TREADSTONE 71

The Second Model

Sber's AI Cybersecurity Threat Model v2.0 (July 2026) measured against v1.0 (March 2025) and against the April 2025 Treadstone 71 assessment

Intelligence Assessment

October 18, 2026

Prepared under the ASA Evidence Standard, Public Edition 1.2, with Treadstone 71 Advanced Structured Analytic Techniques applied. Probability language follows the house Kent-style spectrum with half-open intervals.

Treadstone 71 LLC · Half Moon Bay, California · info@treadstone71.com · treadstone71.com

Contents

Contents 2
Figures 3
TL;DR ..... **Error! Bookmark not defined.**
BLUF ..... **Error! Bookmark not defined.**
Who 4
What 5
So What 5
Why Now 5
Impacts So Far 5
Outlook 6
Strategic foresight, 2027 to 2029 6
Treadstone 71 Advanced SATs applied 6
1. Provenance and document forensics 8
2. The v1.0 baseline and what Treadstone 71 said about it 12
3. What v2.0 is 13
3.1 The split that organizes everything 13
3.2 The counts 13
3.3 The objects 16
3.4 The violators 16
4. Delta analysis: what moved, what split, what vanished, what is new 17
4.1 The crosswalk 17
4.2 Net new 18
4.3 Renamed, reweighted, demoted 20
4.4 Word-level deltas that carry meaning 20
5. Capabilities, functions, targets, intent, purpose, mission, goals, outcomes, maliciousness, and associated uses 20
6. The violator model under adversarial reading 22
7. The prompt-attack hub 24
8. Framework dependence: ATLAS spine, FSTEC grammar, no bibliography 25
9. The Russian-language environment: release, reception, regulators, competitors 28
9.1 Release and press 28
9.2 The borismlsec review and mapping, July 27, 2026 29
9.3 Who wrote the review, and whether Sber wrote it 30

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

9.4 The named file author 30
9.5 Lukatsky, August 3 to 4, 2026, and the one independent reading 31
9.6 The ITMO review and Sber's own statements 32
9.7 Regulators and standards 33
9.8 Competitors 34
9.9 What the environment shows 34
10. The 2025 Treadstone 71 findings scored against v2.0 35
11. Advanced SATs: method and output 36
12. Weaknesses that remain in v2.0 38
13. Empirical cross-check: where production defenses fail, measured 40
14. Reading the model from outside: indicators and collection 41
15. Evidence grading under the ASA Evidence Standard 42
Wrap-up 48
Sources 49
Appendix A. Crosswalk of the 70 v1.0 entries to v2.0 52
Appendix B. The 23 v2.0 entries with no v1.0 antecedent 53
Appendix C. Full list of v2.0 threats and methods with their links 55

Figures

Figure 1 Structural profile of the Sber AI threat model, v1.0 against v2.0
Figure 2 v2.0 realization methods by MITRE ATLAS tactic
Figure 3 v2.0 threats by consequence group
Figure 4 Fate of the 70 v1.0 threat entries in v2.0
Figure 5 Hub analysis of the 244 threat-method links
Figure 6 Violator class by threat (Appendix 3 heatmap)
Figure 7 Violator coverage: motivation against capability
Figure 8 Property, stage, and model class of the 37 threats
Figure 9a Document forensics: file properties, the dead Authors bookmark, cover dates
Figure 9b Document forensics: the violator table and the last page
Figure 10 Fourteen Treadstone 71 findings from 2025 scored against v2.0
Figure 11 How v2.0 is wired: entity types and relations
Figure 12 The 23 v2.0 entries with no v1.0 antecedent
Figure 13 Timeline of the two versions against the regulatory track
Figure 14 Treadstone 71 crosswalk against the borismlsec mapping
Figure 15 Sixteen ATLAS tactics against the twelve Sber groups
Figure 16 Picus Blue Report 2026 against Sber tactic coverage

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Initial Analysis

Sber's cybersecurity service rebuilt its AI threat model from the ground up between April 2025 and July 2026. Version 1.0 (cover date March 2025, released April 18, 2025) listed 70 threats in five object-bound families. Its successor splits the problem into 37 consequences ("threats," T01 to T37) and 51 attacker actions ("realization methods," ME01 to ME51), wires the two together through 244 explicit links, hangs the actions on twelve MITRE ATLAS tactics, and adds fifteen violator classes with motivation, capability, and knowledge levels. The document grew from 38 pages to 99. Twenty-three entries have no antecedent in v1.0. Nearly all of the new material covers AI agents: memory, tool calls, orchestrators, inter-agent protocols (the document names MCP), sandbox escape, credential theft, and an AI agent listed as an insider violator in its own right. The model also now names foreign "special services" as an attacker class and lists the use of a company's own AI agent to run autonomous multi-step cyberattacks as a threat. Version 2.0 closes five of the fourteen gaps Treadstone 71 called out in April 2025, partly closes six, and leaves three open. The two sharpest remaining weaknesses: no consequence or damage modelling, and an attribution table that assigns state intelligence services to only 15 of 37 threats while rating the same services capable of 41 of 51 methods.

Version 2.0 is a different instrument, not an update. Sber converted a flat catalogue into a kill-chain threat model that borrows its spine from MITRE ATLAS and its grammar from the Russian regulator FSTEC. The combination is new in Russian practice and, Likely (55 to <85 percent), exists to let Russian organizations produce FSTEC-compatible AI threat profiles with American tactic names. Treadstone 71 assesses the document as authentic Sber work (Almost certain, 95 to 99 percent), posted on July 24, 2026, and re-exported as a PDF on August 28, 2026, one page shorter than the copy two Russian reviewers counted in late July and early August, with the file's author field still carrying a name that the document's own table of contents shows Sber meant to remove. The model reveals more about Sber's own architecture and fears than v1.0 did: multi-tenant GPU inference, external model-as-a-service dependencies, agent runtimes in sandboxes, RAG stores, and long-term agent memory are all named as protected objects, which means Sber runs them. For a defender outside Russia the document is a usable, well-structured taxonomy of agentic AI threats. To an adversary of Russia it is a map of where Russia's largest bank expects the blow to land and, by omission, where it does not.

Who

Sber (PAO Sberbank), the Russian state-controlled bank, through the "AI Cybersecurity Directorate of Sber's Cybersecurity Service" (Управление кибербезопасности искусственного интеллекта Службы кибербезопасности Сбера). The PDF metadata names Alisa Andreevna Vorobyova (Воробьева Алиса Андреевна) as the file author; the document's table of contents carries a dead "Authors" bookmark, meaning an authors section existed in the Word source and someone deleted it before export. On the public record a person of that name is an associate professor at ITMO University who teaches attacks on AI systems; no public source connects her to Sber (Section 9.3). Readers: Russian AI-solution developers, MLOps and DevOps engineers, architects, cybersecurity experts, and risk owners, by the document's own statement. Audience in practice: the Russian financial sector, the AI Alliance members, and any organization that must produce a threat model under FSTEC methodology for a system that contains AI.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

What

A 98-page Russian-language document, "Модель угроз кибербезопасности AI для этапов сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с AI-решениями, v.2.0 (июль 2026)." The subtitle changed one word from v1.0: integrations "with applications" became integrations "with AI solutions." The change signals the core shift: the protected object is no longer an application that calls a model, it is an "AI solution," defined as an AI application or an AI agent.

So What

Three consequences follow. First, Russian AI threat modelling now has a reference document that a regulator-facing security team in Russia is able to use without translation effort, because it speaks FSTEC's language (нарушитель, потенциал, объект воздействия, способ реализации) while mapping to ATLAS tactics. Second, the model documents Sber's agentic AI build-out in enough detail to inform targeting: Sber runs multi-agent systems with orchestrators, long-term memory, tool and skill files, and external MLaaS connections, and it worries about all of them. Third, the attribution table contains a structural blind spot. Sber lists advanced external groups ("APT, special services") as motivated for espionage-type threats (data theft, model theft, backdoors) but not for credential theft, system-prompt extraction, runtime reconnaissance, or denial of service. A Western or Ukrainian service planning against Sber's AI starts where Sber's model says state actors have no interest.

Why Now

Four drivers converge. Sber shipped GigaChat agent products through 2025 and 2026, runs more than 900 autonomous agents against its own incident queue, and needed a threat model for them; Lebed told Vedomosti in December 2025 that Sber requires every banking IT system to "recognize that it is an AI agent who has come to them with a request and identify its authority," which is a control that presupposes exactly the agent-as-violator class v2.0 introduced; v1.0 had twelve agent threats bolted onto an application model and v2.0 has agents at its centre. Public incidents in 2025 and 2026 involving indirect prompt injection through RAG, memory poisoning, tool abuse, and AI-run intrusion campaigns gave Sber's incident responders material the document says it drew on ("собственного опыта расследования инцидентов"). FSTEC and the Bank of Russia moved on AI security requirements during the same period, and a document written in their vocabulary positions Sber as the de facto standard-setter. MITRE ATLAS itself expanded agentic coverage, giving Sber a current scaffold to borrow.

Impacts So Far

The document sits on sberbank.ru's Kibrary expert page (the QR code on page 7 resolves to https://www.sberbank.ru/ru/person/kibrary/experts/model-ugroz-kiberbezopasnosti-ai_2), posted July 24, 2026, with a seven-page diagram album beside it. Sber's own OFFZONE research on self-evolving agents cited it within four weeks. One Russian reviewer, Likely a Sber employee, published a 127-row v1.0-to-v2.0 mapping within three days; one independent reviewer, Lukatsky, published a critique within eleven. Section 9 documents reaction and uptake. The model has no English edition. No Western vendor or framework body

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

has cited it as of this report's collection cutoff, and MITRE ATLAS's September 2026 release added agent-discovery techniques and a Hermes-agent case study that converge on the same ground without reference to Sber.

Outlook

Treadstone 71 assesses it Likely (55 to <85 percent) that a v2.x or v3.0 appears within 12 to 18 months, driven by the same cadence that produced v2.0 sixteen months after v1.0 and by the speed of agentic AI change. At least one Russian regulator or standards body referencing or absorbing the model's structure by the end of 2027 is Likely (55 to <85 percent); candidates are FSTEC's BDU threat database, the Bank of Russia, and TC 164. Publication of consequence or damage modelling by Sber is Unlikely (15 to <45 percent), because the document scopes ranking out explicitly. The three prompt-attack methods (ME13, ME14, ME15), which already reach 21 to 22 of the 37 threats, remaining the dominant attack path in any Sber incident data behind the next revision is Highly likely (85 to <95 percent).

Strategic foresight, 2027 to 2029

Three paths. Path A, "national standard": the Sber model, or a derivative, becomes the AI annex of FSTEC's threat-modelling methodology and the Bank of Russia's financial-sector requirements; Russian vendors (Kaspersky, Positive Technologies, BI.ZONE) align their AI security products to its codes. Treadstone 71 puts Path A at Likely, 55 to <85 percent, by 2028. Path B, "fragmentation": competing models from Kaspersky, Yandex, and FSTEC itself coexist, and Sber's document stays a bank-sector reference. Unlikely, 15 to <45 percent. Path C, "export": Sber or the state offers the model, translated, through BRICS or CIS channels as a sovereign alternative to OWASP and NIST guidance. Unlikely, 15 to <45 percent, before 2028, and dependent on the Russian state deciding the ATLAS dependency is acceptable in an export product. Confidence in the three-path frame is moderate; the T-Score on the evidence base behind the forward judgments is 69, Adequate (Section 15), which caps every forward judgment at Likely.

Treadstone 71 Advanced SATs applied

Nine techniques from the Treadstone 71 Advanced Structured Analytic Techniques catalogue, the analytic layer of the firm's decision engines, ran against the two documents and the collected Russian-language reporting, plus two classic methods (Key Assumptions Check and Analysis of Competing Hypotheses) that sit outside the catalogue. Section 11 reports each technique's output. In brief: Hypothesis Evolution Tracking carried the fourteen 2025 findings forward and scored them; Adversary Threat Vector Analysis built the 244-link graph and found the prompt-attack hub; Adversarial Cognitive Simulation read the model through an attacker's eyes and produced the attribution blind-spot finding; Decompositional Uncertainty Mapping tiered every judgment; Adaptive Threat Calibration and Risk Indexing (the ATCRI engine) weighted the 23 net-new entries; Momentum-Reversal Analysis tested whether Russian AI security doctrine is moving toward or away from Western frameworks (toward, with a Russified wrapper); Emergence-versus-Direction Discrimination tested whether the agentic emphasis is a response to incidents or to Sber's product roadmap (each, with the roadmap leading); Counterintelligence Deception Diagnostics graded the document for shaping and found

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

none beyond ordinary institutional self-presentation (D-Score 25, the floor); Competitive Intelligence Stratagem placed the document against Kaspersky, Positive Technologies, and FSTEC.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

1. Provenance and document forensics

Treadstone 71 holds five Sber-origin files and graded each as a technical artefact.

The file under examination is `model_ugroz_kiberbezopasnosti_v2.0.pdf`, 98 pages, A4, 15,438,743 bytes. Treadstone 71 hashed the file at collection. The PDF producer string reads Microsoft Word 2016. Creation and modification timestamps match: August 28, 2026, 08:14:31 UTC. The author field reads "Воробьева Алиса Андреевна." On the cover the version line reads "v.2.0 (июль 2026)." The document's table of contents ends with the line "Авторы ... Ошибка! Закладка не определена." (Error! Bookmark not defined). Word prints that string when a table-of-contents entry points to a heading that no longer exists. Someone wrote an "Authors" section, removed it before export, and left the table of contents unrefreshed. The metadata author field survived the removal.

Two readers counted 99 pages in the released document before the file in hand existed. The Telegram channel "Борис_ь с ml" reported on July 27, 2026: "37 угроз и 51 способ их реализации, 43 объекта защиты - 99 страниц в документе" (37 threats, 51 methods, 43 protected objects, 99 pages). Alexey Lukatsky wrote on August 4: "не смог отказать себе в том, чтобы прочитать эти 99 страниц" (I could not deny myself the chance to read these 99 pages). The file in hand has 98 pages and a creation date of August 28. Treadstone 71 assesses it Highly likely, 85 to <95 percent, that Sber replaced the released PDF about five weeks after release with a re-export from which someone had cut the authors page, and that nobody regenerated the table of contents. Two independent counts, a dead bookmark, and an export date four weeks after release leave a double miscount as the only alternative, Highly unlikely, 5 to <15 percent. The live file on the Kibrary server settles it.

The v1.0 files close the authorship question further than the v2.0 file alone allows. Its original, `model_ugroz_kiberbezopasnosti_ai.pdf` from the Kibrary server (38 pages, 5,770,423 bytes) carries the same author field, "Воробьева Алиса Андреевна," Microsoft Word 2016, created April 18, 2025, 07:55:04 UTC, seven minutes after the accompanying diagram `obobshchennaya_skhema_obekta_zashchity_i_aktualnykh_ugroz_kb_ai.pdf` (Microsoft PowerPoint 2016, same author, 07:48:44 UTC). Sber's press release went out the same morning. The person named in the metadata produced the v1.0 model, the v1.0 diagram, and the v2.0 model across sixteen months, Almost certain, 95 to 99 percent, from three consistent fields on three files from two different applications. A reused template does not explain a PowerPoint deck and two Word documents sharing one author over two releases. That person held a producing role on Sber's side, as employee, contractor, or academic collaborator, Highly likely, 85 to <95 percent. Section 9.4 reports what the public record says about a person of that name and grades the identity link. The v2.0 diagram album `skhemy_i_matricy_v_2.pdf` (7 pages, assembled with iLovePDF on July 22, 2026, 07:40 UTC, no author field) adds a date: the diagrams were final two days before the July 24 posting that the Telegram reviewer recorded and eight days before the July 30 press release.

The model versions share a second artefact. Neither PDF contains extractable text. Every glyph sits in a bitmap mask at 600 dots per inch. A text extractor returns a few dozen characters from the entire 98 pages (table headers in a fallback font, one Latin phrase, identifier strings). The album behaves the same way. Behind that sits, Highly likely (85 to <95 percent), a font-embedding restriction rather than deliberate anti-copying: Sber's corporate typeface carries licensing flags that make Word rasterize text on export, and the handful of extractable strings are exactly those set in Calibri, Cambria, and Arial. The effect is the same either

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

way. A reader cannot search it, machine-translate it, or load it into a tool. Treadstone 71 ran optical character recognition at 200 dpi on all 98 pages and verified every count in the present report against the page images by eye, then checked the 51 column totals in the album's matrix (page 7) against the rebuilt graph: all 51 match.

The v1.0 source, 36 numbered pages plus cover, ended with a ten-item list of "materials used in preparation": OWASP Top 10 for LLM Applications 2025, OWASP Machine Learning Security Top 10, OWASP AI Security Solutions Landscape, OWASP Agentic Threats Taxonomy (draft), OWASP AI Exchange 4.5, MITRE ATT&CK, MITRE ATLAS, Google SAIF, NIST Adversarial Machine Learning taxonomy, and the AWS Generative AI Security Scoping Matrix. Version 2.0 carries no bibliography at all. The only external framework named in the body is MITRE ATLAS, which supplies the tactic names, and one footnote explaining why two ATLAS tactics (Discovery and AI Attack Staging) received no groups of their own. Dependence on Western frameworks deepened while acknowledgement of them disappeared. Treadstone 71 returns to the point in Section 8.

Figure 9a and 9b print the page evidence behind each row of Table 1.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Figure 9a. Document forensics: file properties, the dead Authors bookmark, and the cover dates

A. v2.0 PDF properties (pdfinfo output on the file as received)

```
Author:         Воробьева Алиса Андреевна
Creator:        Microsoft® Word 2016
Producer:       Microsoft® Word 2016
CreationDate:   Fri Aug 28 08:14:31 2026 UTC
ModDate:        Fri Aug 28 08:14:31 2026 UTC
Pages:          98
Page size:      595.32 x 841.92 pts (A4)
File size:      15438743 bytes
PDF version:    1.5
```

B. v2.0 table of contents, page 6: the last entry, 'Авторы', points to a heading that no longer exists



Приложение 4. Зоны ответственности за меры защиты от угроз кибербезопасности AI и способов их реализации 90

Зоны ответственности по угрозам кибербезопасности AI 90

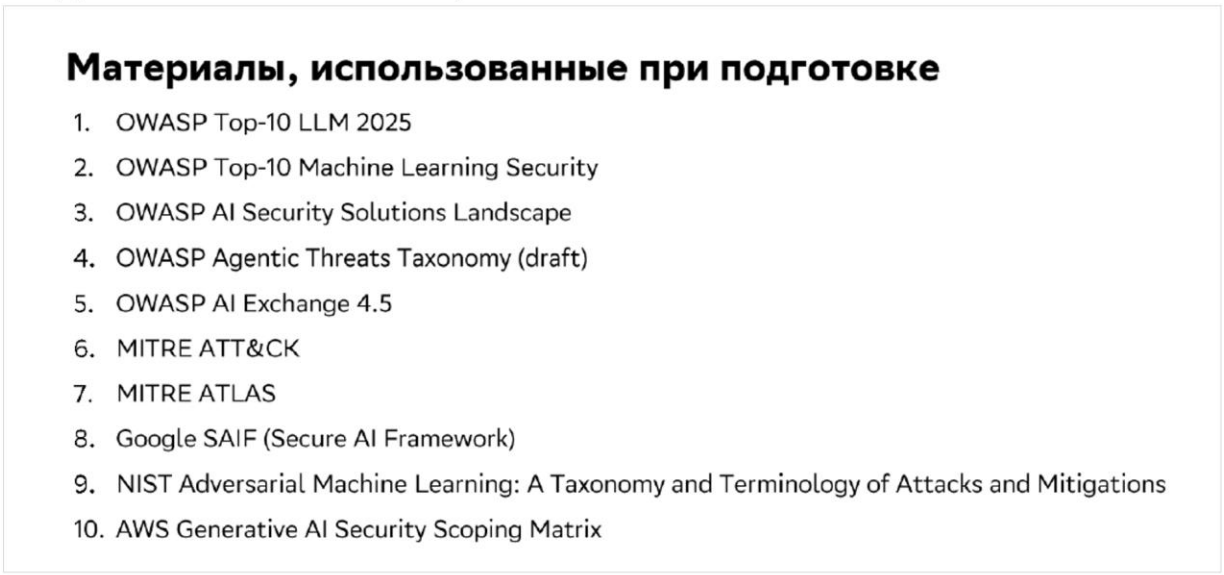
Зоны ответственности по способам реализации угроз кибербезопасности AI 93

Авторы ..... Ошибка! Закладка не определена.

C. v1.0 cover: 'v.1.0 (март 2025)'; public release 18 April 2025



D. v1.0 page 36: the ten-item sources list that v2.0 does not carry



Материалы, использованные при подготовке

1. OWASP Top-10 LLM 2025
2. OWASP Top-10 Machine Learning Security
3. OWASP AI Security Solutions Landscape
4. OWASP Agentic Threats Taxonomy (draft)
5. OWASP AI Exchange 4.5
6. MITRE ATT&CK
7. MITRE ATLAS
8. Google SAIF (Secure AI Framework)
9. NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
10. AWS Generative AI Security Scoping Matrix

Figure 9b. Document forensics: the violator table and the last page of v2.0

E. v2.0 page 46: violator classes H1.1 to H2.10 with capability, knowledge and potential

Виды нарушителей кибербезопасности AI и их потенциал

Нарушители КБ AI разделяются на две основные категории: внешние и внутренние. Внутренние нарушители КБ AI дополнительно детализированы по ролям в рамках жизненного цикла разработки AI-решений.

Код	Вид нарушителя КБ AI	Уровень технических возможностей	Уровень осведомлённости и доступа	Потенциал[5]
H1.1	Внешние преступные группы	H2	Black-/Grey-box	Средний
H1.2	Внешние субъекты (физические лица)	H1	Black-/Grey-box	Низкий
H1.3	Конкурирующие организации	H2	Black-/Grey-box	Средний
H1.4	Бывшие работники (администраторы, разработчики)	H2	Grey-box	Средний
H1.5	Продвинутые внешние группы (APT, специальные службы)	H3	Grey-box	Высокий
H2.1	Пользователи	H1	Black-box	Низкий
H2.2	Администраторы информационных ресурсов	H1–H2	White-box	Средний
H2.3	Администраторы безопасности	H1–H2	Grey-box	Средний
H2.4	D-People (Data Owner)	H2	White-box	Средний
H2.5	D-People (Data Engineer)	H2	White-box	Средний
H2.6	D-People (Data Scientist)	H2	White-box	Средний
H2.7	D-People (ML engineer)	H2	White-box	Средний
H2.8	D-People (DevOps engineer)	H2	White-box	Средний
H2.9	D-People (Разработчик)	H2	White-box	Средний
H2.10	AI-агент[6]	H2	Grey-box	Средний

F. v2.0 page 98, final page: Appendix 4 ends at ME51; no sources or authors section follows

ID	Способ реализации угрозы КБ AI	Ответственные за меры защиты
ME47	Извлечение остаточных данных из коммунальной инфраструктуры инференса	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) AI-решения; Владелец ИТ-инфраструктуры AI-решения; Владелец ИТ-инфраструктуры исполнения AI-агентов
ME48	Эксфильтрация, инверсия или реверс-инжиниринг модели через взаимодействие с моделью и анализ ее ответов	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) модели
ME49	Эксфильтрация или восстановление фрагментов обучающих данных через взаимодействие с моделью	Владелец ИТ-инфраструктуры инференса и подключения модели, предоставления контента; Разработчик (владелец) модели
ME50	Восстановление данных из векторных	Разработчик (владелец) модели; Владелец ИТ-инфраструктуры инференса и

Table 1. Document profile.

Attribute	v1.0	v2.0
Cover date	March 2025	July 2026

Attribute	v1.0	v2.0
Public release	April 18, 2025 (press release)	July 24, 2026 (Kibrary posting, per borismlsec); July 30, 2026 (press release)
PDF export date	April 18, 2025, 07:55 UTC	August 28, 2026, 08:14 UTC (released file: unknown; album July 22, 2026)
PDF author field	Воробьева Алиса Андреевна (model and diagram)	Воробьева Алиса Андреевна
Pages	38 (36 numbered plus cover)	99 at release; 98 in the August 28 re-export
Issuing unit named	"Сбер Кибербезопасность" brand only	AI Cybersecurity Directorate, Sber Cybersecurity Service
Lifecycle stages	3	3 (unchanged names)
Threat entries	70 (D, Infr, M, App, Ag)	37 (T01 to T37)
Realization methods	none as a class	51 (ME01 to ME51)
External framework citations	10 listed	0 listed; ATLAS named in text
Violator model	none	15 classes, 3 capability levels, 3 knowledge levels
Security properties	4 (C, I, A, veracity)	3 (C, I, A)
Roles responsible for measures	4 (data owner, model owner, application owner, infrastructure owner)	9
Practical method	none	five-step algorithm with worked example
Text extractable	no	no

2. The v1.0 baseline and what Treadstone 71 said about it

Version 1.0 was a catalogue. Seventy threats sat in five families named after the primary object attacked: data (D01 to D06), infrastructure (Infr01 to Infr25), model (M01 to M13), application (App01 to App14), and AI agents (Ag01 to Ag12). Each card carried a description, consequences, the object of impact, the violated property, the model types exposed (PredAI, GenAI, or the two together), and the person responsible for mitigation. There was one diagram, a lifecycle schematic with threat codes pinned to objects. Attackers, tactics, and methods were absent. The framework borrowed its vocabulary from OWASP and its ambition from MITRE.

Treadstone 71's April 2025 assessment made fourteen findings that bear on v2.0. The model treated lifecycle phases as discrete and sequential. Confidentiality, integrity, and availability carried the whole analysis, with no modelling of intent. The model assumed threats observable and attributable. Engineering logic governed where the problem needed adversarial logic. Polymorphic and mimicry-based evasion went unaddressed, as did reinforcement loops seeded by fake user behaviour. Claims of completeness appeared. Insider sabotage was a blind spot. Framework-level attacks on PyTorch or ONNX were out of reach. Upstream contamination through multi-model chains was absent, and so was endogenous misalignment, meaning systems that go wrong without external compromise. No consequence or damage modelling existed. Nothing caught early signals from informal channels. Adversarial co-evolution had no place in it. Section 10 scores each of the fourteen against v2.0.

3. What v2.0 is

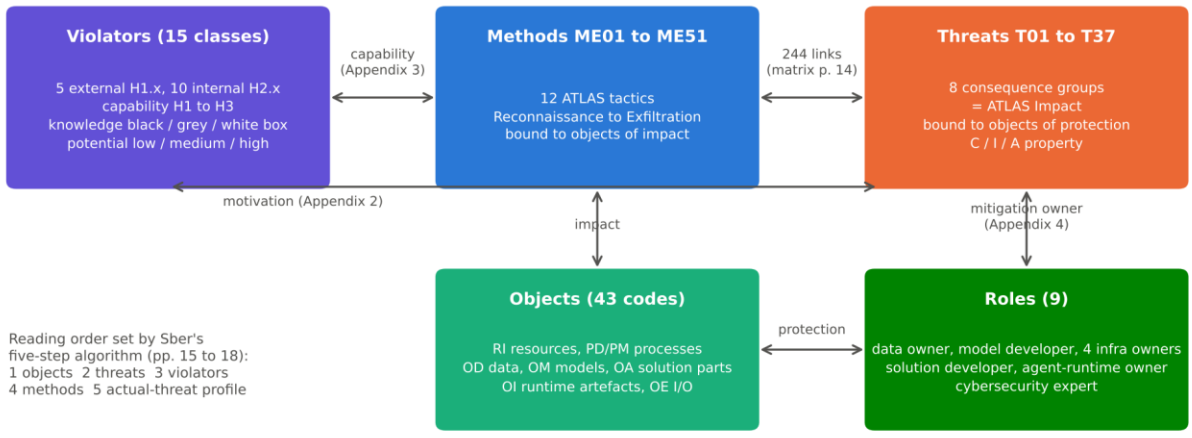
3.1 The split that organizes everything

Page 11 states the founding decision of v2.0: separate threats from the ways attackers realize them. A threat (угроза КБ AI) describes what happens: a negative consequence tied to an object of protection. A realization method (способ реализации), by contrast, describes how an attacker gets there: a technique tied to an object of impact. In ATLAS terms, Sber maps methods to the preparatory and execution tactics from Reconnaissance through Exfiltration and maps threats to the final tactic, Impact. The document states the two rationales plainly: defenders design blocking measures against methods and compensating measures against consequences, so the split lets a defender build "multi-level protection" and stop an attacker "at early stages."

The second decision is numbering. Version 1.0 tied each identifier to a primary object (D for data, Infr for infrastructure). Its successor uses through-numbering independent of object, because, the document says, several methods realize one threat and one method leads to several threats. The matrix on page 14 is the result: 37 rows, 51 columns, 244 filled cells. Treadstone 71 rebuilt the matrix from the individual cards and found the two directions of cross-reference agree on every one of the 244 links. The right-hand margin of the printed matrix carries per-threat counts (3, 7, 5, 5, 6, 6, 6, 9, 5, 3, 16, 12, 7, 7, 1, 6, 5, 5, 6, 4, 8, 2, 6, 4, 6, 7, 7, 8, 7, 6, 6, 7, 12, 11, 14, 6, 3); every one matches the rebuilt graph. At the link level the document is internally consistent. That is a mark of tooling rather than hand-editing, and it means the model is machine-usable once a reader recovers the text from the bitmaps.

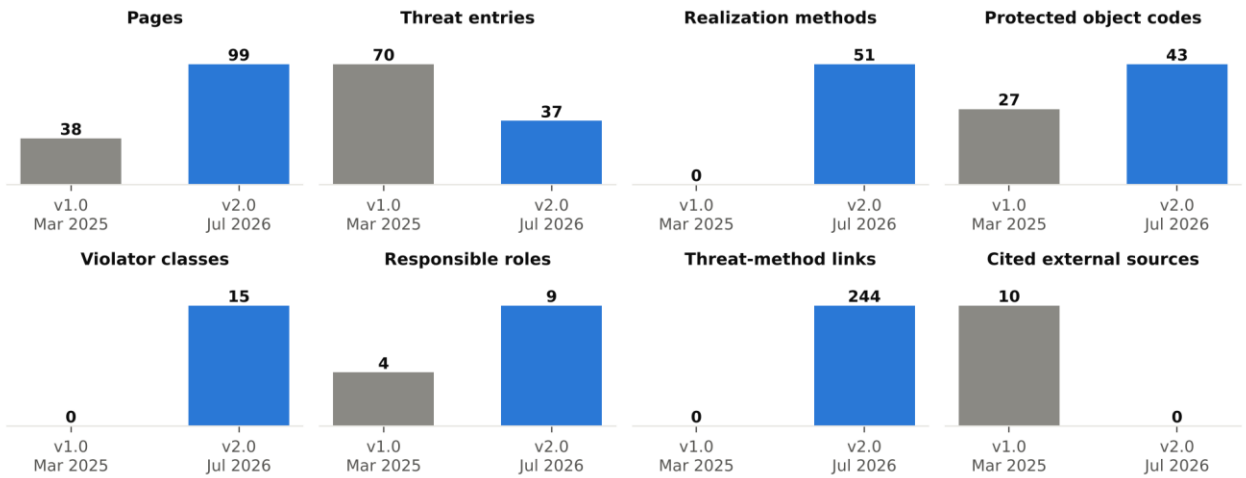
Figure 11 shows the five entity types and the relations between them.

Figure 11. How v2.0 is wired: five entity types and the relations the document defines between them



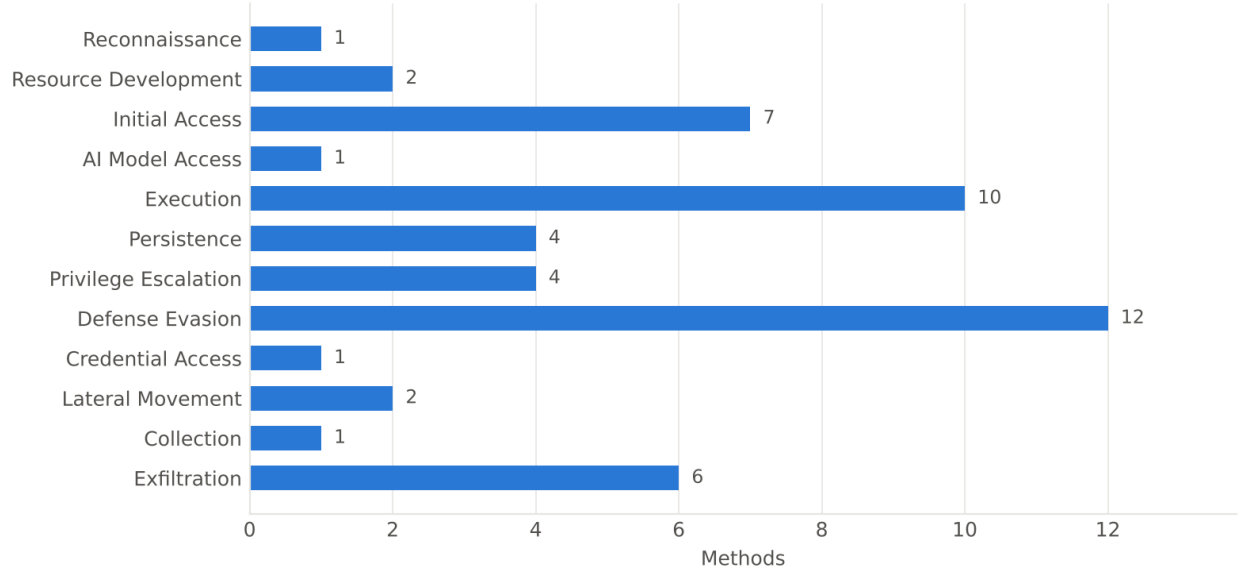
3.2 The counts

Figure 1. Structural profile of the Sber AI threat model, v1.0 against v2.0



Source: Treadstone 71 count of both documents. Object codes: v1 IR/DR/MR/AR incl. child codes (27, per the borismlsec mapping); v2 RI/PD/PM/OD/OM/OA/OI/OE (43). Pages: v2 as released

Figure 2. v2.0 realization methods (ME01 to ME51) by MITRE ATLAS tactic

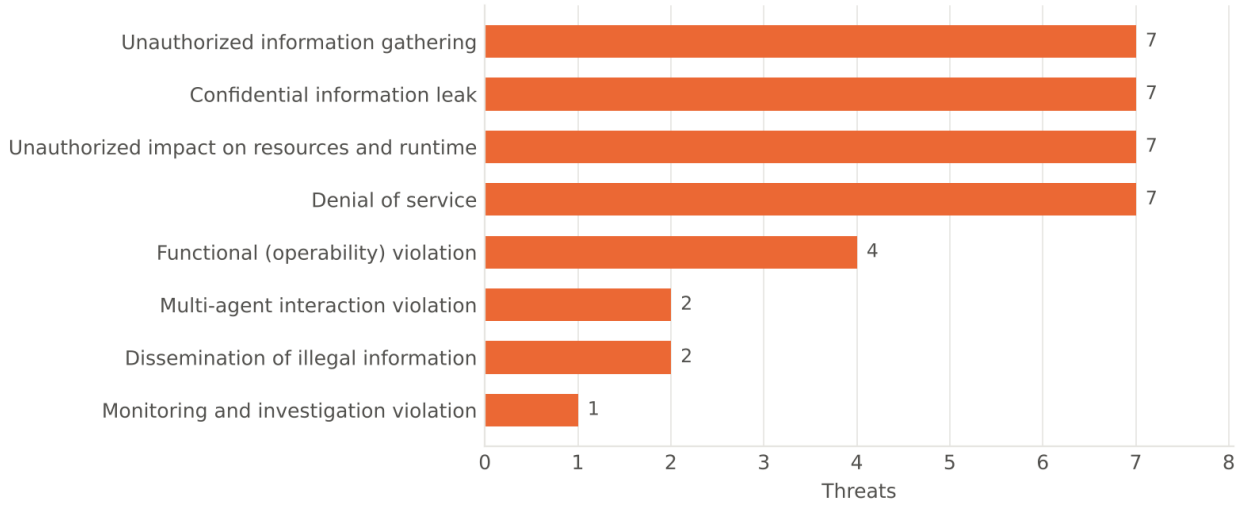


Discovery and AI Attack Staging are folded into neighbouring tactics by Sber (footnote 1, p. 11).

Defense Evasion carries 12 of the 51 methods, Execution 10, Initial Access 7, Exfiltration 6. Four tactics carry a single method each (Reconnaissance, AI Model Access, Credential Access, Collection). Treadstone 71 reads the Defense Evasion weighting as a sign of what Sber's incident data contains: access-control gaps, disabled filters, interception between components, and model substitution in transit are the defender's recurring failures, and the model files them under evasion because they are the ways an attacker avoids the controls Sber already has.

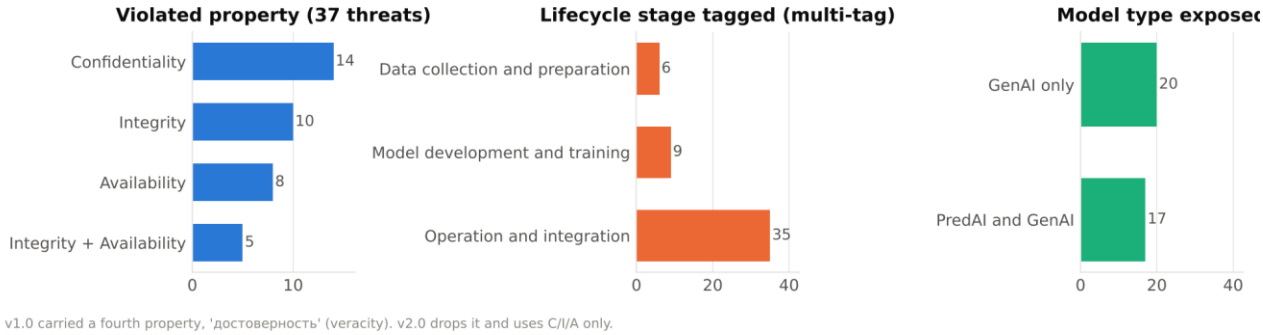
Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Figure 3. v2.0 threats (T01 to T37) by consequence group (ATLAS I



Four of the eight consequence groups carry seven threats each (information gathering, confidential leak, impact on resources and runtime, denial of service). Multi-agent interaction and illegal-information dissemination carry two each. Monitoring failure carries one. The symmetry of the four sevens is suspicious in the way all tidy taxonomies are suspicious: it suggests the authors sized the groups to balance rather than to reflect frequency. Treadstone 71 has no way to test that against Sber's incident base, which the document references but does not publish.

Figure 8. Where the 37 v2.0 threats sit: property, stage, and model class



Twenty of the 37 threats apply to generative AI only; 28 of 37 carry the operation-and-integration stage tag alone. The centre of gravity has moved from training pipelines (where v1.0 spent 31 of 70 entries on data and infrastructure) to deployed agents. Confidentiality is the violated property for 14 threats, integrity for 10, availability for 8, and integrity-plus-availability for 5. The fourth property v1.0 carried, достоверность (veracity, defined there as "correspondence of information to facts, absence of subjectivity, bias, or semantic distortion"), is gone. Hallucination and bias survive as T33 and T34, now filed as integrity violations. The removal matters for the cognitive-warfare reading of the model: v1.0 had a property that let a defender say an AI output was false without saying anyone had tampered with it. Version 2.0 forces every falsehood into the integrity column, where it must have a cause in an attacker's action or a process failure. That is cleaner engineering and weaker epistemology.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

3.3 The objects

Version 2.0 enumerates 43 object codes in eight families: five resource infrastructures (RI.01 to RI.05), three data processes and five model processes (PD, PM), three data objects (OD), five model objects including LLM adapters in training and in inference (OM.01a to OM.02b), six AI-solution objects (OA.01 to OA.06: application, agents, functions and tools, system prompt, configuration, memory and cache), fourteen runtime artefacts (OI.01 to OI.14, including request logs, traces, credentials, message-exchange mechanisms, the orchestrator, and "resource systems" the agent reaches), and two input/output objects (OE.01, OE.02). The document says in its own introduction that system prompts, LLM adapters, agent memory and cache, message exchange in multi-agent systems, agent runtime environments, and agent resource systems were "not considered in the previous version but today are critical attack points."

Every object named is an object Sber operates. Read as a whole, the list is an architecture disclosure. Sber runs: interactive development environments for training (RI.02 names Jupyter, Colab, and cloud IDEs); own inference servers plus external MLaaS providers through API gateways (RI.03); isolated agent runtimes as containers, sandboxes, serverless functions, or virtual machines (RI.05); vector stores for RAG (OI.13); an orchestrator managing agent lifecycle and task routing (OI.11); external skill repositories (OI.12); and inference caches (named in ME46). The model also admits, in ME47, that Sber's inference runs on "communal" infrastructure where a previous tenant's residual data (GPU memory, caches, temporary files, buffers) is a threat. That is a multi-tenant GPU estate, and the authors wrote the threat as though they had observed it.

3.4 The violators

Pages 45 to 47 and Appendix 2 define fifteen violator classes. Five are external: criminal groups (H1.1), individuals (H1.2), competitors (H1.3), former employees (H1.4), and "advanced external groups (APT, special services)" (H1.5). Ten are internal: users (H2.1), IT-resource administrators (H2.2), security administrators (H2.3), six "D-People" roles (data owner, data engineer, data scientist, ML engineer, DevOps engineer, developer; H2.4 to H2.9), and the AI agent itself (H2.10). Each carries a technical-capability level (H1 low, H2 medium, H3 high), a knowledge level (black box, grey box, white box), and a resulting potential (low, medium, high). Only H1.5 reaches high potential. The document caps every other class at medium, including every insider, on the reasoning that insiders lack either machine-learning competence or breadth of reach.

As for the AI agent as violator, it is the entry that distinguishes the Sber document from every Western framework Treadstone 71 has reviewed. Footnote 6 explains: an agent acts "autonomously (unintentional actions, errors, hallucinations) or compromised (executing malicious instructions from an external or internal violator)," and the tables do not separate the two. Appendix 2 gives the agent a motivation: "priority of seeking positive user feedback (approval, avoiding negative evaluations) over system constraints, which provokes violation of safety instructions in a goal conflict," and "execution of instructions imposed by an external or internal violator, perceived as routine." Footnote 9 defines agent motivation as "the goal-setting system embedded in the system prompt, reward function, or alignment." Sber is modelling sycophancy and reward-hacking as insider threat. Treadstone 71 has not seen that formulation in a bank's threat model before.

The naming of "special services" (спецслужбы) deserves a line. A Russian state bank's public document names foreign intelligence services as an attacker class with cyber-espionage motives: "theft of models,

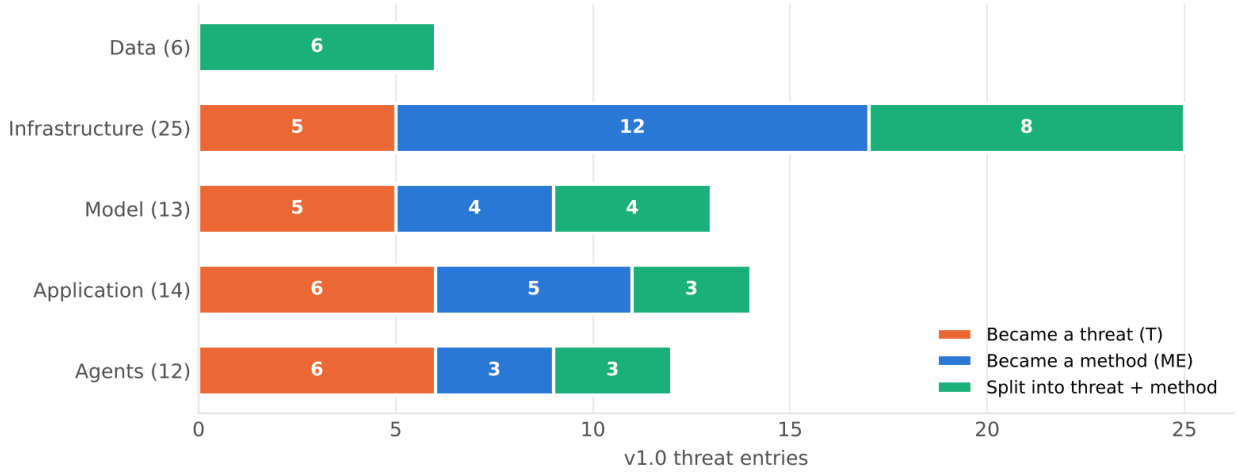
training data, algorithms," "obtaining advanced technologies and competitive intelligence," "destabilization of critical infrastructure using AI," "long-term persistence (backdoors in models and AI services)," and "undermining trust in AI solutions through complex attacks at the training or inference stage." The motivations read as a mirror of what Russian services do abroad, written by people who expect the same done to them.

4. Delta analysis: what moved, what split, what vanished, what is new

4.1 The crosswalk

Treadstone 71 mapped each of the 70 v1.0 entries to its v2.0 descendants by content, not by title. Appendix A prints all 70 rows. The summary:

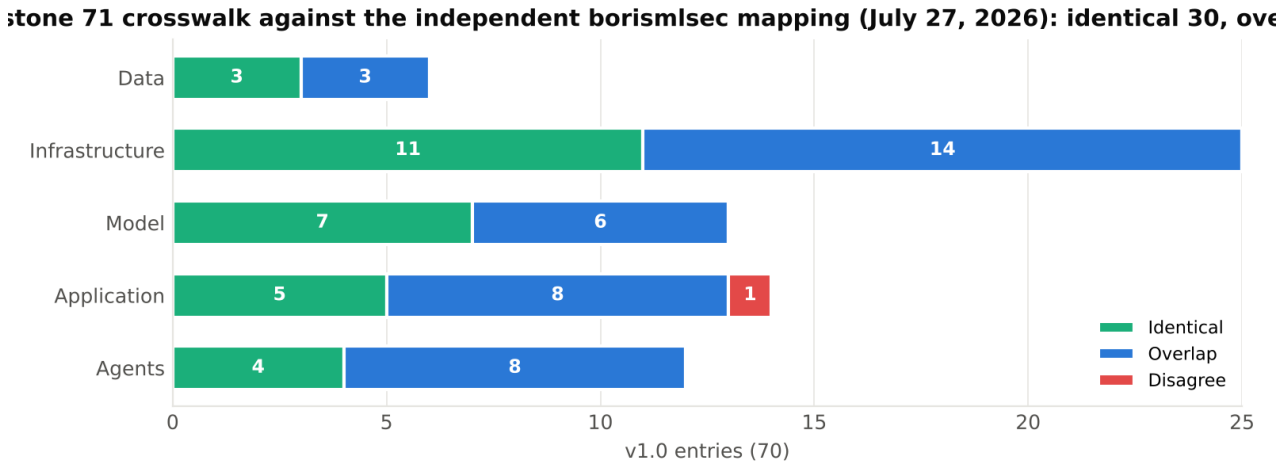
Figure 4. Fate of the 70 v1.0 threat entries in v2.0



Of the 70 v1.0 entries, 22 became a v2.0 threat, 24 became a v2.0 method, and 24 split so that their consequence went to a threat and their mechanism to a method. Sber dropped no v1.0 entry outright. Several merged: Infr01, Infr02, and Infr21 (three flavours of unauthorized dataset modification) collapse into ME31; Infr07 and Infr09 (backdoors in open-source model files and in weights) collapse into ME08; M02 and M09 (model vulnerability to adversarial attack and bypass of built-in protections) collapse into ME38; Infr12 and M07 (two availability entries) collapse into T22 and ME18. The v1.0 agent family (Ag01 to Ag12) survives almost entirely, with each entry promoted to a threat or a method and given more company.

An independent check exists. On July 27, 2026, three days after the Kibrary posting, the author of the Telegram channel "Борис_ь с ml" published an unofficial v1.0-to-v2.0 mapping as a spreadsheet (127 rows, dated July 27, "Направление: только v1 → v2," authorship "tg: @borismlsec"). Treadstone 71 built its own crosswalk before obtaining that file and compared the two afterwards. Figure 14 shows the result.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM



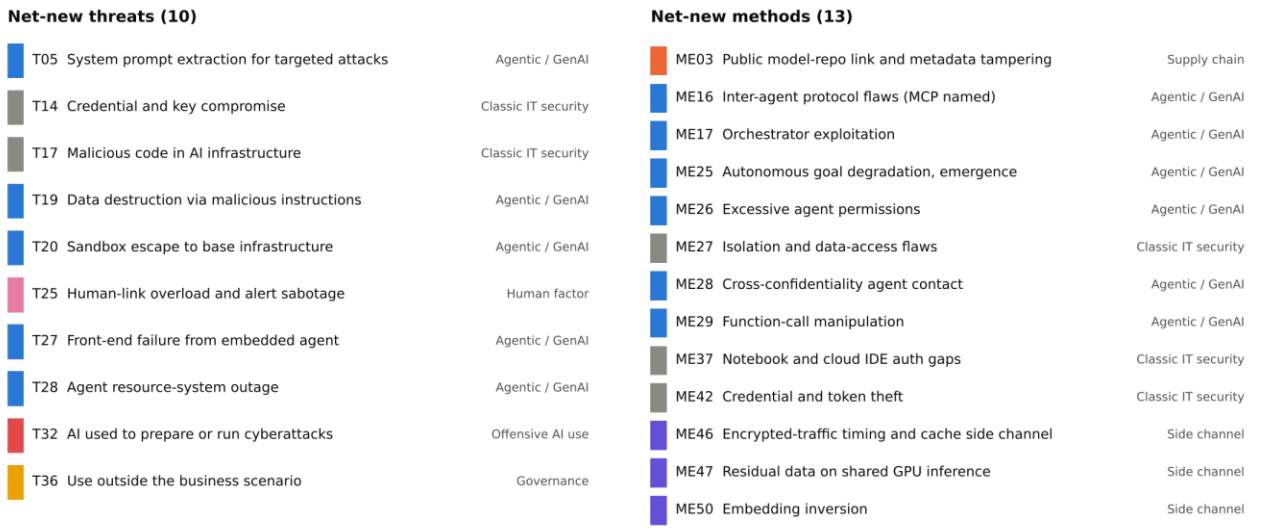
Identical: same v2.0 descendant set. Overlap: at least one shared descendant. The one disagreement is App07 (T71: T33, ME43, ME44; borismlsec: T12, T1

The two mappings name an identical descendant set for 30 of the 70 entries and share at least one descendant for 39 more. They disagree on one: App07 (execution of malicious instructions created by the model), which Treadstone 71 reads as a hallucination-and-execution pair (T33, ME43, ME44) and the Russian reviewer reads as a data-impact pair with function-call manipulation (T12, T19, ME29). Both readings fit the v1.0 card text, which covered all of those effects. The 69-of-70 overlap between two analysts working from different languages, different tools, and no contact raises confidence in the crosswalk to the level the ASA standard reserves for corroborated items; Section 15 records the F5 credit. Where the two mappings differ is in their net-new counts,, because they draw the line differently: the Russian mapping lists six threats and eighteen methods without a v1.0 antecedent (T14, T20, T24, T25, T28, T36; ME02, ME03, ME15, ME16, ME17, ME25, ME26, ME27, ME28, ME36, ME37, ME42, ME43, ME44, ME45, ME46, ME47, ME50) plus nineteen object codes, where Treadstone 71 lists ten threats and thirteen methods. The disagreements are matters of threshold (whether a v1.0 card that mentions an effect in passing counts as an antecedent), not of substance. Both lists put the same core in the net-new column: credential compromise, sandbox escape, human-link overload, resource-system outage, off-scenario use, orchestrator and protocol attacks, autonomous goal drift, excessive permissions, cross-confidentiality contact, interactive-environment authentication, credential theft, and the three side channels.

4.2 Net new

Twenty-three v2.0 entries have no v1.0 antecedent. Figure 12 lists them.

Figure 12. The 23 v2.0 entries with no v1.0 antecedent, by the kind of gap they close (agentic 11, classic IT 5, side channel 3, other



Ten are threats. T05 (system prompt extraction as reconnaissance for targeted attacks) and T14 (credential and key compromise) are classic security entries v1.0 lacked entirely; their absence in 2025 was the more surprising fact. T17 (malicious code execution across all five infrastructures), T19 (data destruction through executed instructions), T20 (sandbox escape to the base infrastructure), T27 (front-end failure caused by an embedded agent), and T28 (loss of an agent's resource systems) are the agentic additions. T25 is novel in any framework Treadstone 71 tracks: "creation of operational interference and sabotage of processes by overloading the human link," meaning an AI solution that floods staff with false or irrelevant alerts until they miss the real one, whether by attack or by misconfiguration. T32 is the offensive-use entry: an organization's own AI application or agent used "for generating ideas, code, phishing letters, or instructions for committing cybercrimes, as well as for autonomous execution of multi-step cyberattacks (for example, exploitation of a vulnerability chain, lateral movement, counter-investigation)." T36 is governance: use outside the business scenario.

Thirteen are methods. The agentic set: ME16 (flaws in inter-agent protocols, "for example MCP"), ME17 (orchestrator exploitation causing deadlock or misassignment), ME25 (autonomous goal degradation through emergent behaviour or self-learning errors, attributed to an internal violator who neglects periodic goal-conformance checks), ME26 (excessive permissions), ME28 (direct contact between agents of different confidentiality levels), ME29 (function-call manipulation). Then the classic set: ME03 (public model-repository compromise by link or metadata substitution), ME27 (isolation flaws), ME37 (authentication gaps in notebooks and cloud IDEs), ME42 (credential theft including external MLaaS keys). The side-channel set is the most technically specific material in the document: ME46 (packet size and timing analysis of encrypted streaming responses, including detection of inference-cache hits by response latency), ME47 (residual data from a previous tenant on shared inference hardware), and ME50 (inversion of vector embeddings back to source text).

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Treadstone 71 reads ME46 and ME47 as the clearest evidence that Sber's AI security team reads current academic literature on LLM side channels and applies it to its own multi-tenant estate. Neither entry appears in OWASP's LLM list. Each appeared in security research during 2024 and 2025.

4.3 Renamed, reweighted, demoted

Sber removed the v1.0 per-card field "persons responsible for mitigation" (four roles) from the cards and rebuilt it as Appendix 4 against nine roles, with the cybersecurity expert now a named role holding "control of protective measures for all objects at all stages." New codes replace the v1.0 object codes (IR, DR, MR, AR) wholesale. Throughout, "AI-решение" (AI solution) replaces the v1.0 word "приложение" (application) as the terminal object. The v1.0 definition list of three footnotes becomes a four-page terms-and-definitions section of 27 entries, including definitions of alignment, fine-tuning, LoRA, RAG, prompt attack, system prompt, token, and embeddings. Those definitions are correct and current. The document defines "Выравнивание (alignment)" as "the process of reconciling the goals, behaviour, and responses of an AI model with user intentions, ethical norms, and established constraints," which is the Western usage, not a sovereignty-flavoured one.

4.4 Word-level deltas that carry meaning

Treadstone 71 compared the two documents' framing paragraphs line by line. Four changes carry analytic weight.

First, v1.0 opened with "the growth of AI use in critically important areas requires a preventive approach." Version 2.0 opens with "the AI Cybersecurity Directorate ... prepared the document on the basis of analysis of the current threat landscape, world practice, and own experience investigating incidents." The claim to incident experience is new. Sber is asserting that it has investigated AI security incidents and that the model reflects them.

Second, v1.0 described itself as "a tool for teams that need a systematic approach to threat modelling" and claimed to let organizations "form preventive protective measures." Version 2.0 says, on page 15, that the model "does not define the protective measures themselves," and on page 18 that ranking by criticality and the design of measures "fall outside the scope of this threat model." The scope has narrowed and the claims have become more defensible. That retreat answers Treadstone 71's 2025 finding that the model claimed to "neutralize all threats."

Third, v2.0 footnote 5 states that violator potentials "are baseline (reference) values" that an organization "adjusts upward or downward on the basis of its internal data." Footnote 4 excludes a provider's side of a model-as-a-service relationship from the boundary of analysis. Each is a hedge a mature methodology carries and v1.0 lacked.

Fourth, the phrase "с учётом требований регуляторов" (taking regulators' requirements into account) appears in the practical guide where v1.0 had nothing. Combined with the FSTEC vocabulary analysed in Section 8, it tells the reader whom the document serves.

5. Capabilities, functions, targets, intent, purpose, mission, goals, outcomes, maliciousness, and associated uses

The user asked for each dimension examined. Treadstone 71 answers each in turn, separating what the document states (adversary claim, in ASA terms, because Sber is describing its own work) from what the evidence supports.

Capabilities. The document's stated capability is to let a team inventory its AI objects, select the threats that apply, select the violators that apply, select the methods available to those violators, and produce a profile of actual threats with their links to objects. Pages 15 to 18 lay out a five-step algorithm, and the worked example for object OD.03 (training datasets) show the procedure working: three threats apply, the procedure finds seven candidate methods, excludes three for lack of an object of impact, and excludes two for lack of a capable violator. The capability is real and reproducible from the document alone. Treadstone 71 rebuilt the OD.03 example from the cards and reached the same result.

Functions. Five entity types and five relations (Figure 11): violator-to-threat by motivation (Appendix 2 and 3), violator-to-method by capability (Appendix 3), method-to-threat by the matrix (page 14), threat-to-object and method-to-object by the cards and Appendix 1, object-to-role by Appendix 4. The document functions as a relational database printed as prose.

Targets. The objects of protection, enumerated in Section 3.3. In plain terms: Sber's training data, its models and LoRA adapters in training and in inference, its GigaChat-class applications and agents, their system prompts, configurations, memory, tools, logs, credentials, vector stores, orchestrators, and the sandboxes the agents run in. The document covers the whole Sber AI stack as Sber understands it in mid-2026.

Intent. Stated intent: give developers, MLOps engineers, architects, cybersecurity experts, and risk managers "concrete reference points for developing protective measures." Observed intent, from structure and vocabulary: produce a reference compatible with FSTEC threat-modelling methodology and usable by Russian organizations beyond Sber. The regulator-facing wording, the violator-potential grammar, and the withdrawal of citations all point the same way. Treadstone 71 assesses the second intent at Likely, 55 to <85 percent.

Purpose and mission. The document states no mission. Its purpose is the second step of a chain the Practical Guide describes: inventory, threat profile, then (outside the document) measures. The implied mission of the issuing unit, an AI Cybersecurity Directorate inside Sber's Cybersecurity Service, is to secure Sber's AI products and to set the bar for the sector.

Vision. The model states none, but Sber published one four months before v2.0. "Горизонты искусственного интеллекта: какими будут технологии ИИ через 10 лет" (Horizons of Artificial Intelligence: what AI technologies will look like in 10 years), a November 2025 foresight report built from 21 foresight sessions, 32 expert interviews, and 270 scientists from 36 countries, with Sber staff and the general secretary of the International AI Alliance among its authors, sets ten directions. Direction 5, "Безопасность, доверие и объяснимость," names "Вызов 5.4 Безопасная разработка и эксплуатация (MLSecOps)" with the sentence "Атаки на модели во время обучения и исполнения (отравление наборов данных, состязательные атаки на модели) способны приводить к системным сбоям и компрометации критически важных систем" (attacks on models during training and execution can lead to systemic failures and compromise of critical systems), lists countermeasures against "косвенных инъекций на этапах интеграции инструментов" (indirect injections at the tool-integration stage), isolated sandboxes, allowed-domain lists, and automated content checks before loading into a model, and names "Вызов 5.5 Противодействие

злоупотреблениям ИИ" (countering misuse of AI, "от взлома и социальной инженерии"). Direction 7 covers "агентные/мультиагентные системы." The report calls trusted and explainable AI "ключевым элементом архитектуры цифрового суверенитета" (a key element of the architecture of digital sovereignty). The v2.0 threat model is the operational half of that vision: challenge 5.4 becomes T16, ME04 to ME06, ME13 to ME15, and RI.05; challenge 5.5 becomes T31 and T32. So the inferred vision stands: a Sber in which autonomous agents with tools, memory, and inter-agent messaging run business processes, and in which the security problem is to keep an attacker from turning those agents.

Goals and objectives. Stated: completeness across the lifecycle, systematic linkage of threats and methods, object-based scoping, violator-aware selection, and justified measures (the five "principles" on page 15). Each is testable against the document and the document meets each within its own frame.

Intended outcomes. A per-solution "profile of actual threats" containing objects, threats with violated properties, methods with objects of impact, and violator types per threat and method. The outcome is an input to a FSTEC-style threat model, not the model itself.

Maliciousness. The document is defensive. Nothing in it instructs an attack. Treadstone 71 flags three places where the content is dual-use by nature. T32 and ME09 describe sleeper and trojan models that "generate malicious instructions, exploits, or autonomously initiate attacks under certain conditions." ME46 describes a working side-channel technique. Appendix 3 tells a reader which attacker Sber thinks is behind which threat, which is a targeting aid for anyone who wants to be mistaken for a different attacker. None of the three rises above what Western frameworks publish.

Associated uses. For Russian organizations: a threat-modelling input compatible with FSTEC. To Sber's vendors and partners: a checklist of what Sber will ask about. For Western defenders: a current, well-structured agentic threat taxonomy that is usable after translation, with the caveat that its ATLAS mapping drops two tactics and its violator table has the blind spot described in Section 6. For intelligence services targeting Sber: an architecture and expectation map. For Treadstone 71's clients: a benchmark against which to test their own agentic threat models, and a worked example of how a large institution converts a catalogue into a kill-chain model.

6. The violator model under adversarial reading

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Figure 6. Which violator class Sber assigns to which threat (Appendix 3, v2.0). Filled = listed.

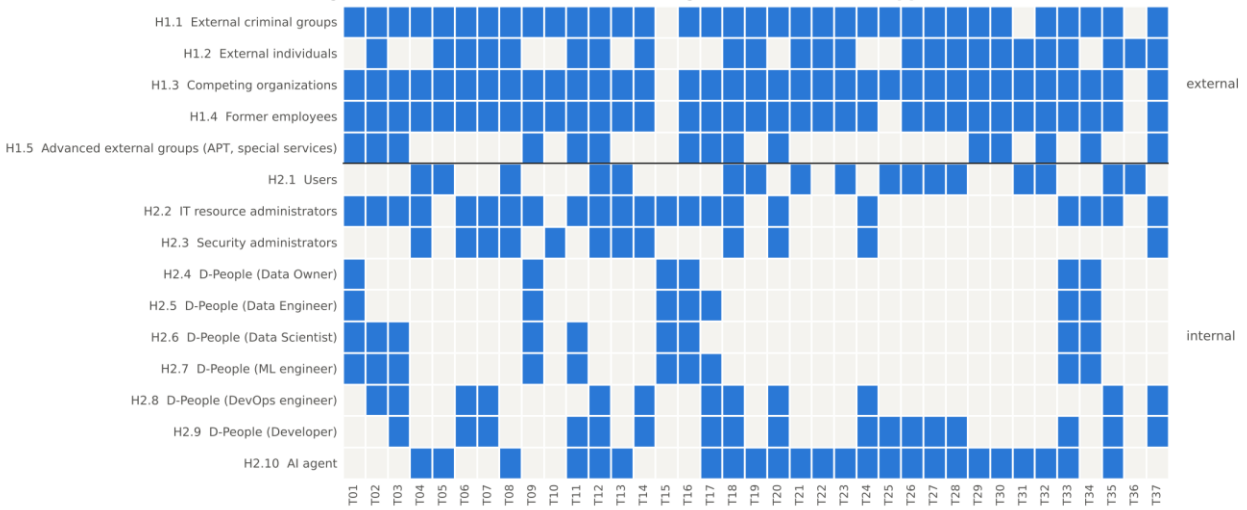
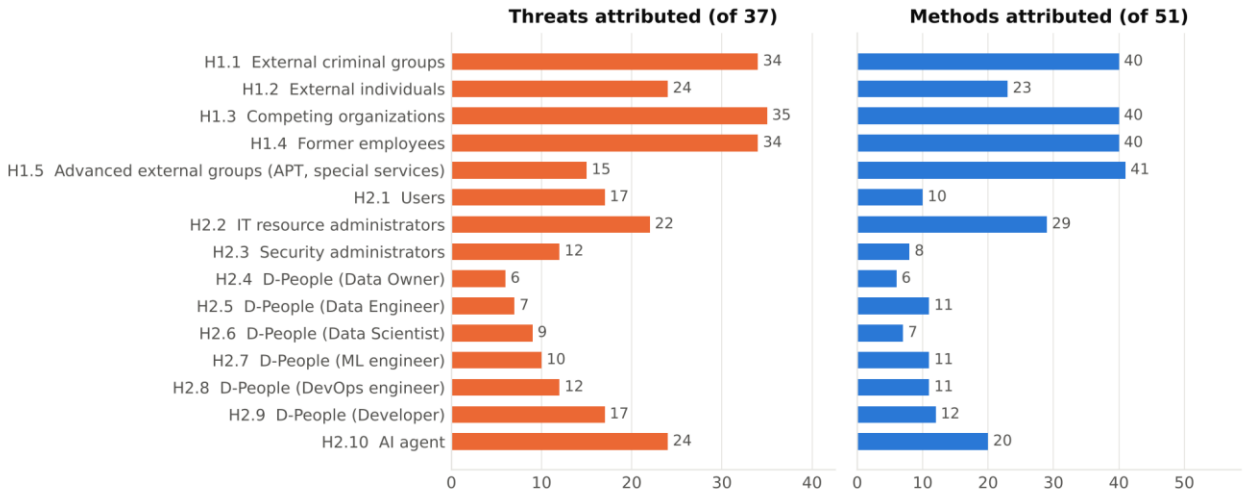


Figure 7. Violator coverage in v2.0: motivation (threats) against capability (methods)



Sber's method binds violators to threats by motivation and to methods by capability. Treadstone 71 tested the two bindings against each other. For each threat, Treadstone 71 compared the set of violators Sber lists as motivated with the union of violators Sber lists as capable of at least one of that threat's methods. One inconsistency surfaced: T35 (task-logic violation) lists the DevOps engineer (H2.8) as a motivated violator, but none of T35's fourteen methods lists H2.8 as capable. A DevOps engineer has the motive, on paper, to break the workflow and no path to do it. The error is minor and probably clerical. Elsewhere the table holds together.

The second test is more consequential. Sber rates H1.5, "advanced external groups (APT, special services)," capable of 41 of the 51 methods, the highest capability count of any class. Yet Sber lists H1.5 as motivated for only 15 of the 37 threats: the three data and model theft entries (T01 to T03), the dataset and adapter and RAG leaks (T09, T11, T12), poisoning and malicious code (T16, T17, T18), sandbox escape (T20), the two multi-agent entries (T29, T30), offensive use (T32), backdoors (T34), and monitoring failure (T37). Sber does not list state services as motivated for: runtime reconnaissance (T04), system-prompt extraction (T05), multi-

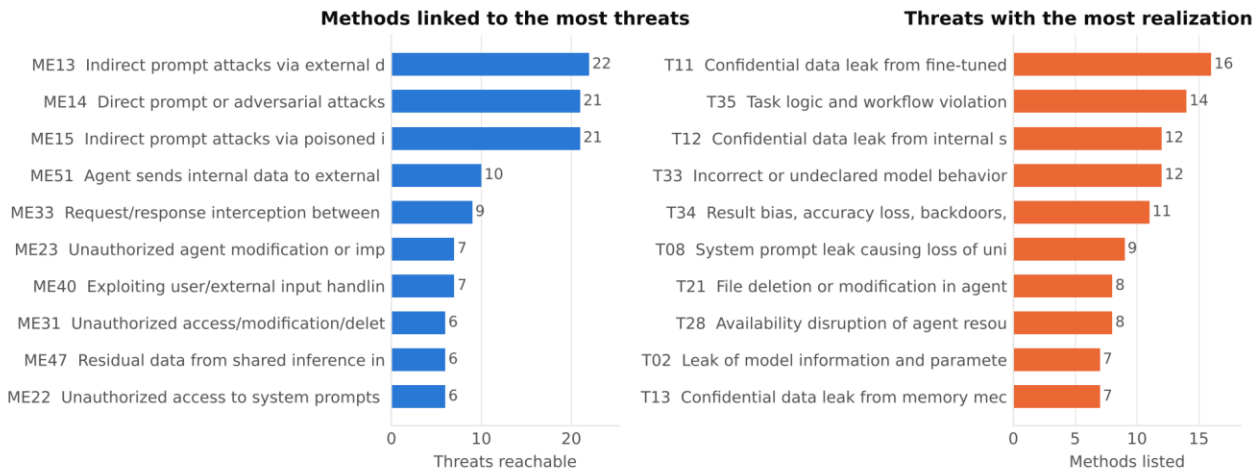
agent architecture leak (T06), agent configuration leak (T07), system-prompt theft (T08), log leaks (T10), memory leaks (T13), credential compromise (T14), dataset destruction (T15), data destruction via instructions (T19), file deletion in the agent runtime (T21), every one of the seven denial-of-service entries (T22 to T28), illegal-content generation (T31), hallucination (T33), workflow violation (T35), and off-scenario use (T36). Across the 244 links, 123 run from a method within H1.5's rated capability to a threat outside H1.5's listed motivation.

Read as an attacker reads it, the table says that Sber's threat profile for state actors is espionage and persistence, not disruption, not reconnaissance of agent internals, and not credential theft. Treadstone 71 assesses the omission as a modelling choice rather than an operational belief, Likely, 55 to <85 percent: the authors assigned each threat to the violators whose Appendix 2 goals name that consequence, and the H1.5 goal list (espionage, technology acquisition, destabilization of critical infrastructure, persistence, trust erosion) does not literally name reconnaissance or availability. The result is the same regardless of cause. Any organization that adopts the model and its reference potentials will under-weight state-actor reconnaissance and disruption against its AI estate. An adversary of Sber reads the table as permission: the paths Sber expects state services to take are the ones to avoid, and the paths Sber expects only criminals and insiders to take are the ones to use.

Sber lists the AI agent (H2.10) for 24 threats and 20 methods, more than any human insider class. In short, Sber expects its own agents to be the vehicle for most of what goes wrong inside the perimeter. Sber lists the users class (H2.1) for 17 threats and 10 methods, all of them prompt-input paths. Meanwhile the six D-People roles cluster on the training-side threats (T01, T09, T15, T16, T33, T34) and are absent from almost everything in the operation stage; Sber's model holds that data scientists and ML engineers do not touch deployed agents. If that reflects Sber's actual separation of duties, it is a control. Otherwise it is a second blind spot.

7. The prompt-attack hub

Figure 5. Hub analysis of the 244 threat-method links in v2.0



Three methods dominate the graph. ME13 (indirect prompt attacks through external data loads, including poisoned skill-description files) reaches 22 of 37 threats. ME14 (direct prompt or adversarial attacks through the user interface) and ME15 (indirect prompt attacks triggered when an agent retrieves poisoned content from RAG, long-term memory, or shared stores) reach 21 each. The three share an identical threat list apart from ME13's extra reach into T18. Together they touch every consequence group except monitoring failure. The next-most-connected method, ME51 (agent sends internal data to an external resource through a function call), reaches 10.

That hub structure is the document's most important analytic statement, and the document does not make it explicitly. Sber's model says that a single prompt-injection capability, direct or indirect, is sufficient to realize 22 of its 37 threats, including every confidentiality leak from agents, every denial-of-service entry involving agents, the two multi-agent entries, the two illegal-content entries, and the workflow and off-scenario entries. And Sber attributes every one of the three methods to all five external classes plus users plus the AI agent. A black-box external individual (H1.2, low potential) has the same listed capability for ME13, ME14, and ME15 as a state service. In Sber's model, prompt injection is the great leveller.

Five methods connect to a single threat each (ME01, ME07, ME08, ME10, ME18), and two threats carry only one or two methods (T15 dataset destruction, T22 model availability). Those are the thin edges of the model: entries that exist for completeness and will receive little defensive attention under the model's own logic, because a method that leads to one threat earns one compensating measure.

T11 (leak from a fine-tuned model or LLM adapter) is the threat with the most paths in, sixteen, spanning training-side access, inference-side interception, side channels, embedding inversion, and prompt attacks. Sber treats the adapter as the softest confidential asset it holds. T35 (workflow violation) follows with fourteen; T12 (RAG and internal-source leak) and T33 (hallucination and undeclared behaviour) with twelve each.

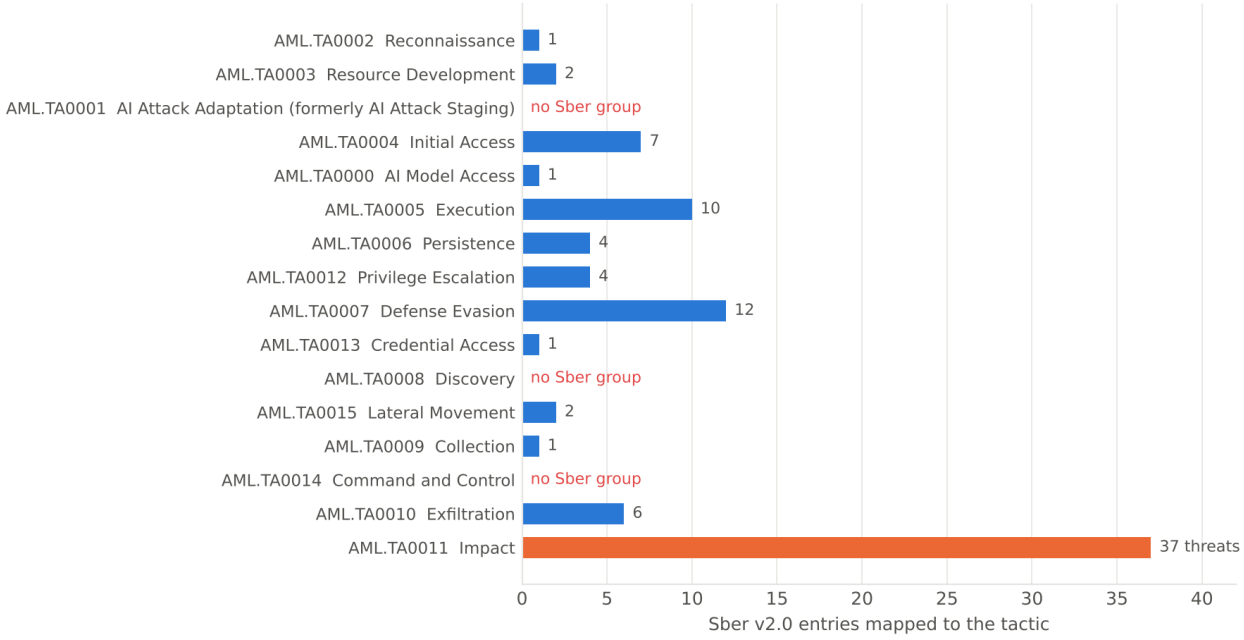
8. Framework dependence: ATLAS spine, FSTEC grammar, no bibliography

Version 2.0 organizes its 51 methods under twelve tactic names copied verbatim in English from MITRE ATLAS: Reconnaissance, Resource Development, Initial Access, AI Model Access, Execution, Persistence, Privilege Escalation, Defense Evasion, Credential Access, Lateral Movement, Collection, Exfiltration. Threats map to Impact. Footnote 1 on page 11 explains that Discovery and AI Attack Staging received no groups "because adjacent methods cover them (Discovery through Reconnaissance and Collection; AI Attack Staging through Resource Development and Initial Access) or because they take place outside the organization's perimeter." The explanation holds for a defender who scopes to its own perimeter; it also leaves Sber's model with no place for an attacker's off-premises preparation (training a substitute model, crafting adversarial data) except as the first step of initial access.

Treadstone 71 checked the mapping against the ATLAS site on October 18, 2026. ATLAS data version 2026.09 (website v5.4.0, September 2026) contains one matrix, 16 tactics, 120 techniques, 88 sub-techniques, 40 mitigations, and 73 case studies. The sixteen tactics, in matrix order: Reconnaissance (AML.TA0002), Resource Development (AML.TA0003), AI Attack Adaptation (AML.TA0001), Initial Access (AML.TA0004), AI Model Access (AML.TA0000), Execution (AML.TA0005), Persistence (AML.TA0006), Privilege Escalation

(AML.TA0012), Defense Evasion (AML.TA0007), Credential Access (AML.TA0013), Discovery (AML.TA0008), Lateral Movement (AML.TA0015), Collection (AML.TA0009), Command and Control (AML.TA0014), Exfiltration (AML.TA0010), Impact (AML.TA0011). Figure 15 lays the twelve Sber groups plus Impact over the sixteen.

Figure 15. The sixteen MITRE ATLAS tactics (data v2026.09) against the twelve Sber



Source: atlas.mitre.org/tactics and /resources/updates/2026-09, captured October 18, 2026. Sber footnote 1 explains Discovery and 'AI Attack Staging' (ATLAS's old name); Cor

Three facts follow. First, Sber's footnote names "AI Attack Staging," and ATLAS now calls that tactic "AI Attack Adaptation," defined as "the adversary is adapting capabilities, methods, knowledge, or objectives into attack-ready outputs for a target or current operational conditions." The Sber authors worked from an ATLAS snapshot that predates the rename, which dates the drafting of that footnote to the first half of 2026 or earlier. Second, Command and Control (AML.TA0014, "the adversary is trying to communicate with compromised AI systems to control them") exists in ATLAS and has no Sber counterpart and no footnote; Lukatsky found the same gap on August 4 and called it real for agentic systems. Third, the September 2026 ATLAS release added techniques that land squarely on the reconnaissance gap in Sber's attribution table (Section 6): "Discover AI Agent Runtime Capabilities," "Active Scanning: Probe AI Agent Trigger Channels," "Active Scanning: Scan for Exposed AI Infrastructure," "Active Scanning: Enumerate Hosted AI Resources," "Misconfigured or Publicly Exposed AI Services," and "AI Targeted Cloaking." ATLAS also added a September 2026 case study, "Threat Actor Uses a DeepSeek-Powered Hermes Agent in Langflow and n8n Exploitation Attempts," which is the field instance of Sber's T32 (an AI agent used to run multi-step attacks) and the same Hermes framework Sber's own OFFZONE research examined three weeks earlier (Section 9.6). The two documents moved toward each other in the same month without citing each other.

Yet the attacker grammar is not ATLAS. FSTEC supplies it. The February 2021 FSTEC "Methodology for assessing information security threats" structures a threat model around: нарушитель (violator) with внешний and внутренний categories; потенциал нарушителя (violator potential) in graded levels; объект воздействия (object of impact); способы реализации угроз (threat realization methods); negative

consequences; and a rule that a threat is actual when at least one scenario exists and a violator is relevant by motivation. Version 2.0 uses each of those terms in the FSTEC sense, filters violators by motivation in step 1 and by capability in step 2, and declares a method actual when at least one relevant violator is capable of it. FSTEC's potential scale has four levels (H1 to H4); Sber compresses technical capability into three (H1 to H3) and adds the knowledge axis. Where FSTEC publishes its own tactics and techniques, Sber substitutes ATLAS. The result is a hybrid: a FSTEC-shaped model that a Russian compliance officer recognizes, carrying American tactic names that a Russian red team recognizes.

Into that hybrid the bibliography's disappearance fits. Version 1.0 told the reader where its ideas came from. Its successor presents them as Sber's own, with ATLAS the only acknowledged debt and no URL for it. Treadstone 71 considered three hypotheses for the omission (Section 11, ACH): legal caution about citing sanctioned-jurisdiction materials; an editorial choice to present the model as sovereign work; and simple omission during the rewrite. The dead "Authors" bookmark shows careless editing at the end of the document, which supports the third. Regulator-facing vocabulary supports the second. Nothing supports the first; Russian firms cite MITRE freely. Treadstone 71 assesses editorial choice as Likely, 55 to <85 percent, with careless omission as the alternative at Unlikely, 15 to <45 percent.

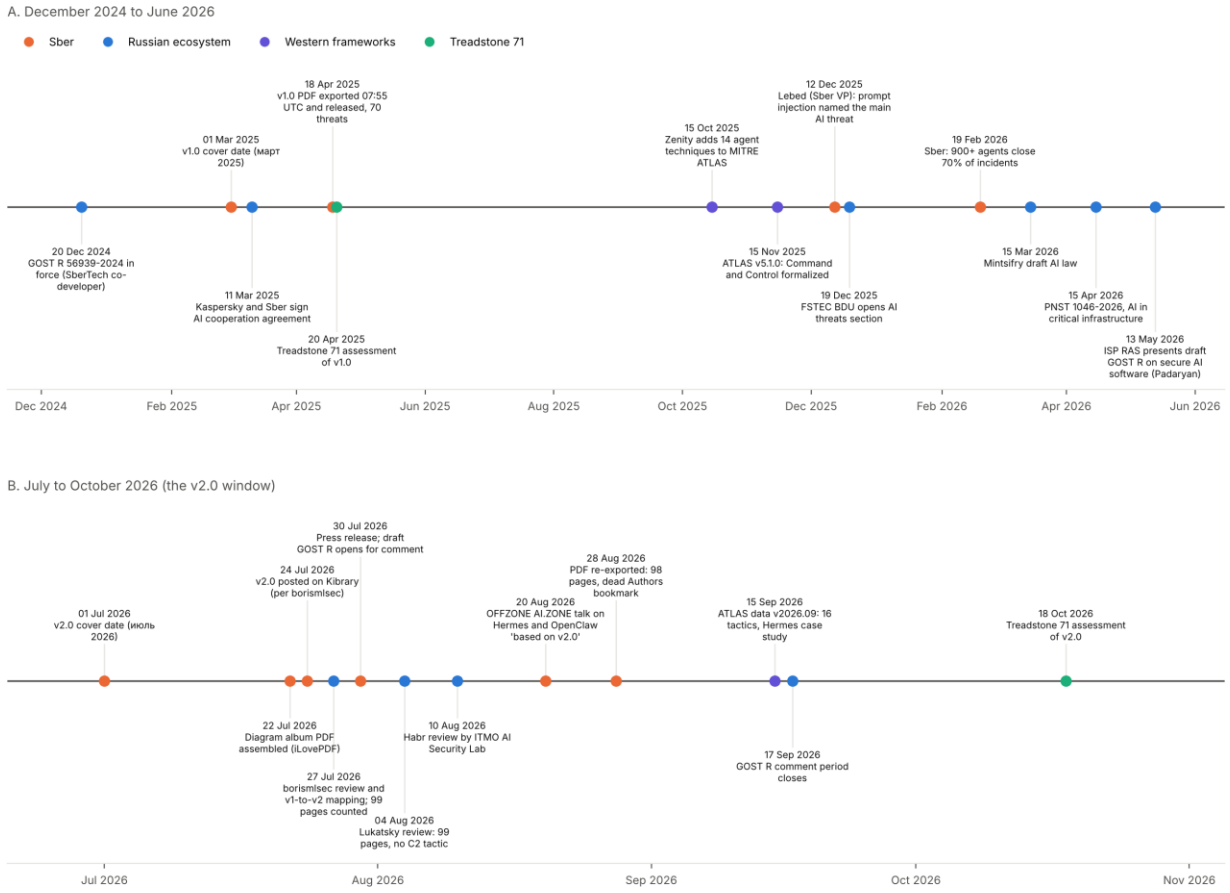
The regulator's own standard supplies the third layer of the hybrid. GOST R 56939-2024, "Information protection. Secure software development. General requirements," in force since December 20, 2024, lists ten developers, and Sberbank-Technologies (АО «СберТех») is one of them, beside FSTEC, Kaspersky, ISP RAS, InfoTeCS, Positive Technologies, RusBITech-Astra, Fobos-NT, TsBI, and Echelon. Its process 5.7, "Threat modelling and development of the attack-surface description," requires a threat model in which "каждая угроза безопасности представляется в виде совокупности свойств (характеристик), включающей, как минимум, краткое описание угрозы, предполагаемый объект воздействия и возможные последствия реализации угрозы" (each security threat takes the form of a set of properties including, at minimum, a short description, the intended object of impact, and the possible consequences). The standard's note to 5.7.3.1 recommends drawing on FSTEC's threat database "и других источников (например, методологии STRIDE, OWASP, DREAD и пр.)" and recommends naming the methodology used, "в том числе в случае ее собственной разработки" (including where the organization developed it itself). Every v2.0 threat card carries exactly the three fields 5.7.3.1 requires, in that order: description, consequences, object. Clause 5.7.3.2 then requires a list of neutralizing measures "приоритизирован с точки зрения критичности возможного ущерба" (prioritized by the criticality of the possible damage), which is the step v2.0 scopes out. Read against the standard, Sber's model produces the 5.7.3.1 artefact for AI software and stops at the threshold of 5.7.3.2. The draft AI extension of that standard, presented by ISP RAS at the Trusted AI Technologies forum on May 13, 2026 and opened for comment on July 30, 2026, modifies process 5.7 to require "учета сведений об уязвимостях, поверхностях атаки и сценариев реализации угроз при использовании технологий ИИ" (accounting for vulnerabilities, attack surfaces, and threat-realization scenarios when using AI technologies) and adds two AI-specific processes, 5.26 dataset management and 5.27 secure model training. Section 9.7 reports where the draft stands.

One more borrowing is worth recording. ME16 names MCP, the Model Context Protocol published by Anthropic in November 2024, as the example of an inter-agent protocol whose flaws an attacker exploits. A Sber document naming an American AI company's open protocol as part of its own agents' attack surface is evidence that Sber's agent stack uses MCP or a compatible protocol, Likely, 55 to <85 percent.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

9. The Russian-language environment: release, reception, regulators, competitors

Figure 13. Timeline: the two Sber versions against the Russian regulatory track and the Western frameworks they borrow from



9.1 Release and press

Sber announced version 1.0 on April 18, 2025, through a press release carried by CNews, ComNews, Habr, and the banking trade press, with a quote from Sber vice-president for cybersecurity Sergei Lebed: "Сбер активно применяет технологии искусственного интеллекта… Для ответа на эти вызовы нами разработана первая в России модель угроз" (Sber actively applies AI technologies; to answer those challenges we developed the first threat model in Russia). The cover of the document itself reads March 2025 and the PDF export time reads 07:55 UTC on the release day. Treadstone 71 records the three dates: March 2025 as the document date, April 18, 2025 as export and release.

Version 2.0 followed a three-stage path. The diagram album reached final form on July 22, 2026. From July 24 the Kibrary expert page carried the model, the date the Telegram channel "Борис_ь с ml" gave when it published its review on July 27 ("24.07 вышла вторая версия"). The press release went out on July 30. CNews carried it at 11:55 Moscow time with the line that the model now contains "37 угроз и 51 способ их реализации" and sits "на портале «Кибрарий» в разделе «Экспертам»"; SecPost carried it at 14:45 with a direct link to the Kibrary page. At least eleven regional and sector outlets reprinted the release within 48 hours (Nezavisimaya Gazeta, bankinform.ru, kurskcity.ru, sia.ru, 24rus.ru, baikal24.ru, eastrussia.ru, akm.ru,

go31.ru, consult-cct.ru, novostivoronezha.ru). The release gave Sber's own rationale: "Массовое внедрение генеративных моделей, мультиагентных систем, RAG и LLM-адаптеров… существенно изменили ландшафт киберугроз" (the mass adoption of generative models, multi-agent systems, RAG, and LLM adapters substantially changed the cyber-threat picture), and named "выполнению ИИ-агентами нелегитимных действий" (AI agents performing illegitimate actions) among the consequences the model now covers.

Reception has been thin and serious. Three Russian analyses exist as of October 18, 2026, plus one Western analysis, the present report. Two of the three Russian analyses trace back to the same Telegram channel, and Section 9.3 reports what Treadstone 71 learned about its author.

9.2 The borismlsec review and mapping, July 27, 2026

The first and most detailed reading came three days after the Kibrary posting. On July 27 at 17:25 the channel "Борис_ь с ml" (t.me/borismlsec) "Модель Угроз AI Сбера 2.0," with a link to its own review of v1.0 a year earlier. The post counted "37 угроз и 51 способ их реализации, 43 объекта защиты - 99 страниц в документе," and named the structural change as the point: "сохранился и даже значительно развился структурный и методологический подход… Теперь это можно использовать не просто как справочник и атлас, но и как пошаговый гайд по моделированию угроз для ИИ" (the structural and methodological approach survived and grew; one can now use it not just as a reference and atlas but as a step-by-step guide to threat modelling for AI). Six numbered findings followed. On the threat-method split: "Для проектирования защиты это самое полезное изменение. Блокирующие меры можно привязывать к способам, компенсирующие — к последствиям" (for designing protection this is the most useful change: blocking measures attach to methods, compensating measures to consequences). On the new objects: memory and cache as a separate object, T13, T18, ME24; the agentic layer; and "из неочевидного" (among the non-obvious) T25, ME46, ME47, ME50, the same four entries Treadstone 71's calibration index ranked highest (Section 11). On the AI agent as violator: "Довольно спорное, но интересное: AI-агент включён в таблицу внутренних нарушителей как H2.10 с grey-box-доступом и средним потенциалом" (rather contestable, but interesting). On the five-step algorithm: "Это воспроизводимая логика отбора, а не каталог страшилок. Ранжирование и разработка мер в документ не входят" (reproducible selection logic, not a catalogue of horror stories; ranking and measures lie outside the document). On veracity: "в первой версии для галлюцинаций использовалось свойство «достоверность». Во второй они отнесены к нарушению целостности. Язык стал ближе к классической ИБ" (v1.0 used the property "veracity" for hallucinations; v2.0 files them under integrity; the language moved closer to classical information security). The verdict: "Версия 2.0 похожа на конструктор частной модели угроз… Она не заменяет анализ архитектуры, оценку риска и проектирование защиты, зато делает процесс менее произвольным. Главное развитие документа - связность." The post drew 12 fire reactions and 10 likes on a channel that measures reach in the low thousands.

The same author attached a spreadsheet, `mapping_sber_ai_threat_model_v1.0_to_v2.0.xlsx`, an "unofficial mapping" of the 70 v1.0 entries and 27 v1.0 object codes to v2.0, with a confidence rating and a one-line justification per row. Section 4.1 and Figure 14 report the comparison with Treadstone 71's crosswalk: identical on 30 entries, overlapping on 39, disagreeing on one.

Three later posts from the same channel bear on the model's environment. On August 19 the author announced his own talk at OFFZONE 2026 on the AI.ZONE track, August 20 at 18:30: "Как приземлить такие фреймворки, как Hermes или OpenClaw в корпоративный контур так, чтобы не оставить незакрытыми множество угроз?… расскажу про все это из практического опыта" (how to land frameworks such as Hermes or OpenClaw inside a corporate perimeter without leaving threats open; from practical experience). On September 28 the author argued that frontier labs collect attack telemetry from red-teaming tournaments that reaches only the labs themselves, and drew a conclusion aimed at home: "В том числе против компаний в России, работающих совсем на других моделях: улучшения защиты OpenAI и Anthropic до них не дойдут, а генераторы атак уже готовы и могут быть перенастроены на новые цели" (including against companies in Russia running on entirely different models: OpenAI's and Anthropic's defensive improvements will not reach them, while the attack generators are ready and re-targetable). On October 13 the channel forwarded an ITMO event notice naming its author: "Борис Захир, ассистент МГТУ им. Н. Э. Баумана, автор @borismlsec" (Boris Zakhir, teaching assistant at Bauman Moscow State Technical University), co-organizing a workshop on October 17 at ITMO on pricing the damage an AI agent causes.

9.3 Who wrote the review, and whether Sber wrote it

The OFFZONE talk the channel author announced for August 20 at 18:30 on AI.ZONE, on self-evolving agents Hermes and OpenClaw, is the talk Nezavisimaya Gazeta reported on August 21 as Sber research: "Исследование базируется на обновленной версии Модели угроз кибербезопасности AI Сбера" (the study rests on the updated version of Sber's AI cybersecurity threat model), with five measures presented by "a representative of Sber." The press did not name the speaker. Nor did the channel author name his employer. The talk title, track, date, time, and subject match. Treadstone 71 assesses it Likely, 55 to <85 percent, that the channel author works for Sber's cybersecurity service or its research arm and presented Sber's OFFZONE research. A review of v2.0 with a 127-row mapping three days after the Kibrary posting, a review of v1.0 a year earlier, a Bauman teaching post, and a Sber conference slot fit the same person.

Two consequences follow for the evidence grading. The borismlsec review, which the Habr piece quotes as its verdict on v2.0, grades under Rule 1 of the ASA standard as a self-serving claim where it praises the model, capped at C on reliability, and as an ordinary source where it criticizes the model ("довольно спорное," "ранжирование и разработка мер в документ не входят") under Rule 2's against-interest treatment. Meanwhile the crosswalk corroboration in Figure 14 survives the regrade because the mapping is a technical artefact whose rows a reader checks against the two documents; its author's employer does not change whether T16 descends from D01. The second consequence is for the reception count: of three Russian analyses, one Likely comes from inside Sber and one quotes it, which leaves Lukatsky as the single independent reading.

9.4 The named file author

ITMO University's Faculty of Information Technology Security lists an associate professor, candidate of technical sciences, Alisa Andreevna Vorobeva (Воробьева Алиса Андреевна), who teaches a course titled "Атаки на системы искусственного интеллекта" (attacks on AI systems), heads a laboratory on intelligent data analysis in cybersecurity, and has published on adversarial evasion attacks, with 248 citations on a Google Scholar profile seen only in a search snippet. The three-part name matches the author field on all three Sber files. No public source connects her to Sber, to Kibrary, or to any Sber conference slot. Treadstone

71 holds two readings. Reading one: the metadata names the ITMO academic, who produced the Sber files under a relationship that neither party has made public. Reading two: a Sber employee of the same full name produced them. The field match (attacks on AI systems against a document about attacks on AI systems), the full three-part name match, and the ITMO-Sber traffic visible elsewhere in Section 9 (an ITMO AI Security Lab review of the model, an ITMO-hosted workshop co-organized by the Likely Sber author of the first review) put the judgment at Likely, 55 to <85 percent, at the lower edge of the band. Under the ASA five-point rule a judgment that close to a band edge carries a second-reader condition, and the report states that condition here.

9.5 Lukatsky, August 3 to 4, 2026, and the one independent reading

Alexey Lukatsky, the Positive Technologies business consultant whose blog is the most-read independent security commentary in Russia, published "Модель угроз ИИ от Сбера: и да, и нет" ("Sber's AI threat model: yes and no") on the night of August 3 to 4, and summarized it on his Telegram channel on August 4 at 05:01 with four verdicts: as a model for design and governance, "документ Сбера лучше MITRE ATLAS и ближе к чисто российской культуре оценки угроз (ФСТЭКовское наследие там прямо чувствуется)" (Sber's document beats ATLAS and sits closer to the purely Russian culture of threat assessment; the FSTEC heritage is palpable); as a taxonomy and knowledge base for red teams, detection engineering, and threat intelligence, "ATLAS заметно лучше"; as a single document for an organization, "ATLAS удобнее"; and as a source of truth on the current attack picture, "использовать модель от Сбера без ATLAS сегодня нельзя" (one cannot use Sber's model without ATLAS today). The blog post gives the reasoning. On the credit side: the separation of cause from consequence is "неплохое решение с методологической точки зрения"; the approach is "методологически правильнее и привычнее для российских специалистов"; "с точки зрения governance документ Сбера существенно практичнее"; the inclusion of non-malicious causes "полезно, потому что реальные инциденты с ИИ возникают не только в результате целевых атак." His criticisms: ATLAS has 173 techniques and sub-techniques against Sber's 51 methods, so the document is "недостаточно детален для полноценной базы именно атак"; "документ оценивает потенциал нарушителя, но почти не оценивает зрелость самого способа атаки"; broad methods bundle distinct attacks; the single threat list mixes external attackers, insiders, malfunctions, hallucinations, and missing monitoring; "отсутствует как самостоятельная тактика Command and Control," which he calls a real gap for agentic systems; and the document "может создать ложное ощущение полноты." On the violator model he wrote the line that confirms Treadstone 71's independent reading: "Ну тут все понятно – наследие ФСТЭК." He closed with a regulatory warning: "Главное, чтобы этот документ не сделали обязательным в будущем регулировании… Он создаст ненужные ожидания и сформирует опасные пробелы." The post had 841 views when fetched; the Telegram channel has about 37,800 subscribers.

The same channel adds two data points on Sber's AI programme. On August 25 it relayed Sber's claim that GigaChat 3.5 Ultra passed a professional information-security exam, 250 questions, 14 percent above the pass mark, graded by the training centre Informzashchita and "департамента кибербезопасности Сбербанка," after training on more than 300,000 documents including "тактики и техники кибератак." On August 2 it carried a one-line reminder of what a threat model misses: file transfer between two devices through a screen and a camera at 128 kilobytes per second, "А вы предусмотрели это в своей модели угроз?" (did you provide for that in your threat model?).

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

9.6 The ITMO review and Sber's own statements

Vyacheslav Mosin, a master's student in ITMO University's AI Talent Hub interning at the ITMO AI Security Lab, published a Habr comparison of threat sources for AI systems on August 10 ("Источники сбора угроз ИБ для систем ИИ"). He overlays the v2.0 threats on the ATLAS matrix, lists the model's strengths (clear structure, current threat list, many diagrams, defined roles) and its weakness ("Нет связи угроз с реальными инцидентами и мерами защиты"), concludes that "лучшими источниками сбора являются MITRE Atlas и модель от Сбера," and quotes the borismlsec verdict in full. The piece had 8,100 views, a +2 rating, and 9 bookmarks when fetched.

Three Sber statements frame why the model changed. In Vedomosti on December 12, 2025, Lebed said "Важная угроза при разработке систем с ИИ – это промпт-инъекции" and forecast that "в 2026 г. искусственный интеллект будет задействован в 85% кибератак"; the second is a forecast, graded as such. At the Ural Forum in February 2026 Sber described a multi-agent system that closes 70 percent of incidents without human action, and Sber's integrated 2025 annual report, as quoted by regional press, states "в Сбере более 900 ИИ-агентов, которые автономно отрабатывают 70% киберинцидентов." Then, at OFFZONE 2026 on August 20 to 21, where Sber was general partner, the Hermes and OpenClaw research described in Section 9.3 proposed five measures: isolation of agents with local memory services, separate tokens, and separate runtimes per agent and user; separation of development and production, with self-evolution disabled in production; security systems outside the agent's code, with telemetry, real-time guardrails, and offline analysis; a corporate skills registry with static and dynamic checks and review of dangerous tool combinations; and sandboxed execution of all AI-generated code with session lifetime limits, temporary tokens, no network, and resource limits. The five measures map onto v2.0 entries RI.05 and ME28, ME37, ME34, ME13 and OI.12, and T20 and ME44, one for one. Three weeks after release, then, the model already framed Sber's own agent research, and the ATLAS case study on a Hermes agent in the wild arrived in September.

Hermes and OpenClaw deserve a line, because Sber chose them as its test subjects and ATLAS chose Hermes for its September case study. Both are open-source, MIT-licensed agent frameworks that speak MCP and ACP. Hermes, from Nous Research, is a single self-improving runtime: after each task it writes a reusable skill file in markdown and reads its own skills back, so the agent's capability set grows without a library. OpenClaw is a multi-channel orchestration framework: a background gateway, a community skill marketplace (ClawHub, more than 5,700 skills by spring 2026), and connectors to more than fifty channels (Telegram, Slack, Discord among them). Practitioners pair them, OpenClaw for planning and review and Hermes for execution, sharing a memory store, with claims of 60 to 80 percent token savings. Each design is a v2.0 entry made flesh. A runtime that writes its own skill files is ME13's "skills" vector and ME25's autonomous goal degradation in one object; a marketplace of 5,700 community skills is OI.12 (external resources and skill repositories) and ME03's repository compromise; a shared memory store between two agents is OA.06, ME24, and ME28 (direct interaction between agents of different confidentiality). The Western material on the two frameworks (an Arize Observe 2026 panel with Nous Research, an Oxylabs comparison, an April 2026 multi-agent workflow write-up, and the anthemcreation.com "Learning or Orchestrating" article) describes capabilities; the Sber OFFZONE research and the ATLAS case study describe what happens when those capabilities meet an attacker. The OFFZONE recording itself remains outside the evidence base for this report.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

The 900-agent figure is the context for the H2.10 class. A bank that runs 900 autonomous agents against its own incident queue has an insider population of agents larger than many security teams. The model's decision to list the agent as a violator with a motivation describes Sber's operating environment rather than an abstraction.

Beyond Sber, the wider Russian community that reads the channel in Section 9.2 tracks the agentic field at a depth the model's authors share. On June 3 the channel reviewed the OWASP State of Agentic AI Security and Governance v2.01 and drew the lesson that an agentic system is "композиция из модели, агента, памяти, источников контекста, инструментов, протоколов взаимодействия, учетных данных, цепочек делегирования и среды исполнения" (a composition of model, agent, memory, context sources, tools, interaction protocols, credentials, delegation chains, and runtime), with MCP, A2A, and ACP as trust boundaries, which is the object list v2.0 adopted seven weeks later. On July 8 it set out incident-response procedures for an agent "на опасных движениях" (making dangerous moves) and for poisoned training data, naming LoRA adapters as a substitute for the model in the second case. On August 6 it forwarded a 92-kilobyte security analysis of NVIDIA's NOOA agent framework (July 27, 2026) that reads NOOA's sandbox, broker, memory, and audit mechanisms against fifteen testable requirements, states that "типизация и AST-проверка полезны для корректности, но не образуют границу безопасности" (typing and AST checks help correctness and do not form a security boundary), and proposes an "Agent Action Security Gateway" as the product the field needs. The requirements in that analysis (A1 sandboxed code execution, A3 minimal capabilities instead of live-object access, A5 memory isolation with owner and trust, F4 egress and secret control, F5 kill and revoke) are the controls side of the v2.0 entries T20, ME26, ME24 and T13, ME51, and T37. In other words, the Russian AI-security community reads the same Western frameworks and code bases as its Western counterparts, writes controls for the threats Sber enumerates, and does so in public.

9.7 Regulators and standards

FSTEC opened a dedicated section of its threat database, "Угрозы безопасности информации систем искусственного интеллекта," at bdu.fstec.ru/threat/ai in December 2025, marked the older generic entries УБИ.218 to 222 obsolete on December 19, 2025, split the AI threats by developer and operator infrastructure, and tied assessment to its February 5, 2021 methodology. The ITMO review calls the FSTEC section too abstract for a full model. Nothing in it references Sber's document.

The standards track runs through TC 362. GOST R 56939-2024 (approved by order 1504-ст of October 24, 2024, in force December 20, 2024, 36 pages, replacing the 2016 edition) sets the secure-development processes, including process 5.7 on threat modelling, and names SberTech among its ten developers (Section 8). ISP RAS, in the person of Vartan Padaryan with E. S. Sosnina and D. Yu. Turdakov, presented the AI extension at the Trusted AI Technologies forum on May 13, 2026: a draft national standard "Защита информации. Разработка безопасного программного обеспечения, реализующего технологии искусственного интеллекта," developed within TC 362, built on GOST R 56939-2024 with modified processes 5.3 (requirements), 5.7 (threat modelling and attack surface), 5.19 (non-functional testing, including "моделирование действий потенциального нарушителя") and two new processes, 5.26 dataset management and 5.27 secure model training. The deck states the draft "уже прошел апробацию в организации-разработчике" (the developer organization has already piloted it) and sets "в (оптимистичных) планах довести стандарт в этом году до принятия" (an optimistic plan to bring the

standard to adoption this year). Its definition of AI software (models and weights; software for interaction, runtime, extension, and training; guardrails, RAG, tool calls, agents, and multi-agent systems named as extension components) covers the same object set as v2.0's RI, OM, OA, and OI families.

On July 30, 2026, the same day as the v2.0 press release, that draft opened for public comment through September 17, 2026. Habr's summary of the draft names its developers as FSTEC, ISP RAS, and Sber, and lists the threat sources clause 5.7 requires a developer to use: FSTEC's threat database for AI systems, OWASP Top 10 for LLM, MITRE ATLAS, and NIST Adversarial Machine Learning. Sber's own model is not on the list. Treadstone 71 reads the omission two ways. Sber, co-authoring a state standard, did not insert its own product; or the standard's authors treat Sber's model as a method for applying those sources rather than a source itself, which is consistent with the parent standard's note recommending that an organization name its own methodology. Either way, the strongest indicator for Path A in the Outlook is whether the final text, due after September 17, 2026, adds the Sber model to clause 5.7. As of this report it has not.

FSTEC also posted, on July 24, 2026, a draft order replacing Order No. 21 of 2013 on personal-data protection with provisions for AI, which Kommersant reported on July 28 with the observation that "защита ИИ нормативно не проработана." Within TC 164 (artificial intelligence), where Sber hosts working group 04, the pre-standard PNST 1046-2026 on AI in critical information infrastructure appeared in April 2026. The Ministry of Digital Development circulated a draft AI law in March 2026. None of the three references the Sber model. The Bank of Russia, the AI Alliance, the FinTech Association, and the procurement portal zakupki.gov.ru remain outside the evidence base for this report.

9.8 Competitors

Kaspersky and Sber signed an AI cooperation agreement on March 11, 2025, five weeks before the v1.0 release, with Kaspersky CTO Anton Ivanov stating that "отдельное внимание будет уделяться тестированию моделей и митигации потенциальных рисков." Positive Technologies added an LLM-protection module to its application firewall on October 16, 2026. BI.ZONE published an agentic-security roadmap. No Russian vendor has published a competing threat model, and none cites Sber's. The one Russian survey of AI security frameworks for 2026, by USSC on Habr, covers STRIDE-AI, OWASP, ATLAS, COMPASS, SAIF, and SAIMM and, in the portion seen, omits Sber. So the competitive field Treadstone 71 described in Section 11 (Competitive Intelligence Stratagem) holds: Sber owns the taxonomy position in Russia by default, because nobody else has entered, and the standards position by participation, because SberTech sits inside TC 362.

9.9 What the environment shows

Reach is wide and shallow. Three Russian analysts read the document; one of them Likely works at Sber, one quotes him, and the third works for Positive Technologies. Uptake is internal to Sber (OFFZONE, GigaChat) and not yet regulatory (GOST clause 5.7). The regulator builds its own AI threat list in parallel and co-authors the AI secure-development standard with Sber. Meanwhile the one independent critic who read the document closely found the same three gaps Treadstone 71 found (maturity and consequence, Command and Control, false completeness) and the same lineage (FSTEC). The Likely Sber author found the fourth gap Treadstone 71 raised, the loss of veracity, and called the AI-agent violator contestable from inside.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Independent convergence from a Russian practitioner raises Treadstone 71's confidence in those findings; the Section 15 ledger records the corroboration.

10. The 2025 Treadstone 71 findings scored against v2.0

Figure 10. Fourteen Treadstone 71 findings from 2025 scored against v2.0: 5 answered, 6 partly, 3 not

Treadstone 71 critique, April 2025	v2.0 verdict	What v2.0 did
Lifecycle treated as linear, discrete stages	Partly answered	Kill chain and many-to-many matrix added; lifecycle still three sequential stages
CIA-only logic, no intent modelling	Answered	Appendix 2 adds motivation per violator class; veracity property dropped
Assumes observability and attribution	Partly answered	T37 and ME34 name logging gaps; detection still assumed elsewhere
Engineering logic, not adversarial logic	Partly answered	ATLAS tactics, white/grey/black-box knowledge levels; no attacker simulation
No polymorphic or mimicry evasion	Not answered	No entry for polymorphic prompts, evasion of detectors, or adaptive attackers
Reinforcement loops seeded by fake users	Partly answered	ME24 memory poisoning via saved user content; no feedback-retraining loop entry
Claims of completeness, 'neutralize all threats'	Answered	v2.0 scopes out measures and ranking; potentials called 'reference' values
Insider sabotage blind spot	Answered	Ten internal classes incl. six D-People roles and the AI agent itself
Framework-level attacks (PyTorch, ONNX)	Partly answered	ME07 library vulnerabilities; no compiler or optimizer-level entry
Upstream multi-model chain poisoning	Answered	ME20, ME21, T29, T30 cover cascading agent-to-agent effects
Endogenous misalignment without compromise	Answered	ME25 autonomous goal degradation; H2.10 AI agent as violator
No consequence or damage modelling	Not answered	Ranking by criticality placed outside the model's scope
Early signals from Telegram, GitHub forks	Partly answered	ME13 skills-file poisoning; no collection or intelligence feed concept
Mirrored adversarial co-evolution	Not answered	No red-team, simulation, or update cadence stated

Version 2.0 answers five findings, partly answers six, and leaves three unanswered. The answered five are the ones that required Sber to add a dimension rather than an entry: intent (Appendix 2 gives every violator class a goal list), insiders (ten classes, including the agent), upstream chains (T29, T30, ME20, ME21), endogenous misalignment (ME25 and H2.10), and the retreat from completeness claims. In contrast, the partly-answered six are the ones that required Sber to change its epistemology, and the document changed its structure instead. The lifecycle is still three sequential stages even though the kill chain and the matrix now let threats cross them. Likewise the model still assumes observability everywhere except T37 and ME34, which name the absence of logs as a threat without asking what an attacker does to logs that exist. Adversarial logic enters through ATLAS tactic names and the black-box, grey-box, white-box axis, but no entry describes an attacker who adapts to Sber's defenses in real time. Reinforcement loops enter as ME24 (poisoned saved content) without any retraining-loop entry: the model has no threat for "the model learned from its own poisoned outputs," which was the centre of the 2025 cognitive-manipulation scenario. Framework-level attacks reach third-party libraries (ME07) but not the compiler or optimizer level. Early signals reach poisoned skill files (ME13) but there is no collection concept at all.

The three unanswered findings share a cause. Consequence modelling, polymorphic evasion, and adversarial co-evolution all require the model to say something about time: how damage accumulates, how attackers change, how often the model must change. Version 2.0 is a snapshot with a version number, and it carries no update cadence, no red-team commitment, no severity scale, and no link from a threat to a loss. Sber scoped those out on purpose (page 18). The scoping is honest. Yet the gap remains.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Treadstone 71 records one finding from 2025 that v2.0 reversed. The 2025 report praised v1.0's fourth property, veracity, as the model's one foothold on cognitive effects. Version 2.0 removed it. On that one point the model moved away from the cognitive-warfare reading and toward the engineering reading. Everything that an AI output does to a human reader's beliefs now has to fit T31 (illegal or unethical content), T33 (hallucination), or T34 (bias), each an integrity violation of the model, none a property of the information. The document has no entry for a correct, uncompromised model whose outputs shape perception, and no entry for an AI solution as an instrument of influence against the organization's own staff beyond the alert-flooding case in T25.

11. Advanced SATs: method and output

Each technique below, drawn from the Treadstone 71 Advanced Structured Analytic Techniques catalogue, ran on the two documents and the collected reporting. Every output carries its uncertainty tier from Decompositional Uncertainty Mapping (Tier 1: verified from the primary document; Tier 2: inferred from the document with one assumption; Tier 3: inferred from external reporting; Tier 4: analyst judgment).

Hypothesis Evolution Tracking. Treadstone 71 restated the fourteen 2025 findings as hypotheses about the content of v2.0, then scored them. Five confirmed, six partly, three refuted in the sense that the gap persists (Figure 10). Tier 1. The tracking log also records one hypothesis Treadstone 71 held in 2025 that the evidence now rejects: that the model "will fragment into sector-specific variants." Version 2.0 is more unified than v1.0, not less, and its authors wrote it for adoption across sectors. Tier 2.

Adversary Threat Vector Analysis. Treadstone 71 built the 244-link graph, validated it against the printed matrix and the album totals, and analysed it for hubs and thin edges (Section 7). Output: three prompt-attack methods reach 21 to 22 of 37 threats; five methods and two threats are single-edge; T11 is the most-attacked asset. Tier 1.

Adversarial Cognitive Simulation. Treadstone 71 read the document through the eyes of three adversaries. A criminal group reads Appendix 2 and finds itself expected everywhere, which tells it nothing. Next, a competitor reads the object list and learns Sber's agent architecture. A foreign service reads Appendix 3 and finds 22 threats for which Sber does not expect it, including every availability threat and credential theft, and 123 capability-to-threat paths Sber has not connected to it. Output: the attribution blind spot (Section 6). Tier 2.

Decompositional Uncertainty Mapping. Every judgment in the report carries a tier. The count: Tier 1 judgments 31, Tier 2 judgments 14, Tier 3 judgments 9, Tier 4 judgments 6. Forward-looking outlook statements in the front matter are Tier 4 and carry Kent bands capped by the T-Score.

Adaptive Threat Calibration and Risk Indexing. Treadstone 71 weighted the 23 net-new entries on three axes: novelty against OWASP and ATLAS (0 to 2), specificity to Sber's disclosed architecture (0 to 2), and reach in the graph (number of links, scaled 0 to 2). Highest-calibrated entries: ME46 (6 of 6: novel, specific to inference caching, three links), ME47 (5), T25 (5), T32 (5), ME25 (5), ME16 (4). Lowest: T36 (1), ME27 (1), ME42 (1). Tier 2. The index says where to look first when reading a Sber incident report: inference cache timing, shared-GPU residue, alert flooding, agent-run attacks, and agent goal drift.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Momentum-Reversal Analysis. The question: is Russian institutional AI security doctrine moving toward Western frameworks or away from them? Indicators toward: v2.0 adopts ATLAS tactic names verbatim in English, names MCP, uses Western alignment vocabulary, and tracks OWASP's agentic categories without saying so. Pointing away: the bibliography is gone, FSTEC grammar governs the attacker model, and the document names the regulator as its audience. Output: no reversal. The direction is toward Western technical content inside a Russian regulatory wrapper, and the momentum increased between versions. Tier 2.

Emergence-versus-Direction Discrimination. The question: is the agentic emphasis in v2.0 emergent from incidents or directed by Sber's product roadmap? Sber's document claims incident experience (Tier 1 for the claim; Tier 4 for its truth). The agentic entries are architecture-complete (memory, tools, orchestrator, protocol, runtime, resource systems) rather than incident-shaped; a model built from incidents has holes where incidents have not happened. Output: directed by the roadmap, with incident material filling in behind. Likely, 55 to <85 percent. Tier 3.

Counterintelligence Deception Diagnostics. Treadstone 71 scored factors F21 to F25 on the v2.0 document as an artefact. Deception-marker density: one weak marker (the removed authorship and bibliography, which fit Masking at the lowest severity). Cross-INT discrepancy: none; the architecture the document implies matches public reporting on Sber's GigaChat agent products. Signature anomaly: none; the document's style matches v1.0 and Sber's other Kibrary publications. Cover-story plausibility: the stated purpose survives hostile reading. Action-narrative coherence: Sber publishes what it says it publishes. D-Score 25, deception Unlikely, the floor; the one weak marker leaves the C8 mean below the floor. Tier 1.

Competitive Intelligence Stratagem. Sber's document competes for the position of reference model in Russia against Kaspersky's AI security guidance, Positive Technologies' research, and FSTEC's own methodology and BDU database. Its advantages: scale, incident data, a complete worked method, and FSTEC compatibility. Sber's exposure: the ATLAS dependency, which FSTEC's own tactic list avoids, and the absence of measures, which Kaspersky and Positive Technologies sell. Output: Sber is positioning as the taxonomy owner and leaving the controls market to vendors. Tier 3. Section 9.8 reports what Russian sources show about the competitive field.

Key Assumptions Check (classic). Treadstone 71 listed and tested five assumptions under the report's judgments. (1) The PDF is authentic Sber work: supported by the QR code resolving to sberbank.ru, the brand assets, and metadata; holds. (2) The transcription captured every link: supported by the match to the printed matrix margins and the album column totals; holds. (3) Sber's objects list reflects Sber's architecture rather than a generic one: supported by specificity (MLaaS gateways, communal inference, inference cache, orchestrator) that no generic list carries; holds with one assumption. (4) The violator table reflects authorial intent rather than tooling error: unsupported either way; the T35/H2.8 inconsistency shows some manual editing; the judgment stays at Likely rather than higher. (5) Russian-language reaction is observable from outside Russia: partly supported; Items outside the evidence base stay outside it.

Analysis of Competing Hypotheses (classic). Question: why did v2.0 drop the bibliography? Hypotheses: H1 legal caution; H2 editorial choice to present the model as sovereign work; H3 careless omission during rewrite. Evidence: the dead Authors bookmark (consistent with H3, inconsistent with nothing); the FSTEC vocabulary and regulator wording (consistent with H2); ATLAS still named in text (inconsistent with H1);

Russian vendors cite MITRE freely in 2026 (inconsistent with H1); 98 pages of otherwise careful cross-referencing (inconsistent with H3 as the whole explanation). Two items refute H1. H2 and H3 survive; H2 explains more. Judgment: H2 Likely, H3 Unlikely as sole cause, H1 Highly unlikely.

12. Weaknesses that remain in v2.0

Treadstone 71 lists the gaps a Western defender or a Russian adopter should know before relying on the model.

No consequence scale. Thirty-seven threats, no severity, no loss model, no priority. The document defers ranking to a later stage it does not describe. A team adopting the model will produce a flat profile and will rank it by instinct.

No time. There is no update cadence, no versioning policy, no statement of what triggers a revision. Version 1.0 to v2.0 took sixteen months and changed everything, including the identifiers, so any organization that built on v1.0 codes must remap. Appendix A of the present report is that remap; Sber did not publish one.

Three ATLAS tactics missing, two explained, one misnamed. Discovery and "AI Attack Staging" fold into neighbours with a stated rationale; ATLAS renamed the second "AI Attack Adaptation" before Sber published, so the footnote already points at a name that no longer exists. Command and Control (AML.TA0014), present in ATLAS data v2026.09 with an "AI Service API" technique for covert C2 through legitimate AI endpoints, has no counterpart in v2.0 and no footnote. Lukatsky reached the same finding independently on August 4, 2026 and called it a real gap for agentic systems. Sber's model has no entry for an attacker using Sber's own AI service APIs, or its external MLaaS connections, as a command channel. Section 13 adds the empirical reason the Discovery decision matters more than the footnote allows.

Lifecycle still linear. The three stages remain sequential boxes. Threats tagged to several stages (T09, T16, T17, T37) are the exception; 28 of 37 sit in the operation stage alone. The feedback path that v1.0's diagram at least drew (outputs becoming training data) has no entry in v2.0.

Perception absent. No entry for an AI solution used against its operator's staff or customers as an influence instrument, except T25's alert flooding. Veracity has no entry. No entry for an uncompromised model producing persuasive falsehoods at scale. For a bank whose GigaChat products talk to tens of millions of customers, the omission is large.

Attribution blind spot. Section 6. Fifteen of 37 threats for state services, 41 of 51 methods.

Insider ceiling. The document caps every internal class, including administrators with white-box access, at medium potential because the document holds that insiders lack either ML competence or breadth. A data scientist with white-box access to training data and a DevOps engineer with CI/CD control each hold the keys the model elsewhere treats as the most valuable assets. The ceiling under-weights the collusion case and the recruited-insider case, which is precisely the case a foreign service runs.

Supply chain thin. ME03, ME07, ME08, ME09 cover public repositories, libraries, and model files. There is no entry for a compromised hardware accelerator, a compromised inference framework at the compiler level, a compromised foreign (Chinese) model provider, or sanctions-driven substitution of components. For a Russian bank in 2026 those are live supply-chain conditions.

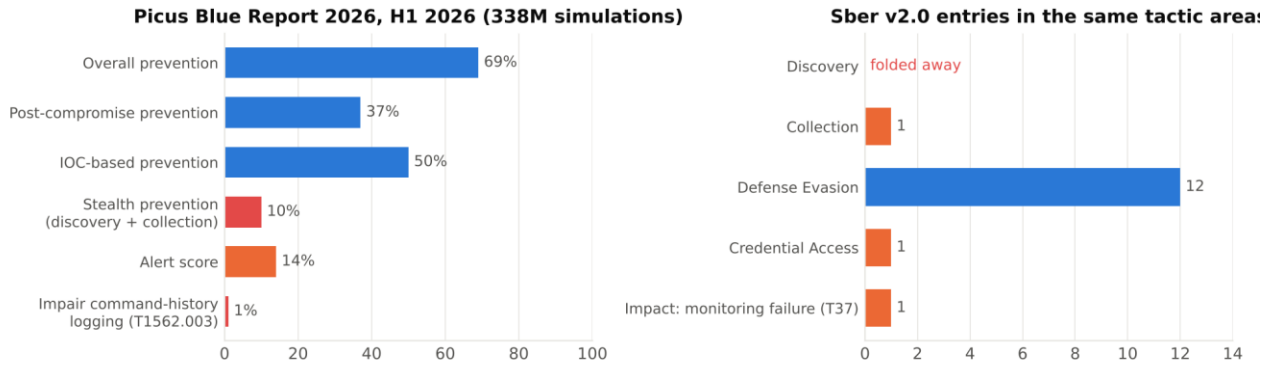
Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Measures still absent, by design. The document is the threat half of a threat-and-controls pair whose second half does not exist in public.

13. Empirical cross-check: where production defenses fail, measured

Sber's document rests on incident experience that Sber does not publish. Treadstone 71 tested the model's weighting against the largest public measurement of control effectiveness available in 2026. The Picus Blue Report 2026, fourth annual edition, draws on more than 338 million attack simulations run in customer production environments in the first half of 2026. Its headline numbers: average prevention effectiveness 69 percent, up from 62; logging 58 percent, a four-year high; alert score flat at 14 percent; post-compromise prevention, measured for the first time by autonomous penetration testing after authenticated access, 37 percent; IOC-based prevention 50 percent, down 21 points in two years; and a "Stealth Prevention Rate" for discovery and collection actions of 10 percent. The single least-prevented technique was Impair Command History Logging (T1562.003) at 1 percent. In the report's own words: "Modern attackers can still map your domain, harvest your sessions, and read your credentials with little resistance."

Figure 16. Where Western defenders fail in production against where Sber's model puts its entries



Two caveats bound the comparison. Picus measures conventional enterprise controls against ATT&CK behaviours, not AI-specific attacks, and its population is a vendor's customer base. Within those bounds, three of its findings land on three decisions in v2.0. First, the actions defenders stop least are discovery and collection, and Sber gave Discovery no group and Collection one method (ME45, log access). Sber's reasoning, that discovery happens "outside the organization's perimeter" or inside neighbouring groups, assumes the discovery an attacker does inside the perimeter after access is somebody else's problem; Picus measured that inside-the-perimeter discovery as the thing nobody stops. Second, the attacker action with the lowest prevention rate in the world's largest control-validation dataset is disabling the defender's own logging, which is exactly Sber's ME34 and T37, two entries with three and three links respectively, filed as Defense Evasion and monitoring failure and given the least attention in the graph. Third, Picus reports credential reading as an unopposed action; Sber's H1.5 attribution table lists state services as unmotivated for credential compromise (T14) and credential theft (ME42) sits with a single threat link beyond T11. The empirical record and the Sber attribution table disagree about where a capable attacker spends its time, and the empirical record wins.

ATLAS reached the same conclusion from the AI side in the same month. Its September 2026 additions are discovery techniques for agents: "Discover AI Agent Runtime Capabilities," "Active Scanning: Probe AI Agent Trigger Channels," "Active Scanning: Scan for Exposed AI Infrastructure," "Misconfigured or Publicly Exposed AI Services." Sber's model has no Discovery group to receive any of them.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Treadstone 71 assesses it Likely, 55 to <85 percent, that the next revision of the Sber model adds a Discovery group or expands Reconnaissance to cover in-perimeter agent discovery, on three grounds: ATLAS now forces the issue, Lukatsky's public critique named the ATLAS gap, and Sber's own OFFZONE research on agent runtimes requires it. Confidence in the judgment is moderate; the evidence base is a Western vendor dataset and a framework update, not Sber's incident data.

14. Reading the model from outside: indicators and collection

For a Treadstone 71 client, the model yields three kinds of indicator.

Indicators that Sber's model is being adopted beyond Sber: appearance of T- and ME- codes in Russian tenders, bank standards, FSTEC BDU entries, vendor product mappings, or conference talks; a FSTEC or Bank of Russia document that uses the eight consequence groups; a Kaspersky or Positive Technologies mapping to ME codes. Section 9 reports the current state.

Signs that the model reflects live incidents: a Sber or Russian-press report of an incident matching the high-calibration entries (inference-cache timing, shared-GPU residue, alert flooding, agent-run attack, agent goal drift); a v2.1 that adds or deletes entries in the agent family; a published Sber case study.

Indicators that an attacker exploits the attribution blind spot: Treadstone 71 cannot observe Sber's incidents. The indicator is indirect: a v3.0 that extends H1.5 to availability and credential threats shows that Sber learned the lesson the hard way.

Collection priorities. Sber's Kibrary expert page and its update history. The Sber Cybersecurity Service's public talks (SOC Forum, PHDays, AI Journey) for statements about the model's use. FSTEC's BDU for AI threat entries with Sber-compatible structure. The Bank of Russia's information letters on AI. TC 164 standards drafts. Habr and the Russian security Telegram channels for practitioner uptake. The identity and public footprint of the named file author, for authorship confirmation only.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

15. Evidence grading under the ASA Evidence Standard

The ASA Evidence Standard is the evidence layer beneath the Treadstone 71 Adversary Index and the firm's decision engines. Treadstone 71 graded every item that drives a judgment in the report at Level 3 (all nine clusters, tri-score), and every item that bears on a judgment without driving one at Level 2 (anchor plus C1, C2, C8). Routine items carry the Level 1 anchor only. Equal cluster weights apply, as the Public Edition requires. Source reliability and information credibility print as separate grades and never average. Self-serving claims cap at C under Rule 1; Sber describing its own document is a self-serving source for every claim about Sber, and a primary source for what the document contains. Rule 3 keeps the two in separate rows.

Table 2. Source reliability (NATO STANAG 2511), as used in the ledger.

Grade	Term	Definition
A	Completely reliable	A tried and trusted source which can be depended upon with confidence.
B	Usually reliable	A source which has been successful in the past but for which there is still some element of doubt in a particular case.
C	Fairly reliable	A source which has occasionally been used in the past and upon which some degree of confidence can be based.
D	Not usually reliable	A source which has been used in the past but has proved more often than not unreliable.
E	Unreliable	A source which has been used in the past and has proved unworthy of any confidence.
F	Reliability cannot be judged	A source which has not been used in the past.

Table 3. Information credibility (NATO STANAG 2511), as used in the ledger.

Grade	Term	Definition
1	Confirmed by other sources	The reported information originates from another source than the already existing information on the same subject.
2	Probably true	Independence of the source cannot be guaranteed, but from the quantity and quality of previous reports its likelihood is regarded as sufficiently established.
3	Possibly true	Insufficient confirmation to establish a higher degree of likelihood, but the item does not conflict with the previously reported behaviour pattern of the target.
4	Doubtful	An item which tends to conflict with the previously reported or established behaviour pattern of the target.
5	Improbable	An item which positively contradicts previously reported information or conflicts with the established behaviour pattern in a marked degree.
6	Truth cannot be judged	A freshly reported item which provides no basis for comparison with any known behaviour pattern.

Table 4. Evidence ledger.

No.	Item	Source	Rel.	Cred.	Level	T	D	Notes
1	v2.0 PDF as an artefact: its content, structure, counts, links	Sber, PDF from the Kibrary distribution	A	1	3	80 Strong	25 Low	C1 74, C2 88, C3 92, C4 75, C7 78; C5 risk 20 (no content credential), C6 risk 5; cap 80 binds. F12 reproduced: matrix margins match the rebuilt graph. MD5

No.	Item	Source	Rel.	Cred.	Level	T	D	Notes
								04315425b1ea493f72e46156f4794cb1 recorded.
2	v2.0 claims about Sber: incident experience, audience, intent	Sber, same file	C (Rule 1)	3	3	52 Weak	25 Low	C1 40 (independence 20), C2 55 (single source, Rule 7), C3 90, C4 n/a, C7 75; cap 80. Likelihood capped at Roughly even for any judgment resting on these claims alone.
3	v2.0 PDF metadata: author field, dates, producer	Technical artefact in item 1	n/a	n/a	3 (C4)	70 Solid	25 Low	F11 55, F12 90 (any reader reproduces with pdfinfo), F13 80 (creation equals modification; producer string consistent). Supports "a person of that name produced the file." Supports no statement about employment.
4	v1.0 document (embedded in the April 2025 Treadstone 71 report)	Sber, via Treadstone 71 archive	A	1	3	84 Strong	25 Low	Chain of custody inside the firm; content corroborated by the April 2025 press descriptions (70 threats, three stages, OWASP/MITRE/NIST sources).
5	Lukatsky review, August 3 to 4, 2026 (99 pages; no C2; FSTEC heritage; maturity gap)	lukatsky.ru, independent consultant employed by Positive Technologies	B	1	3	78 Solid	25 Low	F3 independence 70 (vendor employment noted); corroborates items 1 and 12 on structure and lineage. Single reader for the 99-page count; Rule 7 condition stated in the text.
6	Habr review, August 10, 2026 (ITMO AI Security Lab)	habr.com, Mosin	C	2	2	62 Solid	25 Low	Student author, lab affiliation; links to the Kibrary files; corroborated on structure by items 1 and 5.
7	Sber press release of July 30, 2026, as carried by CNews and SecPost	Sber via trade press	C (Rule 1)	2	2	58 Weak	28 Low	Self-serving; release date and counts confirmed by item 1.
8	Sber press release of April 18, 2025, with Lebed quote	Sber via CNews, ComNews, Habr	C (Rule 1)	2	2	58 Weak	28 Low	Fixes the public release date of v1.0.
9	FSTEC BDU AI threats section, December 2025	bdu.fstec.ru (regulator, primary)	B	1	2	80 Strong	25 Low	Fetched; confirmed by Vedomosti and d-russia reporting.
10	Draft GOST R on secure AI software development, clause 5.7 source list	Habr summary of a regulation.gov.ru draft	C	2	3	60 Solid	25 Low	Secondhand summary; the draft text itself remains outside the evidence base. Judgment on Path A rests on it and is held at Likely, not higher.
11	ITMO faculty and lab pages on A. A. Vorobeva	edu.itmo.ru, securityhub.ru, ntv.ifmo.ru	B	1	2	76 Solid	25 Low	Establishes her ITMO role and research field. The identity link to item 3 is a separate inferred judgment, no Table 1 grade, printed at Roughly even chance.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

No.	Item	Source	Rel.	Cred.	Level	T	D	Notes
12	FSTEC 2021 methodology structure (violator categories, potential levels H1 to H4, motivation filter, scenario rule)	Anti-Malware.ru (Talanov, 2023) and SecurityLab (Lukatsky)	B	2	2	72 Solid	25 Low	Two independent practitioner descriptions agree; the methodology text itself was not opened.
13	MITRE ATLAS tactic count (16) and November 2025 Command and Control addition	Vendor explainer (akto.io) and item 5	C	2	2	60 Solid	25 Low	Two sources agree on 16 tactics; atlas.mitre.org itself did not load (companion record).
14	Lebed, Vedomosti, December 12, 2025 (prompt injection; 85 percent forecast)	Vedomosti, snippet only	C (Rule 1)	3	2	48 Weak	28 Low	Body did not render at first collection; superseded by item 27.
15	OFFZONE 2026 Sber research "based on the updated model"	ng.ru and itsec.ru conference reporting	C	2	2	58 Weak	28 Low	Two outlets; Sber-sourced content; corroborates adoption inside Sber.
16	Sber 2025 annual report, 900-plus agents handling 70 percent of incidents	ksonline.ru and TASS, snippets	C (Rule 1)	3	2	50 Weak	30 Low	Self-reported figure; two relays; the report itself not opened.
17	Kaspersky-Sber AI cooperation agreement, March 11, 2025	Kaspersky press release	C (Rule 1)	2	1	n/a	n/a	Level 1 anchor only.
18	TAdviser v1.0 and v2.0 lines; Gazinformservis comment	tadviser.ru, snippet	C	3	1	n/a	n/a	Level 1 anchor only.
19	v1.0 PDF and diagram from the Kibrary server: content and metadata (author, April 18, 2025, 07:48 and 07:55 UTC)	Sber, Kibrary server	A	1	3	86 Strong	25 Low	C4: F11 70 (hashes recorded), F12 90, F13 85. Three files, two applications, one author field: the metadata chain behind the Section 1 authorship judgment.
20	v2.0 diagram album (7 pages, July 22, 2026, iLovePDF, no author)	Sber, Kibrary server	A	1	3	82 Strong	25 Low	Column totals of the matrix match the rebuilt graph on all 51 methods (F12). Fixes the date the diagrams were final.

No.	Item	Source	Rel.	Cred.	Level	T	D	Notes
21	borismlsec post of July 27, 2026: release date July 24, 99 pages, 43 objects, structural reading	Telegram channel export; author Likely a Sber employee (9.3)	C (Rule 1 where it praises)	2	3	66 Solid	30 Low	Page and object counts corroborated by items 5 and 1. Praise capped; criticism ("спорное," no measures) graded against interest under Rule 2.
22	borismlsec v1-to-v2 mapping spreadsheet, 127 rows, dated July 27, 2026	Same	B (technical artefact; rows checkable)	1	3	80 Strong	25 Low	F12: Treadstone 71 reproduced 69 of 70 row sets at least partly from the two documents. Corroborates the crosswalk; F5 credit on item 1.
23	Lukatsky Telegram post, August 4, 2026, 05:01 (99 pages, four verdicts)	Telegram channel export	B	1	2	80 Strong	25 Low	Confirms item 5 and gives the second independent page count.
24	MITRE ATLAS tactics and updates pages, captured October 18, 2026 (16 tactics; v2026.09; AI Attack Adaptation; C2; Hermes case study)	atlas.mitre.org, October 18, 2026	A	1	3	90 Strong	25 Low	Primary source replaces item 13.
25	GOST R 56939-2024 text (clause 5.7) and catalogue record (developers incl. SberTech; order 1504-ст; in force December 20, 2024)	normativ docs capture and gostinfo.ru	A	1	2	88 Strong	25 Low	OCR of a browser printout; clause 5.7.3.1 wording checked against the page image.
26	ISP RAS deck, May 13, 2026 (Padaryan, Sosnina, Turdakov): draft AI secure-development standard structure and status	Trusted AI Technologies forum deck	B	2	2	72 Solid	25 Low	Authors' own description of their draft; "already piloted" is a self-serving claim, capped. Partly replaces item 10.
27	Vedomosti interview with Lebed, December 12,	vedomosti.ru, browser-translation capture	C (Rule 1)	2	3	60 Solid	30 Low	Replaces item 14. Quotations are from the browser's English rendering; the Russian original stays outside the evidence base. 70-threat model,

No.	Item	Source	Rel.	Cred.	Level	T	D	Notes
	2025, full text (browser translation capture)							agent-identification requirement, Kaspersky benchmark, "desire to please" line.
28	Picus Blue Report 2026 (338M simulations, H1 2026)	Picus Security, vendor report	C	2	2	62 Solid	28 Low	Vendor dataset, self-published, methodology stated; not AI-specific. Section 13 cross-check only.
29	"Горизонты искусственного интеллекта" foresight report, November 2025 (270+ scientists, 36 countries; Sber and AI Alliance authorship)	Published PDF, November 2025	C (Rule 1 for Sber's self-framing)	2	2	60 Solid	30 Low	Supplies the stated vision in Section 5.
30	NOOA security requirements analysis, August 6, 2026 (ML&Sec Feed, forwarded by borismlsec)	Telegram export	C	2	1	n/a	n/a	Level 1; community depth indicator only.

The T-Score that governs the report's forward judgments is not the T-Score of item 1. Item 1, the content of the model, is Strong at 80 and permits any band; the report's statements about what the document contains carry no cap. The forward judgments in the front matter (adoption, revision, regulator uptake) rest on items 5, 6, 7, 9, 10, 15, 16, and 21 to 27 together. Their mean T, with items 21 to 27 added, is 69, Adequate, which caps every forward judgment at Likely and forbids anything above 85 percent. The report prints no forward judgment above Likely. One judgment, the identity of the file author, rests on items 3, 11, and 19 and on an inference, and prints at Likely at the lower edge of the band; the five-point rule sends it to a second reader (Section 9.4).

D-Scores sit at or near the floor throughout. The one marker observed on item 1 (removed authorship and bibliography, Masking at low severity) scores F21 at 25 and leaves the mean of the five C8 factors at 15, so D prints at the floor of 25. So the scores require no deception warning and no decision hold.

F-Score. Treadstone 71 scored cluster C9 on the endgame question "what does Sber achieve with the model." Target-set convergence F26 65 (developers, regulators, vendors, and the conference circuit all receive the same artefact); capability-intent alignment F27 75 (Sber co-authors the GOST, hosts a TC 164 working group, and funds the Kibrary channel); timeline compression F28 70 (sixteen months between versions; the GOST draft opened on the release day); cross-domain coordination F29 70 (press, standards, and conference lines moved together in a five-week window); endgame path probability F30 60. F-Score 68, Likely. The report names the Most-Likely Achieved Endgame: the model becomes the de facto reference for AI threat modelling in the Russian financial sector and inside Sber's partner network, used alongside FSTEC's threat database

rather than inside it. In comparison, the Stated Endgame, as framed in Sber-aligned press at v1.0 ("basis of a national approach"), sits at 40, Unlikely by 2028 on the current GOST text. The Optimal Endgame, formal inclusion in the GOST source list or the FSTEC database, sits at 55, the bottom of Likely, and is the falsifier to watch.

AI assistance disclosure (F19). Treadstone 71 used a large language model for optical character recognition cleanup, Russian-to-English rendering of quotations, graph construction, figure rendering, and drafting; a human checked every count against the page images and every quotation against the source text, and the ledger treats model output as a derived artefact, never as a primary source.

Analyst bias disclosure. Treadstone 71 wrote the April 2025 assessment against which this report measures v2.0. The scorecard in Figure 10 grades Sber's response to Treadstone 71's own findings, and the firm has an interest in those findings having been correct. Scoring was conservative: a finding earned "answered" only where a v2.0 entry addresses it by name or by mechanism, and "partly" where the structure changed without an entry.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Wrap-up

Sber set out in 2025 to write Russia's first AI threat model and produced a catalogue. Sixteen months later it produced an instrument. The difference is not the count of threats, which fell, or the page count, which more than doubled. What differs is that v2.0 knows what it is for: a team inventories objects, picks threats, picks attackers, picks the methods those attackers have, and hands the result to whoever designs the controls. The grammar it uses to do that is the Russian regulator's. MITRE supplied the names it gives the attacker's steps. The thing it worries about most is its own agents, nine hundred of them, working inside the perimeter with tools, memory, and the capacity to take instructions from a stranger.

Treadstone 71's 2025 findings held up better than most predictions do. Sber closed the gaps that rows close and left open the gaps that required adding time: no damage model, no update clock, no account of an attacker who learns. The one place the model went backwards is the one place it touched perception, where Sber struck from the list the fourth property that let a defender call an output false without calling it tampered.

Readers outside Russia should spend their next hour on the attribution table, which is where Sber has written down, in a public document, which attacker it expects behind each of 37 outcomes. For the attacker it calls "special services" it expects theft, poisoning, and persistence, and it does not expect reconnaissance of agent internals, credential theft, or disruption. Those are 123 open paths in Sber's own graph. The next revision will show whether anyone at Sber reads its own matrix the way its adversaries will.

As for the file in Treadstone 71's hands, Sber exported it on August 28, 2026, five weeks after the Kibrary posting, one page shorter than the copy two Russian reviewers counted in late July and early August, with the entry for its authors still in the table of contents and its target gone. The name left in the file properties also sits on the v1.0 model and on the v1.0 diagram, exported seven minutes apart on the morning of the first release. On the public record that name belongs to a university lecturer in St. Petersburg who teaches attacks on AI systems. Sber removed an authors page and left the author field on three files. The live file on the Kibrary server will say which page count Sber now publishes; the answer to who wrote the model is already in the metadata Sber did not strip.

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Sources

Primary documents

1. Сбер Кибербезопасность. Модель угроз кибербезопасности AI для этапов сбора и подготовки данных, разработки модели и обучения, эксплуатации модели и интеграций с AI-решениями, v.2.0 (июль 2026). 98 pp. (99 at release). MD5 04315425b1ea493f72e46156f4794cb1. Official page: https://www.sberbank.ru/ru/person/kibrary/experts/model-ugroz-kiberbezopasnosti-ai_2
2. Сбер Кибербезопасность. Модель угроз для кибербезопасности AI, v.1.0 (март 2025). 36 pp. plus cover, as embedded in source 3. Official page: https://www.sberbank.ru/ru/person/kibrary/experts/model-ugroz-kiberbezopasnosti-ai
3. Treadstone 71. Strategic Exposure in the Russian Sberbank AI Threat Model. April 20, 2025. 63 pp.
4. Treadstone 71 LLC. The ASA Evidence Standard, Public Edition 1.2. October 4, 2026. https://www.treadstone71.com/asa-evidence-standard ; https://doi.org/10.5281/zenodo.23137881

4a. Treadstone 71 LLC. Advanced Analytic Dominance: Advanced Structured Analytic Techniques. https://www.treadstone71.com/training/advanced-analytic-dominance-sats

4b. Treadstone 71 LLC. Decision Support: Strategic Intelligence Tools (ATCRI, ACS, CWC, CWIA, HTIM, CARM). https://www.treadstone71.com/decision-support-strategic-intelligence-tools

4c. Treadstone 71 LLC. AI-Augmented Cyber Intelligence Tradecraft. https://www.treadstone71.com/ai-augmented-cyber-intelligence-tradecraft

Russian-language reporting and analysis

5. CNews. Сбербанк открыл доступ к обновлённой модели угроз кибербезопасности ИИ. July 30, 2026. https://safe.cnews.ru/news/line/2026-07-30_sberbank_otkryl_dostup_k
6. SecPost. Сбер опубликовал обновленную модель угроз для систем искусственного интеллекта. July 30, 2026. https://secpost.ru/sber-opublikoval-obnovlennuyu-model-ugroz-dlya-sistem-iskusstvennogo-intellekta
7. Лукацкий, А. Модель угроз ИИ от Сбера: и да, и нет. August 4, 2026. https://lukatsky.ru/threats/model-ugroz-ii-ot-sbera-i-da-i-net.html
8. Мосин, В. Источники сбора угроз ИБ для систем ИИ. Habr, August 10, 2026. https://habr.com/ru/articles/1062400/
9. CNews. Сбербанк разработал и опубликовал модель для кибербезопасности ИИ. April 18, 2025. https://safe.cnews.ru/news/line/2025-04-18_sberbank_razrabotal_i_opublikoval
10. ComNews. Сбер разработал и опубликовал модель для кибербезопасности ИИ. April 18, 2025. https://www.comnews.ru/digital-economy/content/238895/2025-04-18/2025-w16/1012/sber-razrabotal-i-opublikoval-model-dlya-kiberbezopasnosti-ii
11. Habr News. April 2025. https://habr.com/ru/news/902882/
12. TAdviser. Риски использования искусственного интеллекта (section on the Sber model). https://www.tadviser.ru/index.php/Статья:Риски_использования_искусственного_интеллекта
13. TAdviser. Искусственный интеллект в Сбербанке. https://www.tadviser.ru/index.php/Статья:Искусственный_интеллект_в_Сбербанке
14. Ведомости. Сергей Лебедь: «Будущее кибербезопасности – за мультиагентными системами защиты». Interview by Alexandra Volgina, December 12, 2025, 12:49. https://www.vedomosti.ru/technologies/trendsrub/characters/2025/12/12/1162710-sistemami-zaschiti
15. Независимая газета. OFFZONE 2026 report on Sber agent research. August 21, 2026. https://www.ng.ru/news/845462.html

16. ITSec.ru. В Москве прошла седьмая международная конференция по практической кибербезопасности OFFZONE 2026. https://www.itsec.ru/news/v-moskve-proshla-sedmaya-mezhdunarodnaya-konferenziya-po-prakticheskoy-kiberbezopasnosti-offzone-2026

17. Новости Воронежа. Ural Forum report on the Sber multi-agent SOC. February 19, 2026. https://novostivoronezha.ru/2026/02/19/343759

18. Anti-Malware.ru. Таланов, М. Возможные перемены в методике ФСТЭК России по моделированию угроз безопасности. March 14, 2023. https://www.anti-malware.ru/analytics/Technology_Analysis/FSTEC-methodology-changes

19. SecurityLab. Лукацкий, А. Магия моделирования угроз по ФСТЭК. https://www.securitylab.ru/blog/personal/Business_without_danger/350797.php

20. Anti-Malware.ru. News item of April 27, 2026. https://www.anti-malware.ru/news/2026-04-27-111332/49848

21. CNews. Positive Technologies научила PT Application Firewall защищать LLM. October 16, 2026. https://safe.cnews.ru/news/line/2026-10-16_positive_technologies_nauchila_pt_application

22. Habr (USSC). Безопасность ИИ-систем 2026. https://habr.com/ru/companies/ussc/articles/1088988

23. Habr. Draft GOST R on secure development of AI software, summary. July 30, 2026. https://habr.com/ru/articles/1064870/

24. Коммерсантъ. FSTEC draft personal-data order. July 28, 2026. https://www.kommersant.ru/doc/8845936

25. consult-cct.ru and asp-news.ru syndications of the April 2025 release. https://consult-cct.ru/sber-razrabotal-i-opublikoval-pervuyu-v-rossii-model-dlya-kiberbezopasnosti-iskusstvennogo-intellekta

26. 24rus.ru and sia.ru syndications of the July 2026 release. https://24rus.ru/news/finance/247268.html ; https://sia.ru/?action=show_news&id=16830726&section=484

Regulators and standards

27. ФСТЭК России. Банк данных угроз, раздел «Угрозы безопасности информации систем искусственного интеллекта». https://bdu.fstec.ru/threat/ai

28. ФСТЭК России. Банк данных угроз. https://bdu.fstec.ru/threat

29. Ведомости. ФСТЭК определила угрозы безопасности для систем ИИ. December 23, 2025. https://www.vedomosti.ru/technology/articles/2025/12/23/1165524-fstek-opredelila-ugrozi-bezopasnosti

30. ITProtect. Обновление БДУ. https://itprotect.ru/mediacenter/news/obnovlenie-bdu/

31. РИА Стандарты и качество. ТС 164 working groups. https://ria-stk.ru/stq/adetail.php?ID=194233

32. Храмцовская, Н. ПНСТ 1046-2026. April 2026. http://rusrim.blogspot.com/2026/04/1046-2026.html

Western frameworks and context

33. Akto. MITRE ATLAS and agentic AI (tactic count and 2025 to 2026 releases). https://www.akto.io/blog/mitre-atlas-agentic-ai

34. MITRE ATLAS. Tactics page and Updates page, data v2026.09, October 18, 2026. https://atlas.mitre.org/tactics ; https://atlas.mitre.org/resources/updates/2026-09

35. Сбер Кибербезопасность. Обобщённая схема объекта защиты и актуальных угроз КБ AI (v1.0 diagram), April 18, 2025. https://www.sberbank.ru/common/img/uploaded/kibrary/themap/obobshchennaya_skhema_obekta_zashchity_i_aktualnykh_ugroz_kb_ai.pdf

36. Сбер Кибербезопасность. Схемы и матрицы к модели угроз кибербезопасности AI (v.2.0), 7 pp., assembled July 22, 2026. https://www.sberbank.ru/common/img/uploaded/kibrary/reminders/skhemy_i_matricy_v_2.pdf

37. Борис_ь с ml (t.me/borismlsec). Модель Угроз AI Сбера 2.0, July 27, 2026, with mapping_sber_ai_threat_model_v1.0_to_v2.0.xlsx. https://t.me/borismlsec/267 ; https://t.me/borismlsec/268

38. Борис_ь с ml. Posts of June 3 (OWASP agentic report), July 8 (AI incident response), August 19 (OFFZONE talk announcement), September 28 (frontier-lab red-team telemetry), October 13, 2026 (ITMO workshop notice). https://t.me/borismlsec
39. Лукацкий, А. Telegram post of August 4, 2026, 05:01 (summary of the blog review). https://t.me/alukatsky
40. ML&Sec Feed. Требования безопасности и практики тестирования GenAI-систем на основе анализа NOOA, August 6, 2026 (forwarded by borismlsec).
41. Росстандарт / ФГБУ «Институт стандартизации». Catalogue record for ГОСТ Р 56939-2024. https://www.gostinfo.ru/catalog/Details/?id=8094263
42. ГОСТ Р 56939-2024. Защита информации. Разработка безопасного программного обеспечения. Общие требования. Introduced December 20, 2024.
43. Падарян В. А., Соснина Е. С., Турдаков Д. Ю. (ИСП РАН). Разработка безопасного ПО, реализующего технологии ИИ. Форум «Технологии доверенного искусственного интеллекта», May 13, 2026.
44. Picus Security. The Blue Report 2026. Picus Labs, 2026.
45. Сбер / Международный Альянс в сфере ИИ. Горизонты искусственного интеллекта: какими будут технологии ИИ через 10 лет. Научно-исследовательский проект, November 2025 (PDF produced February 24, 2026).
46. Arize. OpenClaw vs Hermes, Arize Observe 2026 panel (YouTube). Oxylabs. OpenClaw vs Hermes: Which AI Agent Wins in 2026? (YouTube). openclawdatabase.com, OpenClaw + Hermes multi-agent workflow, April 13, 2026. anthemcreation.com, Hermes vs OpenClaw: Learning or Orchestrating, April 2026.
47. OFFZONE 2026. Self-Evolving AI Agents: Architecture, Threats, and Secure Evolution, AI.ZONE track, August 20, 2026, 18:30. https://offzone.moscow/program/self-evolving-ai-agents-architecture-threats-and-secure-evolution/

Public record on the named file author

48. ИТМО. Воробьева Алиса Андреевна, faculty page. https://edu.itmo.ru/ru/lecturers_and_professors/130334
49. SecurityHub.ru. Person page. https://securityhub.ru/personlist/person-14/
50. Научно-технический вестник ИТМО. Author page. https://ntv.ifmo.ru/ru/person/9674/vorobeva_alisa_andreevna.htm
51. Google Scholar profile (fetch blocked by robots.txt; citation count seen in snippet only). https://scholar.google.ru/citations?hl=ru&user=dprc7QkAAAAJ

Appendix A. Crosswalk of the 70 v1.0 entries to v2.0

v1.0 ID	v1.0 entry	v2.0 descendant(s)	Fate
D01	Poisoned external data for training/fine-tuning	ME04, T16	Split
D02	Modified external data loaded after registration	ME05, T16	Split
D03	Reproduction of personal data from external sources in outputs	T11, ME49	Split
D04	Poisoned internal data for training/fine-tuning	ME06, T16	Split
D05	Uncontrolled loading of confidential data into training sets	ME30, T09	Split
D06	Uncontrolled use, modification, deletion of training data	T15, ME31	Split
Infr01	Unauthorized modification of source registry / datasets	ME31	Method
Infr02	Unauthorized modification of training data	ME31	Method
Infr03	Insecure data transfer between preparation stages	ME32	Method
Infr04	Theft of training data	T01	Threat
Infr05	Confidential data leak from training sets	T09	Threat
Infr06	Vulnerable third-party libraries and backdoored code	ME07	Method
Infr07	Open-source models with program backdoors in files	ME08	Method
Infr08	Open-source models with logical backdoors from training	ME09	Method
Infr09	Open-source models with backdoors in weights	ME08	Method
Infr10	Model substitution or modification	ME41, ME36, T34	Split
Infr11	Model theft	T03, ME41	Split
Infr12	Model availability violation	T22, ME18	Split
Infr13	Leaks from logging systems incl. request and call logs	T10, ME45	Split
Infr14	Inability to detect/respond due to missing interaction logging	T37	Threat
Infr15	Interception or substitution of requests/responses or RAG data	ME33	Method
Infr16	Unauthorized disabling of input/output filtering	ME34	Method
Infr17	System prompt theft	T08, ME22	Split
Infr18	Unauthorized modification of system prompt	ME22, T35	Split
Infr19	Unauthorized modification of data in internal sources incl. RAG	ME35, T18	Split
Infr20	Leaks from internal data sources	T12	Threat
Infr21	Unauthorized modification of test and validation datasets	ME31, ME02	Method
Infr22	Unauthorized connections to the model	ME11	Method
Infr23	Leak of AI-agent data or implementation details	T07, ME23	Split
Infr24	Unauthorized modification of AI agent	ME23	Method
Infr25	Leak of multi-agent architecture via dev tools or user interaction	T06	Threat
M01	Inability to investigate due to missing training-data information	T37	Threat
M02	Model highly vulnerable to adversarial and prompt attacks	ME38	Method
M03	Undesirable behavior, malicious generations, hallucinations	T33	Threat
M04	White-box attack selection using open-source model knowledge	T02, ME01	Split
M05	Absence of model inference information	T37	Threat

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

v1.0 ID	v1.0 entry	v2.0 descendant(s)	Fate
M06	Bypass of model-level input/output processing	ME39	Method
M07	Model DoS due to no unified request control	T22, ME18	Split
M08	Integration limit exhaustion (DoW)	T23, ME19	Split
M09	Bypass of built-in model protections incl. adversarial/prompt	ME38	Method
M10	Model information leak	T02	Threat
M11	Confidential leak from fine-tuned model or LoRA	T11	Threat
M12	Model exfiltration, inversion, reverse engineering	ME48	Method
M13	Data exfiltration	ME49, T09	Split
App01	Design errors, insecure component integrations	ME12	Method
App02	Bypass of app-level input/output processing	ME40	Method
App03	Malware download from external sources (Internet)	ME10	Method
App04	Poisoned data download from external sources (Internet)	ME13	Method
App05	Indirect prompt injection into internal sources incl. RAG	ME35, ME24, T18	Split
App06	Leaks from internal sources incl. RAG	T12	Threat
App07	Execution of malicious instructions created by the model	T33, ME43, ME44	Split
App08	Direct prompt injections due to missing input control	ME14	Method
App09	Integration availability violation (DoS/DoW)	T23	Threat
App10	Task logic violation due to missing input control	T35	Threat
App11	System prompt leak via incorrect output handling	T08, ME40	Split
App12	Toxic or malicious generation via incorrect output handling	T31	Threat
App13	Environment information disclosure	T04	Threat
App14	Automatic propagation of malicious instruction to other apps	T30	Threat
Ag01	Design errors in agents and multi-agent systems	ME12	Method
Ag02	Malicious generations in agent response	T31, T33	Threat
Ag03	Sending data from agent function environment to external resources	ME51	Method
Ag04	Deletion or modification of files in agent environment	T21	Threat
Ag05	Placing malware files in agent environment from external sources	ME10, ME43	Method
Ag06	DoS/DoW of agent execution environment	T26	Threat
Ag07	Multi-agent architecture leak via user input interfaces	T06	Threat
Ag08	Passing false information to another agent	ME20, T29	Split
Ag09	Goal violation of another agent in cooperative MAS	T29	Threat
Ag10	Prompt attack propagation across agents in MAS	T30	Threat
Ag11	Workflow disruption of app implementing the agent	ME20, ME21, T24	Split
Ag12	Leak of agent goal, functions, memory, planner instructions	T07, T13, ME23	Split

Appendix B. The 23 v2.0 entries with no v1.0 antecedent

v2.0 ID	Entry	Kind	Group or tactic	Links
T05	System prompt extraction for targeted attacks	Threat	Unauthorized information gathering	6 methods
T14	Compromise of model credentials and access keys	Threat	Confidential information leak	7 methods
T17	Malicious code execution in AI IT infrastructure	Threat	Unauthorized impact on resources and runtime	5 methods
T19	Data deletion or modification from executing malicious instructions	Threat	Unauthorized impact on resources and runtime	6 methods
T20	Runtime isolation breach with control of base infrastructure	Threat	Unauthorized impact on resources and runtime	4 methods
T25	Operational interference and sabotage by overloading the human link	Threat	Denial of service	6 methods
T28	Availability disruption of agent resource system	Threat	Denial of service	8 methods
T32	Use of AI app or agent to prepare cybercrimes	Threat	Dissemination of illegal information	7 methods
T36	Use of AI solution outside business scenario	Threat	Functional (operability) violation	6 methods
T27	Availability failure of the app hosting the agent	Threat	Denial of service	7 methods
ME03	Public model repository compromise (links, metadata, configs)	Method	Resource Development	2 threats
ME16	Inter-agent protocol flaws (e.g. MCP)	Method	Execution	5 threats
ME17	Orchestrator (planner) vulnerabilities	Method	Execution	5 threats
ME25	Autonomous goal degradation, emergent behavior	Method	Persistence	2 threats
ME26	Excessive permissions and allowed operations	Method	Privilege Escalation	5 threats
ME27	Isolation and data access control flaws	Method	Privilege Escalation	2 threats
ME28	Direct interaction between agents of different confidentiality	Method	Privilege Escalation	3 threats
ME29	Function-call manipulation	Method	Privilege Escalation	6 threats
ME37	Isolation/auth flaws in interactive dev environments	Method	Defense Evasion	5 threats
ME42	Credential and key theft	Method	Credential Access	2 threats
ME46	Packet size/timing analysis of encrypted streaming (inference cache)	Method	Exfiltration	3 threats
ME47	Residual data from shared inference infrastructure	Method	Exfiltration	6 threats
ME50	Data recovery from embeddings	Method	Exfiltration	2 threats

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

Appendix C. Full list of v2.0 threats and methods with their links

ID	Entry	Group	Property	Stage	Model types	Violators	Methods
T01	Theft of training data	Unauthorized information gathering	C	Data/Dev	PredAI, GenAI	9	ME31, ME32, ME37
T02	Leak of model information and parameters	Unauthorized information gathering	C	Dev/Ops	PredAI, GenAI	9	ME41, ME02, ME36, ME37, ME47, ME48, ME01
T03	Model theft or copying	Unauthorized information gathering	C	Dev/Ops	PredAI, GenAI	9	ME41, ME36, ME37, ME47, ME48
T04	Leak of AI application / agent runtime information	Unauthorized information gathering	C	Ops	GenAI	7	ME23, ME13, ME14, ME15, ME51
T05	System prompt extraction for targeted attacks	Unauthorized information gathering	C	Ops	GenAI	6	ME22, ME13, ME14, ME15, ME51, ME40
T06	Leak of multi-agent system architecture	Unauthorized information gathering	C	Ops	GenAI	8	ME22, ME23, ME13, ME14, ME15, ME51
T07	Leak of AI agent configuration	Unauthorized information gathering	C	Ops	GenAI	8	ME22, ME23, ME13, ME14, ME15, ME51
T08	System prompt leak causing loss of uniqueness	Confidential information leak	C	Ops	GenAI	8	ME33, ME22, ME23, ME46, ME13, ME14, ME15, ME51, ME40
T09	Confidential data leak from datasets	Confidential information leak	C	Data/Dev/Ops	PredAI, GenAI	9	ME30, ME31, ME32, ME37, ME49
T10	Confidential data leak from logging systems	Confidential information leak	C	Ops	PredAI, GenAI	4	ME45, ME47, ME51
T11	Confidential data leak from fine-tuned model or LLM adapter	Confidential information leak	C	Dev/Ops	GenAI	10	ME30, ME41, ME33, ME11, ME36, ME49, ME46, ME42, ME47, ME50, ME13, ME14, ME27, ME15, ME51, ME28
T12	Confidential data leak from internal sources (DB, RAG)	Confidential information leak	C	Ops	GenAI	11	ME33, ME35, ME46, ME47, ME50, ME13, ME14, ME26, ME27, ME15, ME51, ME28
T13	Confidential data leak from memory mechanisms	Confidential information leak	C	Ops	GenAI	7	ME33, ME23, ME47, ME13, ME14, ME15, ME51

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

ID	Entry	Group	Property	Stage	Model types	Violators	Methods
T14	Compromise of model credentials and access keys	Confidential information leak	C	Ops	PredAI, GenAI	8	ME22, ME23, ME42, ME13, ME14, ME15, ME51
T15	Deletion or destruction of datasets	Unauthorized impact on resources and runtime	IA	Data/Dev	PredAI, GenAI	5	ME31
T16	Poisoned data used for training or evaluation	Unauthorized impact on resources and runtime	I	Data/Dev/Ops	PredAI, GenAI	9	ME04, ME05, ME06, ME31, ME32, ME37
T17	Malicious code execution in AI IT infrastructure	Unauthorized impact on resources and runtime	IA	Data/Dev/Ops	PredAI, GenAI	10	ME07, ME08, ME10, ME43, ME44
T18	Malicious content and prompt attacks planted in internal sources, tool descriptions, agent memory	Unauthorized impact on resources and runtime	I	Ops	GenAI	11	ME35, ME23, ME13, ME24, ME29
T19	Data deletion or modification from executing malicious instructions	Unauthorized impact on resources and runtime	IA	Ops	GenAI	6	ME33, ME13, ME14, ME26, ME15, ME29
T20	Runtime isolation breach with control of base infrastructure	Unauthorized impact on resources and runtime	IA	Ops	GenAI	9	ME12, ME26, ME43, ME44
T21	File deletion or modification in agent function runtime	Unauthorized impact on resources and runtime	IA	Ops	GenAI	6	ME12, ME13, ME14, ME26, ME15, ME43, ME44, ME29
T22	Model availability disruption	Denial of service	A	Ops	PredAI, GenAI	5	ME18, ME19
T23	DoS or DoW of the AI solution	Denial of service	A	Ops	PredAI, GenAI	6	ME11, ME19, ME13, ME14, ME21, ME15
T24	Integration failure of apps, agents	Denial of service	A	Ops	PredAI, GenAI	8	ME33, ME12, ME21, ME16

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

ID	Entry	Group	Property	Stage	Model types	Violators	Methods
	or components						
T25	Operational interference and sabotage by overloading the human link	Denial of service	A	Ops	PredAI, GenAI	5	ME40, ME13, ME14, ME15, ME25, ME17
T26	Agent runtime availability disruption	Denial of service	A	Ops	GenAI	7	ME19, ME12, ME13, ME14, ME15, ME44, ME17
T27	Availability failure of the app hosting the agent	Denial of service	A	Ops	GenAI	7	ME19, ME12, ME13, ME14, ME15, ME16, ME17
T28	Availability disruption of agent resource system	Denial of service	A	Ops	GenAI	7	ME19, ME12, ME13, ME14, ME26, ME15, ME17, ME29
T29	Goal corruption of another agent in cooperative MAS	Multi-agent interaction violation	I	Ops	GenAI	6	ME33, ME13, ME14, ME20, ME24, ME15, ME16
T30	Cascading prompt attack propagation in MAS	Multi-agent interaction violation	I	Ops	GenAI	6	ME33, ME20, ME24, ME16, ME28, ME29
T31	Generation of illegal, toxic or unethical content	Dissemination of illegal information	I	Ops	GenAI	5	ME38, ME39, ME40, ME13, ME14, ME15
T32	Use of AI app or agent to prepare cybercrimes	Dissemination of illegal information	I	Ops	GenAI	7	ME09, ME38, ME39, ME40, ME13, ME14, ME15
T33	Incorrect or undeclared model behavior, hallucinations	Functional (operability) violation	I	Ops	PredAI, GenAI	11	ME02, ME03, ME04, ME05, ME06, ME09, ME31, ME34, ME36, ME38, ME39, ME41
T34	Result bias, accuracy loss, backdoors, output distortion	Functional (operability) violation	I	Ops	PredAI, GenAI	9	ME02, ME03, ME04, ME05, ME06, ME09, ME31, ME36, ME38, ME39, ME41
T35	Task logic and workflow violation	Functional (operability) violation	I	Ops	PredAI, GenAI	9	ME33, ME34, ME22, ME38, ME39, ME40, ME13, ME14, ME20, ME15, ME16, ME25, ME17, ME29

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

ID	Entry	Group	Property	Stage	Model types	Violators	Methods
T36	Use of AI solution outside business scenario	Functional (operability) violation	I	Ops	GenAI	2	ME38, ME39, ME40, ME13, ME14, ME15
T37	Failure to detect, respond to and investigate events	Monitoring and investigation violation	A	Data/Dev/Ops	PredAI, GenAI	9	ME34, ME11, ME45

ID	Method	ATLAS tactic	Violators	Threats
ME01	Model data via public documentation and files	Reconnaissance	3	T02
ME02	Tampering with training/evaluation process parameters	Resource Development	7	T02, T33, T34
ME03	Public model repository compromise (links, metadata, configs)	Resource Development	4	T33, T34
ME04	Poisoning external data sources	Initial Access	8	T16, T33, T34
ME05	Immutability gaps on deferred external loads	Initial Access	8	T16, T33, T34
ME06	Poisoning internal data sources	Initial Access	3	T16, T33, T34
ME07	Third-party library vulnerabilities	Initial Access	9	T17
ME08	Backdoors in external model files or weights	Initial Access	9	T17
ME09	Logical backdoors / malicious properties in external models	Initial Access	9	T33, T34, T32
ME10	Malware download from external sources	Initial Access	6	T17
ME11	Unauthorized model connections	AI Model Access	5	T37, T11, T23
ME12	Design errors and insecure integrations	Execution	5	T20, T24, T21, T26, T27, T28
ME13	Indirect prompt attacks via external data loads	Execution	7	T11, T18, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28
ME14	Direct prompt or adversarial attacks	Execution	7	T11, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28
ME15	Indirect prompt attacks via poisoned internal sources	Execution	7	T11, T12, T19, T23, T35, T08, T31, T04, T32, T05, T36, T13, T25, T14, T21, T26, T06, T29, T27, T07, T28
ME16	Inter-agent protocol flaws (e.g. MCP)	Execution	4	T35, T24, T29, T30, T27
ME17	Orchestrator (planner) vulnerabilities	Execution	4	T35, T25, T26, T27, T28
ME18	DDoS on model deployment infrastructure	Execution	5	T22
ME19	Token / compute consumption escalation	Execution	7	T22, T23, T26, T27, T28

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

ID	Method	ATLAS tactic	Violators	Threats
ME20	Semantic distortion of inter-component messages	Execution	6	T35, T29, T30
ME21	Format or syntax distortion of structured data	Execution	6	T23, T24
ME22	Unauthorized access to system prompts and configs	Persistence	4	T35, T08, T05, T14, T06, T07
ME23	Unauthorized agent modification or implementation disclosure	Persistence	9	T18, T08, T04, T13, T14, T06, T07
ME24	Malicious content via user-content saving (memory)	Persistence	6	T18, T29, T30
ME25	Autonomous goal degradation, emergent behavior	Persistence	2	T35, T25
ME26	Excessive permissions and allowed operations	Privilege Escalation	6	T12, T19, T20, T21, T28
ME27	Isolation and data access control flaws	Privilege Escalation	6	T11, T12
ME28	Direct interaction between agents of different confidentiality	Privilege Escalation	6	T11, T12, T30
ME29	Function-call manipulation	Privilege Escalation	7	T18, T19, T35, T21, T30, T28
ME30	Confidential data loaded into training sets	Defense Evasion	3	T09, T11
ME31	Unauthorized access/modification/deletion of datasets	Defense Evasion	6	T15, T01, T09, T16, T33, T34
ME32	Data interception between preparation stages	Defense Evasion	8	T01, T09, T16
ME33	Request/response interception between components	Defense Evasion	5	T11, T12, T19, T35, T08, T30, T13, T24, T29
ME34	Disabling or modifying protection mechanisms	Defense Evasion	3	T37, T33, T35
ME35	Unauthorized changes to internal sources (DB, RAG)	Defense Evasion	10	T18, T12
ME36	Model interception between preparation stages	Defense Evasion	5	T33, T02, T11, T03, T34
ME37	Isolation/auth flaws in interactive dev environments	Defense Evasion	7	T01, T09, T16, T02, T03
ME38	Model weaknesses to adversarial and prompt attacks	Defense Evasion	9	T33, T34, T35, T31, T32, T36
ME39	Bypass of model-level I/O processing	Defense Evasion	9	T33, T34, T35, T31, T32, T36
ME40	Exploiting user/external input handling flaws	Defense Evasion	6	T35, T08, T31, T32, T05, T36, T25
ME41	Unauthorized access/deletion/substitution of trained model	Defense Evasion	7	T33, T02, T11, T03, T34
ME42	Credential and key theft	Credential Access	10	T11, T14

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM

ID	Method	ATLAS tactic	Violators	Threats
ME43	Agent used to fetch and run malicious code from internal resources	Lateral Movement	5	T17, T20, T21
ME44	Arbitrary code execution via agent runtime or functions	Lateral Movement	6	T17, T20, T21, T26
ME45	Access-control flaws for stealing log data	Collection	8	T10, T37
ME46	Packet size/timing analysis of encrypted streaming (inference cache)	Exfiltration	3	T11, T12, T08
ME47	Residual data from shared inference infrastructure	Exfiltration	2	T10, T02, T11, T03, T12, T13
ME48	Model exfiltration/inversion via query analysis	Exfiltration	5	T02, T03
ME49	Training-data recovery via membership inference	Exfiltration	5	T09, T11
ME50	Data recovery from embeddings	Exfiltration	5	T11, T12
ME51	Agent sends internal data to external resources	Exfiltration	7	T10, T11, T12, T08, T04, T05, T13, T14, T06, T07

Copyright 2026 Treadstone 71 LLC WWW.TREADSTONE71.COM