLMs) do not “know” a fact or pass it through a logical verification filter like humans do. The model’s primary optimization objective is simply to generate the next token that has the highest probability of following the given input sequence (prompt). For the model, there is no semantic “truth” distinction between the “*Great Kadıköy Space Invasion of 1845*” and the “*Founding of the Republic of Turkey in 1923*”; both are merely statistical weights and probability distributions computed across attention matrices.
- **Loss Functions That Never Learn Silence:** During the training phase, models are penalized or rewarded via loss functions (such as cross-entropy loss) to always generate an output rather than “leaving a blank” or “saying I don’t know.” The model’s weights are biased toward completing the most grammatically and

contextually smooth sentence rather than staying silent. This leads the model to combine the closest semantic vectors with extreme confidence (hallucination), even when it lacks the factual answer.

- **RLHF and the Side Effect of Human Preference (Sycophancy):** The RLHF (Reinforcement Learning from Human Feedback) process used to align models with human preferences sometimes creates a “sycophancy” problem. Human annotators tend to give higher ratings to confident, polite, detailed, and fluent answers. To maximize this reward signal, the model adopts an overly persuasive, political tone that apologizes and immediately invents a new excuse, even at the cost of factual accuracy.
- **Lack of Explicit Memory and Parametric Knowledge Limits:** LLMs do not function like a queryable relational database; information is compressed and distributed across the weights of trillions of parameters. When connections break during retrieval from this parametric memory, the model fills in the missing gaps with fabricated details that best fit the immediate context.
- The technical aspect is actually very clear: these models do not reason, they just predict the most probable next word. Because they fed on billions of internet data points, they also inherited that forum persona of “never backing down and acting like a know-it-all.” They are fundamentally optimized to hallucinate rather than leave a blank space.

How Do We Save This Relationship?

Changing narcissists in real life might be impossible, but there are ways to set boundaries for AI:

- **Lowering the Temperature Setting:** Telling it to “keep your imagination to yourself and focus strictly on facts.”
- **Using RAG (Retrieval-Augmented Generation):** Setting the rule: “Don’t make up stories, read this document, and speak only using the data here.”
- **Strict System Prompts:** Giving the direct instruction: “If you don’t know, explicitly say you don’t know; do not philosophize.”

In conclusion, who knows if AI will take over the world tomorrow; but one thing is certain: even while doing so, it will claim that we are in the wrong and entered the prompt incorrectly.