Chapter 3 — Dark Brains: Criminal Exploitation of AI Models

[Note this chapter was written largely by ChatGPT 5.0 with source and saliency confirmation by Michael J. McCarron, under a query outline provided by the human author.]

Artificial Intelligence has become a double-edged sword – just as cybersecurity defenders harness AI, threat actors are weaponizing it. In particular, **Large Language Models (LLMs)** with removed safety filters – so-called **"dark LLMs"** – are now being used to **automate and scale cybercrime**. Unlike mainstream chatbots that refuse illicit requests, dark LLMs will happily generate malware code, phishing lures, or illicit tradecraft on demand[1][2]. These models operate *without guardrails*, giving criminals a way to get "good answers to bad questions" for hacking and fraud tasks[3][4]. Security researchers warn that this **unfettered AI** is lowering the skill barrier for cybercrime and supercharging malicious campaigns in real time[5][6].

Crucially, dark LLMs are often built on **open-source models** (or jailbroken versions of commercial ones) fine-tuned with malicious data[7][8]. They are marketed on underground forums and darknet marketplaces as **"AI-as-a-service" for criminals**, with subscriptions granting access to these uncensored chatbots[9]. Because no central provider monitors their output, **black-market LLMs have no oversight** – a stark contrast to the tightly filtered APIs of OpenAI or Google. This makes them attractive for threat actors seeking anonymity and unrestricted capabilities. Below we dive into the development of dark LLMs, the major examples in circulation, their uses and threats, and how both criminals and nation-states (like Russia) are leveraging these tools.

What Are "Dark LLMs"?

Dark LLMs are AI models with their safety restraints removed. In essence, they are large language models **devoid of alignment** and content filters, meaning they will produce **any output a user asks for** – including disallowed or illegal content[10][1]. Mainstream LLMs (ChatGPT, Bard, etc.) have guardrails to reject requests for hate speech, hacking advice, violent plans, and so on. Dark LLMs eliminate those safeguards, either by using **uncensored open-source models** or by **jailbreaking** proprietary ones[11][12]. The result is an AI that will freely assist with *hacking, fraud, or other crimes* without ethical restrictions[13][14].

These malicious models often originate from publicly available AI. Developers take an open model (such as Meta's LLaMA or EleutherAI's GPT-J) and fine-tune it on **malware code, hacking tutorials, and other illicit data**[15]. Some dark LLM operators don't even bother training their own model – instead they provide a **wrapper around an existing model** (like a **jailbroken ChatGPT** or an uncensored fork of an open model) [16][17]. In either case, the AI is *re-purposed explicitly for malicious use*. As a 2025 Barracuda report puts it, dark LLMs "provide attackers a leg up" by identifying vulnerabilities, writing exploits, and crafting phishing content that normal AI would refuse to produce[18][19].

Notably, almost **any LLM can be turned "dark"** via prompt exploits if one is clever – an ongoing "jailbreak arms race" exists between attackers and AI providers[20][21]. But the dark LLMs we focus on here are **custom or openly distributed tools** deliberately built *without any guardrails* from the start. They are typically **sold on the dark web** or shared in criminal circles, often touting features like *no logging of user activity*, *fast uncensored responses*, and *illicit capabilities* (malware generation, carding assistance, etc.)[22][23]. In short, a dark LLM is **an "evil twin" of ChatGPT** – the same powerful language generation, but pointed toward unethical tasks.

Timeline of Criminal Dark LLMs

The concept of criminals exploiting AI for nefarious ends is not entirely new, but it accelerated rapidly in the last couple of years. Here's a brief timeline of how dark LLMs developed into their current state:

- **Early 2022:** OpenAI's ChatGPT launched and soon **cybercriminal forums lit up with interest**. At first, criminals attempted to **jailbreak ChatGPT** and other public bots to output banned content. By early 2023, threads on dark web forums shared jailbreak prompts and "DAN" (Do Anything Now) techniques to make ChatGPT produce malware or phishing text[24][25]. This was unreliable and limited – OpenAI kept patching prompts – but it proved the *appetite* for AI-assisted crime.

- **Mid 2023:** Realizing the need for an uncensored alternative, enterprising hackers began **rolling out custom LLM chatbots**. One of the first was **WormGPT**, revealed around July 2023[26]. WormGPT's developer (alias "LastLaste") took the open-source **GPT-J** model (6 billion parameters) and trained it on **malware code and cybercrime data**[27][28]. They started selling access to WormGPT on hacking forums for **€60–€100 per month (or €550/year)**[7]. WormGPT set the template: an **English-language chatbot** with *no ethical limits*, marketed as "ChatGPT for blackhats." It could write keylogger malware, craft convincing phishing emails, and generally answer any illegal query[15].

- **Late 2023:** Following WormGPT's buzz, **copycats and "improvements"** emerged. July 2023 saw **FraudGPT** advertised on dark web markets and Telegram[29]. FraudGPT's seller ("CanadianKingpin") billed it as an **"all-in-one" criminal AI toolkit** with capabilities from malware writing to phishing page generation to finding software vulnerabilities[30][31]. They claimed thousands of sales and charged higher prices ($200/month or $1700/year) for access[32][33]. Around the same time, mentions of **"DarkBard"** (a malicious version of Google's Bard) and **"DarkGPT"**/"DarkBERT" circulated, though some of these were likely scams or exaggerations[34][28]. Researchers also spotted niche offerings like

WolfGPT and **XXXGPT**, purportedly targeting tasks like cryptographic malware and botnet control[35][36].

- **2024:** As generative AI hype grew, so did the *underground interest*. New dark LLM brands appeared. By late 2024, a Telegram-based bot called **GhostGPT** gained traction as a **cheap ($50/week) uncensored AI service**[37][38]. GhostGPT built on the lessons of earlier models – it promised *fast responses, no logging, no jailbreaks needed*, and marketed itself for malware dev and BEC (business email compromise) scams[39][37]. Researchers noted GhostGPT might just be a wrapper around a jailbroken mainstream model, but its popularity spiked on forums, indicating demand for "plug-and-play" dark AI[40][41]. Throughout 2024, criminals also began leveraging **open-source "uncensored" models** (like LLaMA 2 Uncensored, WhiteRabbit Neo, etc.) which can be run locally[42][43]. Discussions on top forums (e.g. XSS, Exploit) included tutorials to build your own private GPT and attacks on AI systems[44]. This period also saw academic demonstrations like **PoisonGPT**, where researchers edited an open model to **embed disinformation** – a warning that open LLMs could be twisted for propaganda[45].

- **2025 and Beyond:** Dark LLM development continues to accelerate. **New variants of WormGPT** have appeared built on cutting-edge bases like Mistral's models and even Elon Musk's **xAI "Grok"**, offering improved performance[16]. Cybercrime tools are increasingly integrating AI; for example, a 2025 malware dubbed **"LameGravity"** (or *LameHug*) used a built-in LLM (Alibaba's Qwen model) to dynamically generate hacking commands on victim machines[46][47] – essentially malware with an AI "brain" for on-the-fly decisions. Underground chatter indicates that more **bespoke criminal AI** projects are underway, often kept private within groups. We are likely at the cusp of an arms race where *attackers no longer need advanced coding skill*, just the budget to rent a malicious AI that will do the heavy lifting. As one security CEO noted, *"AI has transformed cybercrime from a game of skill to a game of scale"*, dropping the cost and effort of launching attacks dramatically[48][49].

Notorious Dark LLMs and Capabilities

Several dark LLMs have gained notoriety on the black market. Below is a briefing on **known (and rumored) malicious AI chatbots** and what each brings to the table:

- **WormGPT:** *The original "blackhat GPT," based on GPT-J 6B*. WormGPT's developer trained it on malware and hacking data, resulting in a chatbot that **writes exploit code, crafts phishing emails, and answers any cybercrime question**[15]. First sold in mid-2023 on forums for ~€100/month, it quickly became a **tool of choice for Business Email Compromise (BEC)** scammers to generate convincing English emails[27][50]. WormGPT v2 was later offered with upgrades like code formatting, multi-language support, and even the ability to switch underlying models[7][51]. Essentially, WormGPT can do everything

ChatGPT refuses to – from producing ransomware strains to giving step-by-step hacking advice.

- **FraudGPT:** *An "all-in-one" fraudster's AI*, spotted by researchers in July 2023[52]. Advertised on Tor forums and Telegram, FraudGPT claims it can **write undetectable malware, create phishing websites, generate scam text messages, find vulnerabilities, and even teach you to hack**[53][54]**.** Its dark web landing page boasted "no boundaries" and thousands of successful sales[55][56]. Price points ranged from ~$90/month up to several hundred for longer subscriptions[29]. There's evidence the same actor behind WormGPT is involved with FraudGPT, suggesting a suite of "evil GPT" products[57]. However, later investigation by Cisco Talos found the FraudGPT service was likely a **scam** – the seller took crypto payments but provided non-working credentials[58]. Still, the *idea* of FraudGPT spurred copycats and demonstrates the market demand for AI-driven fraud tools.

- **DarkBard:** A malicious variant of Google's **Bard** chatbot. DarkBard was mentioned in mid-2023 as being peddled on forums[59]**.** It purportedly offered similar features to FraudGPT (malware, phishing generation) but built on Google's model. It's unclear if DarkBard was ever a functional product or just a buzzword used by a FraudGPT scammer (who claimed to have it)[34]. Regardless, the concept is plausible: fine-tune or jailbreak Google's LLM to remove safeties, and you'd have "Bard gone bad" – potentially powerful given Bard's resources.

- **WolfGPT:** Another entrant in late 2023, described as an **"alternative to ChatGPT minus guardrails"**[60]. WolfGPT was reportedly coded in Python and offered "complete confidentiality" for users, focusing on **cryptographic malware creation and advanced phishing**[35]. It didn't gain as much traction, possibly overshadowed by others. But it illustrates that multiple actors were attempting their own "evil GPT" brands around the same time.

- **GhostGPT:** A **Telegram-based uncensored chatbot** that rose to prominence by end of 2024. GhostGPT is marketed as a **user-friendly crime AI** – no need to set up any model or prompts; you pay the fee and chat with it live on Telegram[37]. According to Abnormal Security, GhostGPT can generate **polymorphic malware code, exploit scripts, and highly personalized phishing emails** with ease[61][62]. Its devs advertise *fast responses and zero logging*, appealing to criminals concerned with speed and secrecy[23][63]. GhostGPT's pricing (around $150/month) undercuts earlier services, and it garnered thousands of views on forums – indicating significant interest[37][64]. Researchers suspect GhostGPT might be using a **jailbroken GPT-4 or similar model under the hood** rather than a wholly new LLM[17][22], but from an end-user perspective it doesn't matter – it's a one-stop *"write me malware now"* bot. By early 2025, GhostGPT and its ilk are considered a **new and growing problem** for security teams[65][66].

- **Others (DarkGPT, DarkestGPT, EscapeGPT, etc.):** The dark web has seen a flurry of other names. **DarkGPT** was advertised on Telegram as an "AI assistant" for querying hacked databases and doing OSINT on leaked data[67]. **DarkestGPT** showed up on a Tor site with subscription pricing in Bitcoin, offering tools and "AI insight" for carding and hacking ops[68]. **EscapeGPT** was noted as yet another variant that uses clever prompt engineering to *escape* safety filters[69]. Many of these fringe projects never gained wide use or turned out to be repackaged versions of existing models[70]. However, their proliferation underscores an important point: **the barrier to creating a custom LLM is low** (open models + some coding), so we can expect many small threat actors to experiment with making their own *"[____]GPT"* for specialized purposes. Security firms have observed forum posts sharing scripts and datasets to facilitate exactly that[71][72].

Figure: Underground advertisement for WormGPT on a Russian cybercrime forum (2023). This screenshot (from Trustwave) shows the seller marketing WormGPT as a "ChatGPT alternative for blackhat" use, with no ethical limits, privacy features, and subscription plans (e.g. €100/month or €550/year)[7][73]. The emergence of WormGPT marked the start of a trend of criminals offering custom AI chatbots as services to others.

How Criminals Use Dark LLMs

Unfettered LLMs have quickly become "force multipliers" for a range of cybercriminal activities, such as ransomware, phishing, etc. Some **key uses and threats** posed by dark LLMs include:

- **Phishing and Social Engineering at Scale:** One of the clearest advantages is writing **fluent, persuasive phishing messages** in any style or language. Dark LLMs can generate *business email compromise (BEC)* scam emails that are **remarkably convincing and strategically worded**, even mimicking a CEO's tone to fool employees[74][28]. They eliminate the tell-tale grammar mistakes that often give away foreign scammers. Criminals also use them to craft **spear-phishing** content tailored to individual targets, pulling details from LinkedIn or breaches and having the AI weave them into personalized lures. A jailbroken or custom model will even output **harassing or coercive language** that legitimate bots would block – useful for extortion emails and impostor scams. According to Rapid7, AI has reduced the cost and effort of phishing and social engineering by up to *95%*, shifting these attacks from low-volume artisan efforts to **high-volume campaigns**[49]. Even multilingual phishing becomes trivial – attackers can prompt the LLM to produce convincing scams in Spanish, French, Chinese, etc., broadening their victim pools.

- **Malware and Exploit Development:** Dark LLMs serve as a **tireless malicious coder** on demand. Need a ransomware program that evades antivirus? Or a script to scan for a specific vulnerability? These models can produce functional code for viruses, keyloggers, backdoors, you name it. WormGPT, for instance, has been used to write **polymorphic malware** – malicious code that the AI can continually mutate (change signatures, obfuscate sections) to evade detection[75][76]. FraudGPT's marketing boasted of "millions of phishing email examples" and "6,000+ malware source code references" built-in[77]. Some models claim to **find exploits** as well: by inputting a snippet of code or an app's description, an uncensored LLM might suggest potential vulnerabilities or even generate a proof-of-concept exploit. While current AI still has bugs, skilled hackers use it to **accelerate bug discovery and development** – essentially outsourcing a junior malware developer. Notably, even less-skilled criminals can now create dangerous software by simply describing what they want (e.g. "a virus that steals PDFs and Excel files and exfiltrates via FTP") and letting the AI handle the syntax. This raises the specter of **more malware, from more sources**, overwhelming defenders.

- **Crimeware Automation & "AI Agents":** Beyond writing static code, criminals are exploring **LLM-powered agents** that automate entire workflows. A dark LLM hooked into tools can act as an **offensive assistant** – for example, scanning a list of stolen credit card numbers and automatically testing which are valid, or controlling a botnet's actions via natural language commands[13][78]. There are reports of dark LLM services integrating with **email systems, vulnerability scanners, and carding APIs** to provide one-stop automation[79][78]. This means a single AI could coordinate tasks like: find vulnerable websites, craft exploit payloads, dispatch phishing emails, and process the stolen data – essentially running a **personal cybercrime campaign** with minimal human oversight. While such "agentic AI" is in early stages, security experts warn it's the *"beginning of AI-driven cyberwarfare"* and could lead to semi-autonomous malware that adapts to environments on the fly[80][81]. An example is the **LameHug malware (2025)** which embedded a large language model (Qwen 32B) inside; once on a victim's PC, it used AI to dynamically generate commands for data theft and system exploration[46][47]. This **adaptive AI malware** is harder to predict and may adjust its tactics per victim, making infections more dangerous and stealthy.

- **Fraud, Social Scams, and Other Crimes:** Dark LLMs are not limited to pure hacking – they also assist in **financial fraud and "social" crimes**. For example, they can generate **fake identities and scripts** for scam call centers, write compelling romance scam messages, or produce deepfake text for impersonation. In underground markets, criminals discuss using AI to automate **investment fraud** (e.g. writing a convincing whitepaper for a fake cryptocurrency, or mass-producing pump-and-dump stock tips on forums). The **"insider trading plans"** or other complex schemes that a savvy fraudster might conceive can potentially be brainstormed by an AI given enough data. **Spam and disinformation** for profit are also in play – using LLMs to auto-generate thousands of posts advertising counterfeit goods, phishing links, or fraudulent services. Essentially, any scam that involves convincing a human at scale (through text, email, chat) can be turbocharged by an LLM's ability to tailor and churn out content in volume. We have already seen cybercriminals bragging about custom AI models to write **fake websites and scam pages** that look professionally made[53]. Combine that with AI's talent for **mimicry** – e.g. copying writing styles or even coding fake legitimate sites – and the line between genuine and fraudulent online content blurs further.

- **Evasion of Detection:** Interestingly, some dark AI tools advertise features to **evade security measures**. For instance, the FraudGPT page listed "code obfuscation" and automated creation of polymorphic payloads as features[82]. LLMs can help criminals refine their output to slip past filters – whether it's rephrasing phishing text to avoid spam triggers, or encoding malicious code in novel ways to evade antivirus. Uncensored models will also cheerfully give advice on how to avoid law enforcement stings or encrypt communications. All this means attacks assisted by AI may be **harder to detect** through traditional defenses. Already, corporate security teams note that AI-generated phishing emails often *bypass legacy email filters* because they read as perfectly benign prose, not the common bad grammar and keywords those filters flag[83][84]. It's an AI vs AI cat-and-mouse game now – with malicious AI generating ever more human-like and varied artifacts, forcing defensive AI to work harder on pattern recognition.

In summary, dark LLMs empower **more attackers to do more damage with less effort**. A novice with a few hundred dollars can unleash a credible phishing campaign or write a new malware strain – things that used to require a skilled team. And an experienced criminal can use AI to **amplify their reach and sophistication**, juggling more attacks than ever before. This democratization of "evil expertise" is precisely why law enforcement and cybersecurity professionals are alarmed. The **threat surface is exploding**: more phishing, more malware, more fraud, potentially at a pace and scale we haven't dealt with before[49][85].

Black Market Ecosystem and Trends

Dark LLMs have given rise to a small but vibrant **black market** ecosystem. Understanding where and how these tools are distributed can help investigators know where to look:

- **Underground Forums:** Much of the action happens on infamous hacking forums (both clearnet and dark web). For example, WormGPT was initially sold via posts on **HackForums (English forum)** and later on **Exploit** (a top Russian-language forum)[7][86]. Forum posts often include screenshots demonstrating the AI's capabilities (e.g. WormGPT writing malware or phishing emails) to entice buyers[87][88]. There are dedicated sections on some forums for **AI and ML** topics, where members exchange jailbreaking tips, share open-source model links, or even post code to build one's own GPT[44][89]. Key forums of interest include XSS (Russian), Exploit, Breach Forums (pre-2023 takedown), HackForums, and newer communities where criminals congregate. Investigators monitoring these forums have observed **users offering "private AI" services**, essentially freelancing their custom model or prompt skills to others.

- **Dark Web Marketplaces:** Some dark LLMs have appeared on Tor hidden service marketplaces – the same sites that sell drugs, stolen data, and hacking tools. For instance, **FraudGPT was advertised on at least two Tor markets** in mid-2023[29][90]. The listing touted it as a "*ChatGPT alternative with no limits*" and gave pricing options in crypto. Likewise, GhostGPT was initially promoted on a dark web site before shifting to Telegram sales[91][92]. These marketplaces sometimes provide escrow, but as with any illicit product, scams are rife – multiple buyers reported being scammed by the FraudGPT seller, who took payment without delivering a working product[58]. This underscores that **trust is a commodity** even among criminals; reputable sellers or those with a history on forums (as WormGPT's dev had[93]) tend to attract more customers.

- **Telegram and Messaging Apps:** A noticeable trend is the move to **Telegram channels** and bots for selling access. The developer of FraudGPT maintained a Telegram account to handle subscriptions (likely to avoid marketplace fees and exit scams)[32][94]. GhostGPT, as noted, is itself delivered via a Telegram chatbot interface[37]. Telegram is popular in cybercrime circles for its relative anonymity and ease of use. We see criminals advertising their AI bots on Telegram channels, providing updates, and taking payment directly (often in cryptocurrency). This complicates law enforcement's job, as the transactions become **peer-to-peer and ephemeral** (channels can be deleted or moved quickly). Other messaging apps like Discord or ICQ have also been rumored for sharing AI tools, but Telegram appears to be the primary venue currently.

- **Pricing and Monetization:** The **price points** for dark LLM services give a sense of their value in the underground. WormGPT v1 started at €100/month[7]; WormGPT v2 was advertised at €550/year, with a "private build" for €5000[95]. FraudGPT ranged from ~$200/month up to $1000+ for longer terms[29], and GhostGPT offered even shorter trials at $50/week[37]. These prices are non-trivial, suggesting that criminals believe the ROI (e.g. from successful scams or breaches enabled by the AI) is worth it. The subscription model also indicates a **Cybercrime-as-a-Service** approach – rather than selling the model itself, sellers keep control and rent out usage, possibly to prevent leaks of the model weights. There's also chatter about **private bespoke models**: for a higher fee, some developers will fine-tune an AI specifically for a client (for example, trained on data targeting a particular industry's systems). This mirrors the way bespoke malware is developed for high-end clients, and could lead to *"boutique AI"* services for organized crime.

- **Quality and Authenticity Issues:** It's worth noting that not all dark LLMs are as capable as advertised. Security analyses have found that many are **just repackaged open models or slightly jailbroken versions of public APIs**[70]. For instance, EscapeGPT was basically ChatGPT with clever prompts, and GhostGPT might be hooking into an existing model in the backend[17]. The **lack of transparency** (no one discloses their model architecture or training data) makes it hard to assess each tool's true sophistication. Additionally, as mentioned, some offerings are outright **scams targeting other criminals** – a longstanding tradition in the dark web (scammers scamming scammers)[96][97]. In one case, the "developer" of FraudGPT simply disappeared after taking payments, hinting that they never had a real model[58]. Nonetheless, enough of these tools *do* exist and function that the threat is not imaginary. Even if a criminal doesn't want to pay for a dubious service, they can always **roll their own model** using open-source weights and community-released "uncensored" datasets[98][43]. The barrier to entry for DIY is perhaps a decent GPU and some know-how – which well-funded gangs have in abundance. In short, the dark market for LLMs is a Wild West: **rapidly evolving, somewhat scam-ridden, but increasingly embedded in the cybercriminal toolkit**.

Influence Operations and Disinformation

Beyond hands-on cybercrime, there's growing evidence that **LLMs without guardrails are being used in influence campaigns** – by both state actors and criminal groups. These AI systems can produce convincing propaganda, fake personas, and automated content at a volume that humans could never match, potentially supercharging disinformation efforts.

State-Sponsored Influence: 2024 marked a turning point where Western authorities openly identified generative AI in foreign influence ops. In July 2024, the U.S. Department of Justice revealed it had disrupted a **Russian government-backed propaganda campaign** that leveraged **an "AI-enabled" bot farm**[99][100]. According to court documents, a private Russian outfit (with Kremlin funding) built a custom AI platform to create and run *fake social media accounts* — complete with AI-generated profile pictures and posts — that pushed pro-Kremlin narratives to American and European audiences[101][102]. Over 1,000 bot accounts were part of this network, and they were *indistinguishable from real users*, even mimicking real U.S. citizens and spreading tailored propaganda about the Ukraine war and other topics[103][102]. This is believed to be the first publicly confirmed case of a nation-state using generative AI for online influence at scale[104]. The AI platform handled content creation and account management, essentially automating a troll farm. The incident underscores how **regimes like Russia are experimenting with LLM-driven influence** – amplifying their disinformation playbooks by generating more content, more quickly, and with plausible authenticity. Western officials have warned that as AI models improve, adversaries will use them to **"rapidly scale up" misinformation efforts** and make fake news campaigns harder to detect[105][101].

Russia is not alone; Chinese influence operations have also been observed adopting generative AI. A 2023 analysis by The Diplomat noted **China-linked spam networks using AI-generated text and deepfake images** to bolster Beijing's narratives on social media[106]. The UK's CETaS (Centre for Emerging Technology and Security) report likewise flagged **Chinese frontier AI innovations** (many open-source) as a boon for criminals and propagandists, since these **"open-weight" models come with fewer guardrails to prevent misuse**[107]. In essence, **unrestricted LLMs enable authoritarian actors to flood information spaces** with convincing fake content – whether that's political propaganda, fake grassroots comments, or forged documents – at a scale and customization level that was previously impossible.

Criminal-Driven Disinformation: It's not just governments – **criminal gangs for hire** can run influence or manipulation campaigns as a service, and they too are turning to AI. This overlaps with cybercrime in cases like extortion or stock manipulation. For example, a criminal crew might be paid to smear a business rival or pump a cryptocurrency – tasks that involve blasting out misleading content and engaging with targets. LLMs can make this far easier: auto-generating hundreds of blog posts, social media comments, or even fake "leaks" to support a false narrative. Already, we've seen **fake news-for-hire services** on the dark web, and adding AI would allow them to scale output while maintaining coherence. There is reporting that **Russian cybercriminal groups sometimes undertake disinformation jobs** on behalf of oligarchs or state-linked clients[108][109]. With AI, these groups could amplify hate speech, election interference, or social discord campaigns at a fraction of the manpower previously needed.

One specific area is **deepfake text and media**: criminals can use LLMs to generate scripts for deepfake videos or create chatbots that impersonate people online. In 2023, for instance, cybercriminals used AI to clone the voice of a company's CEO and nearly pulled off a fraudulent funds transfer by calling a subordinate[110]. While that was voice (deepfake audio), the *script* and setup for such social engineering can be optimized by LLMs. Looking forward, we anticipate **fake persona networks** run by criminals using LLMs to respond in real-time on social platforms, engaging in conversation and persuasion – essentially botnets of "social clones" that are hard to distinguish from passionate humans. For law enforcement and OSINT analysts, this means the usual signs of inorganic activity (repetitive phrasing, same mistakes) might vanish as each AI agent produces unique, human-like output.

It's also worth mentioning **terrorist and extremist propaganda**. There's concern that non-state extremist groups (or lone actors) will leverage open AI models to produce recruitment material, fake manifestos, or how-to guides for attacks. Normally, ChatGPT would refuse requests to glorify terrorism or give bomb-making instructions – but an offline uncensored model would comply. We have already seen **AI being abused to generate child sexual abuse material descriptions and other heinous content** in underground circles[111]. The implication is that *any form of harmful influence or content generation can and will be tried* with these models.

In summary, **influence operations have entered the AI era**. State actors like Russia and China are actively exploring LLMs to turbocharge their propaganda and social manipulation, often by **outsourcing or leveraging criminal networks as proxies**. And conversely, cybercriminal organizations are diversifying into information warfare tactics, using the same dark LLM tools to sow confusion for profit or on contract. This convergence means investigators must watch not only for malware and hacks, but also for subtler AI-generated influence campaigns in forums, social media, and fringe websites. The line between traditional cybercrime and information warfare is blurring, with **"dark AI" sitting in the middle as an accelerant**.

Russia's "Shadow Alliance" with Criminal Hackers

Russia presents a particularly notable case of a state merging forces with cybercriminals in the context of AI and cyber operations. **Russian intelligence agencies have long collaborated with, protected, or co-opted criminal hacker groups** – a relationship often described as a "**shadow alliance**"[109][108]. This arrangement allows the Kremlin to **outsource dirty work** and maintain plausible deniability, while the criminals receive resources and a degree of impunity (as long as they don't target domestic interests).

Europol's 2025 organized crime threat assessment highlighted how **Russian state actors leverage organized crime networks** to destabilize targets in Europe[112][108]. These proxies carry out everything from cyber-attacks and data theft to sabotage and smuggling, effectively acting as extensions of state power[113][114]. Cybercrime gangs based in Russia (e.g. ransomware crews) are often left untouched by Russian law enforcement and are suspected of moonlighting for state-directed missions when called upon[115][109]. A Guardian investigation noted that even if the Russia-Ukraine war were to end, Russian "criminal groups will continue to exert influence" and likely increase black market activities like weapons trade and cyber aggression[116].

In the realm of AI and LLMs, this means **Moscow can tap its pool of cybercriminal talent to develop and deploy dark AI tools**. The **LameHug malware** example is instructive: APT28 (Russian military intelligence) created malware using a Chinese open-source LLM (Qwen) to dynamically execute tasks[46][47]. It shows the willingness to integrate AI into state hacking tools. Now consider that many top ransomware and banking trojan gangs (Evil Corp, TrickBot, REvil, etc.) operate from Russia – these groups could serve as guinea pigs or collaborators for LLM-powered cyber attacks. For instance, a ransomware gang could adopt an LLM to generate more effective phishing lures to gain initial access, or to write custom exploits for each victim's environment. In return, if the FSB or GRU needs an influence campaign or a disruptive attack, they could task these criminals to leverage their AI capabilities for Mother Russia. Western officials have publicly accused Russian security services of *tasking criminal hackers* to carry out attacks on targets like infrastructure or political enemies[117][118]. With AI in the mix, we might see state messaging campaigns coordinated with criminal-run botnets or AI-driven spam networks, blurring who is behind the keyboard.

A concrete example of this synergy was the **Russian AI propaganda bot farm** dismantled by the DOJ in 2024 (mentioned earlier). It involved not only state agents (including an RT employee) but also likely contractors who built the AI platform and managed the bots[101][102]. It wouldn't be surprising if some of those technical experts had roots in the cybercriminal underground – the skillsets overlap (data harvesting, AI modeling, social media manipulation). Indeed, some darknet forums in Russian have threads where users with AI expertise (like machine learning engineers) offer their services, which could be quietly leveraged by state-tied actors.

Additionally, Russia has a history of using **hacktivist fronts and patriot hacker groups** that are essentially criminal actors given political direction. Groups like "KillNet" (a pro-Russian hacktivist group) have engaged in disruptive attacks on Western sites. These groups could incorporate generative AI for greater impact – e.g. automating the creation of fake news posts during an attack to amplify panic, or using LLMs to rapidly translate propaganda to multiple languages when targeting international audiences. We saw a hint of this with reports that **Iran-aligned actors used AI-generated text messages and deepfaked alerts to incite panic in Israel** during conflict[119] – a tactic Russia could certainly mirror via its proxy groups.

In essence, **Russia's fusion of state and criminal cyber capabilities extends to AI**: the Kremlin can utilize criminal-developed LLM tools for its own operations, and conversely, provide safe harbor and data to criminals experimenting with AI. The "shadow alliance" means advances in dark LLMs within the Russian cybercrime ecosystem can quickly find their way into state-sponsored campaigns. This makes the threat extremely agile and hard to attribute – is a given AI-generated phishing campaign just financially motivated, or an espionage operation, or both? It could be *all of the above*. Intelligence officers and investigators should be aware that any significant Russian cybercrime actor dabbling in AI might be doing so with a wink and nod from Russian authorities. Conversely, when analyzing Russian disinformation or cyber attacks, one should consider the potential involvement of **off-the-shelf criminal AI services** behind the scenes.

Conclusion

The rise of dark LLMs represents a **new chapter in cybercrime and security**. These unrestricted AI models, fine-tuned for malice, have lowered the entry barriers for cybercriminals and opened fresh avenues for state-sponsored attackers. In just the past two years, we've seen a proliferation of illicit chatbots – from WormGPT and FraudGPT to GhostGPT and beyond – **enabling everything from mass phishing and malware engineering to automated propaganda**. While some of these offerings are hyped or fraudulent, the underlying trend is real: **powerful language models are now in the hands of threat actors** who operate outside any ethical or legal constraints.

For cybersecurity professionals, police investigators, and intelligence officers, this evolution poses several challenges. We must **update our threat models** – AI-driven attacks mean more volume and sophistication. Phishing emails can no longer be dismissed for bad grammar; malware may morph its signature faster than IOC feeds can keep up. Traditional defenses will catch fewer low-hanging threats as criminals move to AI-curated tactics. At the same time, investigators have new leads to monitor: illicit AI services leave traces (forum posts, Telegram channels, crypto transactions) that can be infiltrated or analyzed. It will be crucial to **track the marketplaces and communities** where dark LLMs proliferate – the HackForums, XSS, Exploits, and emerging venues that serve as bazaars for these tools. Law enforcement might consider undercover buys of AI services to gauge their true capabilities (with the caveat that many sellers scam). Intelligence sharing between agencies is also vital, since an AI tool used for crime in one country could be repurposed for espionage in another.

On the flip side, defenders are not powerless – the community is already deploying **defensive AI** to counter malicious AI. Email security vendors use AI to detect the subtle signals of AI-written phishing[120]. Researchers are developing methods to watermark or identify AI-generated text, which could help flag suspicious content floods. And companies like OpenAI are continuously improving guardrails to make jailbreaks harder (forcing criminals to use their own models at greater expense). Gartner predicts that by 2026, organizations that integrate GenAI into security awareness will see significantly fewer successful social engineering incidents[121][122] – basically using AI to bolster human vigilance. In short, **AI will be fought with AI**, and security teams need to embrace that reality quickly.

Finally, the involvement of **nation-states like Russia leveraging criminal AI** means this is not just a technical issue but a geopolitical one. The use of dark LLMs in influence operations blurs the line between cybercrime and information warfare. We may need new norms or even deterrence strategies for AI misuse – much like chemical or biological agents, AI could be seen as a dual-use technology requiring international oversight when it comes to malicious deployment. The UK's Alan Turing Institute (CETaS) has called for an *AI Crime Taskforce* and proactive measures to **"raise barriers to criminal adoption" of AI tools**[123][124]. This might include everything from AI monitoring on darknet forums to legal consequences for creating pernicious models.

In conclusion, *dark LLMs have arrived* and are evolving fast. Cybersecurity professionals must stay informed about the latest "evil AI" tools circulating in the underground, understand their capabilities, and adjust defenses accordingly. Law enforcement and intel agencies should recognize that the old playbook of chasing lone hackers is now complicated by **AI systems as force multipliers** – and sometimes as independent actors executing parts of an attack. The black market for AI will likely expand, with more custom models and services catering to criminals and authoritarians. It's a daunting picture, but awareness is the first step. By studying how Dark LLMs developed and are used today, defenders can anticipate their moves tomorrow and ensure that the **future of AI in cyberspace is not owned solely by the dark side**.

Sources:

- Burdett, E. (2025). *AI Goes on Offense: How LLMs Are Redefining the Cybercrime Landscape*. Rapid7 Blog[2][15][16][48][49][119].
- Bonderud, D. (2025). *LLMs gone bad: The dark side of generative AI*. Barracuda Networks[10][18][19][1][125].
- Schultz, J. (2025). *Cybercriminal abuse of large language models*. Cisco Talos Intelligence[13][12][126][43][127][128][58].
- Erzberger, A. (2023). *WormGPT and FraudGPT – The Rise of Malicious LLMs*. Trustwave SpiderLabs Blog[7][87][29][53].
- Poireault, K. (2023). *Five Malicious LLMs Found on the Dark Web*. Infosecurity Magazine[50][30][56][35][36][45].
- Burgess, M. (2023). *Criminals Have Created Their Own ChatGPT Clones*. Wired[27][28][93][34].
- Vijayan, J. (2025). *For $50, Cyberattackers Can Use GhostGPT to Write Malicious Code*. DarkReading[39][37][17][41].
- Abnormal Security Threat Intel. (2025). *How GhostGPT Empowers Cybercriminals with Uncensored AI*[22][23][61][62][64].
- O'Carroll, L. (2025). *Russia using criminal networks to drive increase in sabotage acts: Europol report*. The Guardian[109][108][116].
- Reuters. (2024). *US DOJ disrupts Russian AI-enabled propaganda campaign*. The Guardian[100][101][102].
- Paganini, P. (2025). *LameHug: first AI-powered malware linked to Russia's APT28*. Security Affairs[46][47].

---

[1] [3] [4] [10] [18] [19] [60] [125] LLMs gone bad: The dark side of generative AI | Barracuda Networks Blog

https://blog.barracuda.com/2025/06/20/llms-gone-bad-dark-side-generative-ai

[2] [5] [6] [9] [15] [16] [48] [49] [75] [76] [110] [119] [121] [122] How LLMs Like WormGPT Are Reshaping Cybercrime in 2025

https://www.rapid7.com/blog/post/ai-goes-on-offense-how-llms-are-redefining-the-cybercrime-landscape/

[7] [29] [44] [51] [53] [54] [71] [72] [73] [86] [87] [88] [89] [90] [95] WormGPT and FraudGPT – The Rise of Malicious LLMs

https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/

[8] [17] [37] [38] [39] [40] [41] [65] [66] [69] [70] [92] For $50, Attackers Can Use GhostGPT to Write Malicious Code

https://www.darkreading.com/cloud-security/cyberattackers-ghostgpt-write-malicious-code

[11] [12] [13] [14] [20] [21] [42] [43] [58] [68] [77] [78] [79] [82] [91] [97] [98] [126] [127] [128] Cybercriminal abuse of large language models

https://blog.talosintelligence.com/cybercriminal-abuse-of-large-language-models/

[22] [23] [61] [62] [63] [64] [83] [84] [120] How GhostGPT Empowers Cybercriminals with Uncensored AI | Abnormal AI

https://abnormal.ai/blog/ghostgpt-uncensored-ai-chatbot

[24] [25] [30] [31] [32] [33] [35] [36] [45] [50] [52] [55] [56] [57] [59] [94] The Dark Side of Generative AI: Five Malicious LLMs Found on the Dark Web

https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html

[26] [27] [28] [34] [74] [93] [96] Criminals Have Created Their Own ChatGPT Clones | WIRED

https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/

[46] [47] LameHug: first AI-Powered malware linked to Russia's APT28

https://securityaffairs.com/180092/apt/lamehug-first-ai-powered-malware-linked-to-russias-apt28.html

[67] Dark Web Intelligence - X

https://x.com/DailyDarkWeb/status/1772971948256997798

[80] [81] [85] Ukraine Exposes Russia's AI-Powered Hacking: A Glimpse Into the Future of Cyber Conflict - The420.in

https://the420.in/russia-ai-hacking-llm-cybersecurity-shinyhunters-defenders/

[99] [100] [101] [102] [103] [104] [105] US justice department says it disrupted Russian social media influence operation | Social media | The Guardian

https://www.theguardian.com/us-news/article/2024/jul/09/justice-department-russia-social-media

[106] For Beijing's Foreign Disinformation, the Era of AI-Driven Operations ...

https://thediplomat.com/2025/09/for-beijings-foreign-disinformation-the-era-of-ai-driven-operations-has-arrived/

[107] [111] [123] [124] Alan Turing Institute calls for AI Crime Taskforce | UKAuthority

https://www.ukauthority.com/articles/alan-turing-institute-calls-for-ai-crime-taskforce

[108] [109] [112] [113] [114] [115] [116] [117] Russia using criminal networks to drive increase in sabotage acts, says Europol | Cybercrime | The Guardian

https://www.theguardian.com/technology/2025/mar/18/russia-criminal-networks-drive-increase-sabotage-europol

[118] How Russia Uses Organized Crime for Espionage

https://newlinesinstitute.org/strategic-competition/how-russia-uses-organized-crime-for-espionage/

———————————-