日経ビジネス 日経 XTECH

AI リーダーズ100

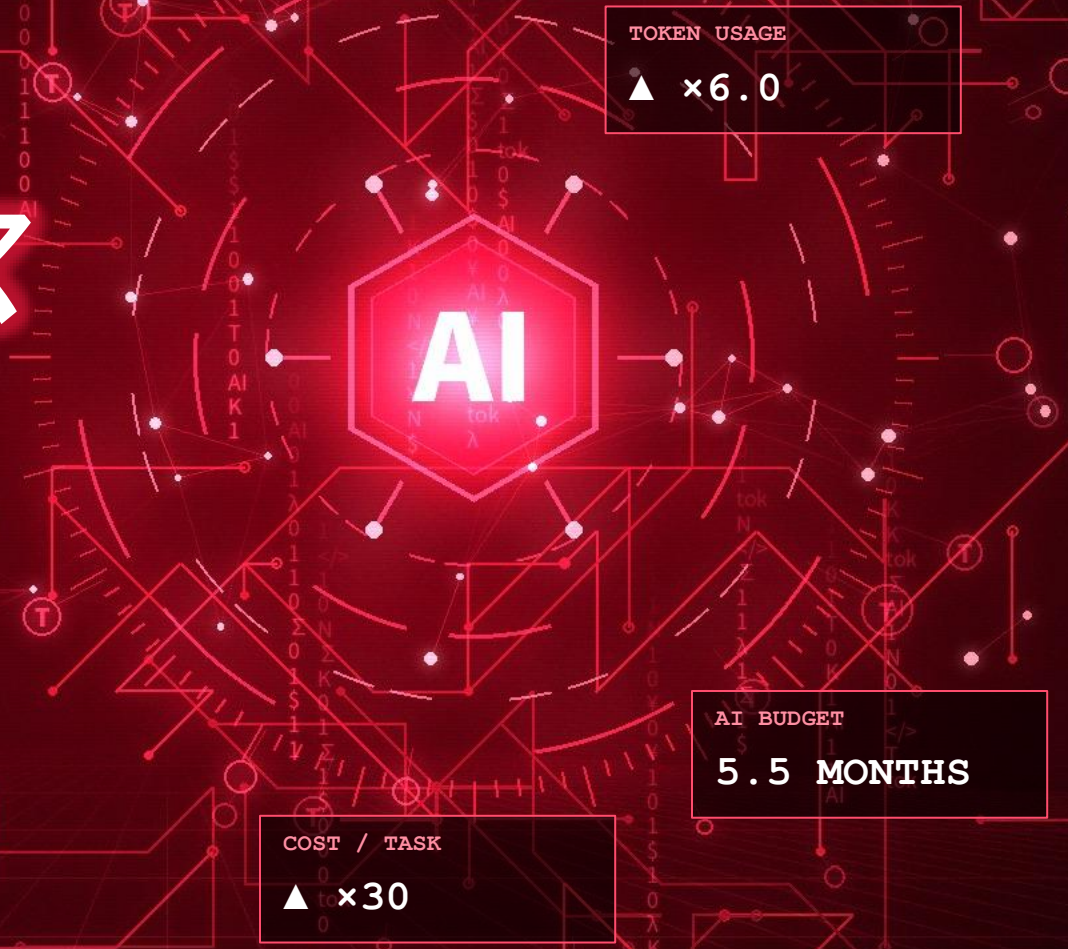
◆ NIKKEI BP EMERGENCY SPECIAL 2026

緊急始動！日経BP特番

AIトークン・クライシス

～「トークン爆発」に、企業はどう備えるべきか～

- 年間AI予算が半年弱で枯渇――日本企業を襲う「トークン爆発」
- 米国でも予算を4カ月で消化、AIコストは“第2の人件費”へ
- 経営 × システム × ソリューションで挑む「トークンマネジメント」



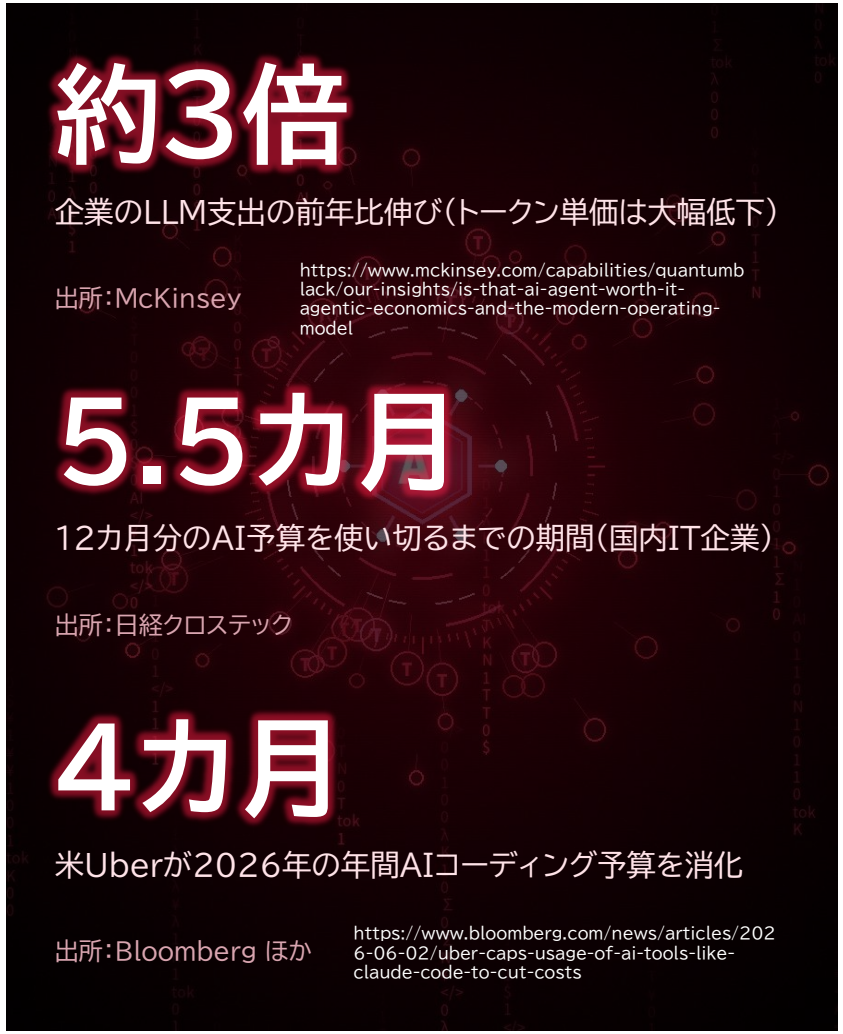
Nikkei BP Special Project

日経ビジネス 日経XTECH

AI活用の“次の壁”は、トークンコストだ

生成AIの業務活用が本格化し、AIエージェントが自律的に仕事をこなす時代が到来しました。その一方で企業を襲っているのが「トークン爆発」です。1トークン当たりの単価は下がり続けているにもかかわらず、利用量の急増で請求額は膨らみ、年間予算を半年足らずで使い切る企業も現れています。AIのコストは、企業にとって見過ごせない経営課題になりつつあります。

本特番では、国内外で起きている「トークン爆発」の実態を整理するとともに、経営者が持つべきAI投資の考え方、システム部門が取るべき技術的対策、そして可視化・上限管理・モデルルーティング、ローカルAI／エッジAIなどの最新ソリューションを紹介。AI活用の成果を落とさずにコストを最適化する「トークンマネジメント」の実践策を、編集部と協賛企業が一体となって発信します。



日経ビジネス電子版
経営層向け緊急特集

日経クロステック
IT・システム部門向け緊急特集

課題：単価は下がるのに、請求は膨らむ

日経ビジネス 日経XTECH

トークン単価

↓

1/280

GPT-4相当の性能の出力単価
（100万トークン当たり）

https://fortune.com/2026/09/23/mckinsey-cheaper-ai-models-bigger-ai-bills-cfo/

企業のLLM総支出

↑

約3倍

2025年末の前年比。
調査企業の93%が
トークン予算を超過

https://www.mckinsey.com/capabilities/quantumblack/our-insights/is-that-ai-agent-worth-it-agentic-economics-and-the-modern-operating-model

使う量の伸びが、単価の下落を上回る

AIが「人が画面で使うチャット」から「自律的に動くエージェント」へと移ったことで、1タスク当たりのトークン消費が桁違いに増加。完全従量課金のため、設定ミスや過剰スペックのモデル運用が“ビルショック”に直結します。

なぜトークンは爆発するのか

チャットからAIエージェントへ

エージェント型のタスクは、通常のチャットに比べ5～30倍のトークンを消費。人が逐一指示しなくてもAIが繰り返し処理を進める。

賢くなるAI、膨らむコンテキスト

長い推論、会話履歴や大量の資料の参照で、1回の処理量が増加。同じチャットで話題を変えると無関係な履歴まで処理される。

「定額だから使い放題」の誤解

定額プランにも利用枠があり、枯渇すると高機能モデルが停止。従量課金では上限がなければ請求は青天井に。

使う人・回数・任せる範囲が同時に拡大

モデル性能の向上と社員のAI習熟が掛け算になり、利用が一気に跳ね上がる。費用の見込みが極めて難しい。

出所：Stanford HAI（AI Index）、McKinsey「State of AI 2026」（Fortune報道）、Gartner（2026年7月）、日経クロステック（2026年9月30日）

日本企業で相次ぐ「想定外」――トークン爆発の実態

日経ビジネス 日経XTECH

企業	何が起きたか	どこまで増えたか
NTTデータグループ	ある社内プロジェクトでOpenAI APIの使用量が急増	**2026年4月の使用量が同年2月の約6倍に**
ファインディ	12カ月分として確保したAI予算を約5カ月半で使い切る	**追加予算後も2026年9月時点で当初想定の約3倍のペース**
クレディセゾン	全社員で共有するChatGPTのクレジットを使い切り、画像生成や推論モデルが一時利用不可に	**2026年6月15日に会社全体のクレジットが上限到達**
マネーフォワード	CursorやClaude CodeなどAI開発ツールの利用拡大でコストが増加	**2025年12月～2026年4月で約2.5倍（月平均約25%増）**
NEC	2026年6月にClaudeを全社導入した後、トークンコストが大幅に増加	**大幅増（数値は非公開）**

「上限に達したら止まるということを、ほとんど誰も意識していなかった」（クレディセゾン）

企業が陥る4つの落とし穴

定額制なら使い放題だと思う
定額料金でもトークンの利用枠は存在する

何でも高性能モデルを使う
簡単な質問に推論モデルを使えばコストがかさむ（AI版「カラーコピー問題」）

同一チャットで別の話題を尋ねる
無関係な履歴まで処理し、トークン効率が悪化

多く使う社員を止めればよいと考える
1人の利用が30～40人の業務を支えている例も

出所：日経クロステック「AIの年間予算が半年弱で枯渇、企業を襲う『トークン爆発』」（2026年9月30日）※クレディセゾンは支払額が予算を超過したわけではない

米国：AIトークンは「第2の人件費」になった

日経ビジネス 日経XTECH

Uber	2026年の年間AIコーディング予算を約4カ月で消化。従業員1人・1ツール当たり月1,500ドルの利用上限を設定（2026年6月報道）
Mercor	経営陣が「一部の組織ではトークンコストが社員の報酬に迫る、あるいは上回る」と公言。AIコストが人件費に並ぶ規模に
Meta	AIトークンコストが2026年に数十億ドル規模に達する見通しと社内メモで警告（報道）
GitHub	Copilotがトークンベースの課金に移行（2026年6月）。エージェント利用が多い開発者の請求額増加が予想される
Goldman Sachs	集中型のモデルゲートウエーでリクエストをタスク種別に分類し、最もコスト効率の良いモデルへ自動で振り分け

93%
トークン予算を超過した企業
McKinsey調査（報道）

40%
AIエージェントを本番で拡大する大企業（前年27%）
McKinsey調査（報道）

98%
AI支出を管理対象にするFinOps実務者
FinOps Foundation 2026

55%
AI向けIaaS支出に占める推論の比率。2026年に初めて学習を上回る
Gartner

使い放題 ▶ **上限・配分・使い分け**

出所：Bloomberg、Unite.AI、Fortune（McKinsey「State of AI 2026」）、FinOps Foundation「State of FinOps 2026」、Gartner（2026年8月）

経営の視点：AIトークンを「投資」として管理する

新たなコスト科目として予算化する

組織が大きくなるほど、AIコストは本部予算でも無視できない割合に。米国ではトークン費が社員の報酬に迫る企業も現れ、人件費と並べて管理する発想が必要に。

上限と責任者を決める

部門ごとに予算を付与して本部長が管理し、社員には業務に応じた月額上限（ティア制）を設定。超過時は理由とモデルを示して承認を得る仕組みで、青天井を防ぐ。

一律に止めず、配分する

高額利用が無駄とは限らない。1人のAIエージェントを30～40人が使う例もある。エンジニアなど事業に直結する利用は別枠で確保し、必要な所に投資を集中させる。

価値ある利用を見極める

白黒コピーで足りる仕事に高価なカラーコピーを使わない。どの利用に費用をかける価値があり、どの処理を安価な方法で代替できるかを経営として判断する。

先進企業の声

“組織の人数が増えるとAIコストは本部予算の中でも無視できない割合になる。本部長が適切に予算管理できるよう、可視化が必要だった”

— マネーフォワード 渡辺慎也氏

“追加し続けると、みんなが欲しい分だけ買い続けなければいけなくなる”

— クレディセゾン 川口裕大氏

“使わせない仕組みを作るのではなく、必要な時には追加投資できるようにしつつ、振り返りと対話ができる状態を作る”

— Sansan 牧野哲也氏

経営に求められるのは「止める」ことではなく、「見える化・配分・使い分け」で成果とコストを両立させること

出所：日経クロステック（2026年9月30日、10月1日）、Unite.AI（2026年9月25日）

システムの視点：3つのレバーでトークンを最適化する

利用回数を減らす

- 利用上限（Spend Limit）とアラート
- 高頻度処理をプログラムで固定化
- セマンティックキャッシュで類似質問を再利用

消費量を減らす

- プロンプト・コンテキストの圧縮
- MCPツール／検索結果の読み込みを絞る
- 推論量（Reasoning Effort）の自動調整
- プロンプトキャッシュで入力費を大幅削減

単価を下げる

- LLMルーターで難易度別にモデルを自動選択
- オープンウエート・ローカル／エッジAIの活用
- 自社用途向けファインチューニング
- 即時性不要の処理はバッチAPI（割引枠）へ

CASE：NEC

ピーク比

9割削減

を2026年12月に目指す

- Claude全社導入でトークンが激増（2026年6月）
- 技術・制度の多面的施策で9月に約5割削減の見通し
- 最大の効果はモデル切り替え：適切な用途でオープンウエートモデルを使い約2割減
- 10月からLLMルーターで全社的に自動選択へ

基盤：AI Gateway（中継・監視レイヤー）

社内システムとLLM APIの間に挟み、部署・プロジェクト別の消費をリアルタイムで可視化。権限管理、上限・自動停止、キャッシュ、ルーティングを中央で制御する。

出所：日経クロステック「AIのトークンコスト激増から9割削減へ、コスパを高める5カ条」（2026年10月1日）、NEC資料

先進企業に学ぶ、トークンコスト管理の5カ条

01

可視化・監視

部門・社員ごとのコストをダッシュボードで全社共有

02

利用上限の設定

部門予算と社員別上限。超過時は本部長承認で追加

03

モデルの使い分け

プロンプト内容から最適なLLMを自動選択

04

入力情報の削減

プロンプト整形・圧縮、類似質問はキャッシュ回答

05

意識醸成と共有

全社会議やNotionで節約ノウハウと好事例を共有

導入ロードマップ例

PHASE 1　現状の可視化

SaaS型ツールを導入し、どの部署・業務でどれだけトークンが発生しているかを1～2週間モニタリング

PHASE 2　ガバナンス設定

部署・職種ごとの月間利用上限（ティア制）とアラートを設け、予期せぬ請求事故を防止

PHASE 3　自動化・最適化

AI Gatewayを導入し、プロンプトキャッシュと難易度別の自動ルーティングを実装

出所：日経クロステック「AIのトークンコスト激増から9割削減へ、コスパを高める5カ条」（2026年10月1日）をもとに作成

トークン対策ソリューションの全体像

日経ビジネス 日経 XTECH

可視化・コスト管理

全社・部門別の利用状況を可視化
マネーフォワード クラウドAIトークン管理　など

トレースとコスト分析
Langfuse　など

プロキシ型でコード改修を抑えて可視化
Helicone　など

AI Gateway・ルーティング

多くのLLMにアクセスを自動振り分け
OrcaRouter(FlashLabs)／LiteLLM(OSS)　など

監視とルーティング機能を備える
Portkey　など

エッジでのキャッシュとアクセス制御
Cloudflare AI Gateway　など

ローカルAI・エッジAI

手元のワークステーションで大規模モデルを推論
NVIDIA DGX Spark／RTX搭載PC　など

NPU搭載端末で小型モデルをオンデバイス実行
AI PC(Copilot+ PC)　など

軽量な国産LLMをオンプレミスで運用
tsuzumi 2(NTT)／cotomi(NEC)　など

ローカルLLMの実行・推論サーバー
Ollama／vLLM（OSS）　など

クラウド・モデル提供者

モデル管理・ガードレール・アクセス制御
Amazon Bedrock(AWS)　など

拠点・部署別のトークン制限と費用按分
Azure API Management　など

プロンプトキャッシュ、割引バッチ処理
各社LLMのキャッシュ／バッチAPI　など

構築支援・コンサル

全社AIガバナンス環境の設計

クラウド上のLLM基盤とコスト最適化(FinOps)設計

マルチモデル設計、業務別モデル選定

PKSHA Technology／アイレット／エクサウィザーズ／クラスメソッド／電通総研／野村総合研究所　など

▶ **本特番の協賛候補領域**　可視化・ゲートウエー・ローカル／エッジAI・クラウド・SI各社のソリューションを、企業の経営層とIT部門に直接届けます

出所：各社公表資料、日経クロステック などをもとに編集部作成。掲載は例示であり網羅的なものではありません

ローカルAI／エッジAIで「自前の推論」へ

日経ビジネス 日経XTECH

なぜ今、ローカル／エッジなのか

コストを固定費化

トークンの従量課金を、ハードウエア投資と電気代に置き換え。利用量が多い定型業務ほど有利に

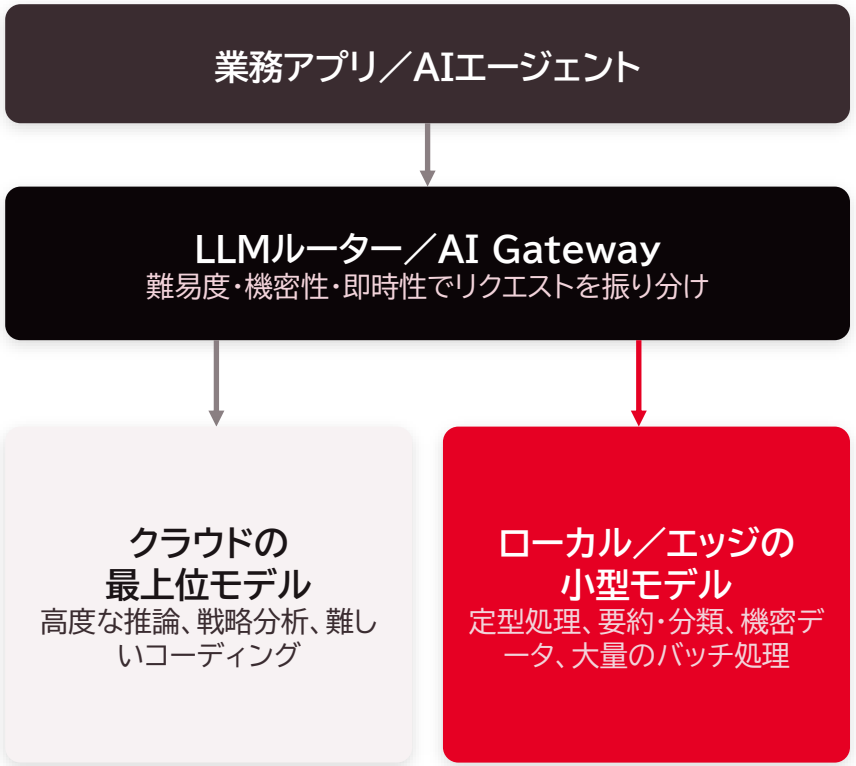
データを外に出さない

機密情報や個人情報を社内・端末内で処理。金融・医療・行政などで需要が拡大

低遅延・オフライン

端末上で即座に応答し、通信環境に左右されない。工場や店舗など現場の業務に

現実解は「ハイブリッド」での使い分け



3倍

2027年までに、タスク特化型の小型モデルの利用が汎用LLMの3倍に

Gartner予測（2025年4月）

https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20250410-ai-model

約2割減

NECが適切な用途をオープンウエートモデルに切り替えた効果（6月比）

日経クロステック

約1/18

オンプレミス推論の100万トークン当たりコスト（クラウドAPI比の試算）

DreamFactory調査（AIO総研の記事より）

導入時の注意点：初期投資と運用スキルが必要／最先端のクラウドモデルとは性能差がある／利用量が少なければクラウドの従量課金が割安。全部をローカルに移すのではなく、用途で使い分けることが鍵

出所：Gartner、日経クロステック（2026年10月1日）、AIO総研「ローカルAI元年」、NTT西日本／エフサステクノロジーズ発表、各社公表資料

AIトークン緊急特番：5大テーマとフォーカス領域

日経ビジネス 日経XTECH

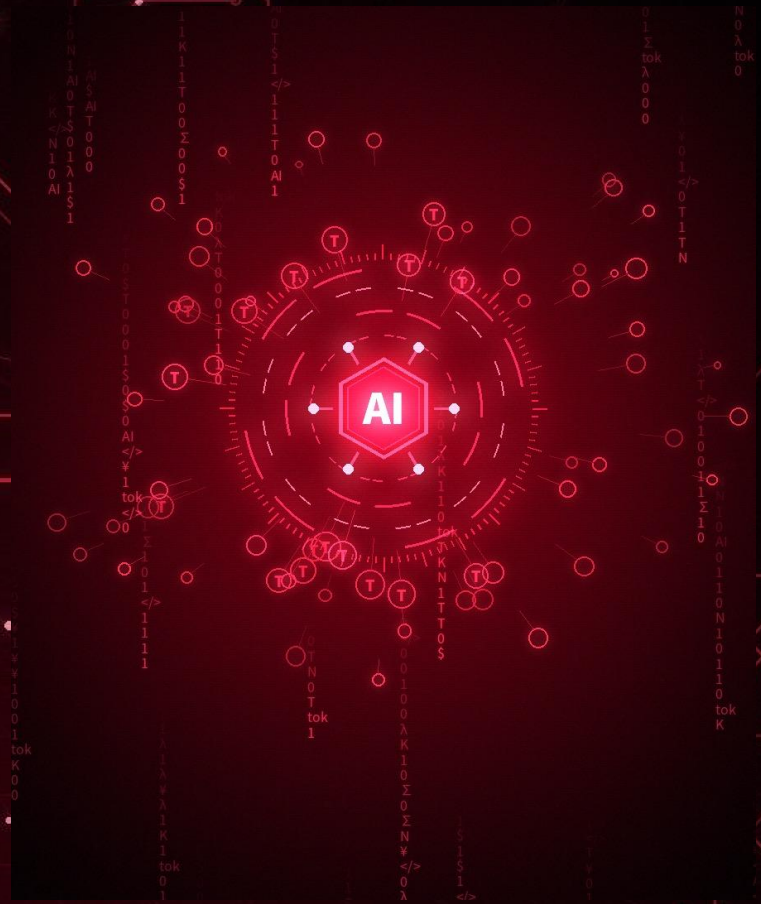
1 「トークン爆発」が経営を揺さぶる
～予算を半年弱で使い切る企業も。AIコストは新たな経営課題に～

2 米国先進企業に学ぶトークンマネジメント
～予算4カ月消化からの立て直しとROI経営～

3 可視化とガバナンス：AI FinOpsの実践
～「誰が・何に・いくら」を把握し、上限を設計する～

4 ルーティング・キャッシュ・ローカルAIでコスパ最大化
～成果を落とさずに9割削減を目指す技術的アプローチ～

5 AIエージェント時代のROI設計
～「使わせない」ではなく「賢く使う」組織と文化へ～



共通のフォーカス領域 経営視点とシステム視点の両面から、AI活用の成果とコストを両立させる実践策を提示します

トークンマネジメントの最前線を、日経BPとともに発信

日経ビジネス　日経 XTECH

なぜ、今日経BPがこの企画を行うのか

信頼される編集力

日経BP専門メディアの取材・分析で、トークン爆発の実態を国内外の事例から正確に伝える

幅広い読者基盤

経営・事業部門から情報システム部門まで、AI投資の意思決定に関わる層にリーチ

メディア連動の拡散力

日経ビジネス電子版・日経クロステックを連動させ、特設サイトや誘導枠で継続発信

協賛いただくメリット

高い訴求力	“今まさに困っている”テーマで、自社ソリューションの価値を強く訴求
質の高いリード獲得	AIコストに関心の高い見込み客を効率的に獲得
ブランド価値の向上	信頼性の高いメディアでの露出で専門性・先進性を印象付け
意思決定層へのリーチ	経営層・CIO・情報システム部門に直接アプローチ
長期的な資産に	特設サイトや記事はアーカイブ化され、継続的に効果を発揮

いま企業に求められる「AIの賢い使い方」を、貴社とともに発信します。ぜひ本特番へのご協賛をご検討ください。

本企画の3つのメニュー

編集部連動のメディア横断企画で、貴社のソリューション・ブランド価値を最大化します

セミナー

① 編集部連動1社限定ウェビナー

5つのテーマから選べる1社限定セミナー。前半は編集部の解説、後半は対談形式で貴社ソリューションを紹介。登録者リスト（想定150件）をご提供

一式300万円（税抜）

タイアップ

② 編集記事連動 特設サイト

日経ビジネス電子版・日経クロステックの編集記事と連動した特設サイトにタイアップ記事を掲載。両メディアから6週間継続して誘導

1万PV保証250万円（税抜）

動画

③ 特別解説番組 動画制作

日経BP媒体発行人または総研研究員が聞き手の解説動画を制作。特設サイト掲載とYouTube広告配信をワンストップで

一式800万円（税抜）

日経BPは、企業の「攻めのAI活用」と「賢いコスト管理」を支えるパートナーとして、最適な情報とつながりを提供します

① 編集部連動1社限定ウェビナー「AI Token Foresight」

日経ビジネス　日経XTECH

以下のテーマから1つを選択していただく1社限定セミナー。前半で編集部の解説、後半で対談形式の貴社ソリューション紹介を行います。

セミナー

▽セッション構成

時間帯	内容	登壇者
16:00～16:30（30分）	■第1部 テーマ・トレンド解説	日経クロステック編集部員 日経BP総合研究所 研究員
16:30～17:00（30分）	■第2部 対談：協賛社様×日経BP 総合研究所の研究員（最新ソリューションや事例の紹介）	ご協賛社様 日経BP総合研究所 研究員

▽告知サイトイメージ



▽テーマ（案）

- 「トークン爆発」が経営を揺さぶる～AIコストは第2の人件費か～
- 米国先進企業に学ぶトークンマネジメント
- 可視化と上限管理で実践するAI FinOps
- モデルルーティングとキャッシュでコスパを高める
- AIエージェント時代のROI設計と組織づくり

▽協賛メニュー

- オンラインセミナー講演枠：番組後半の30分間×1枠提供
- 登録者リスト提供：想定提供数150件（出欠結果、アンケートと紐づけ、開催翌日から4営業日後にご提供）
- 個別アンケートの実施（任意・運営ルールに基づく）
- 講演資料やカタログのDL（PDF・1データ20MB以内）
- 後半30分の貴社講演パートを動画素材としてご提供

▽スケジュール

2027年1月中旬～3月の期間で調整

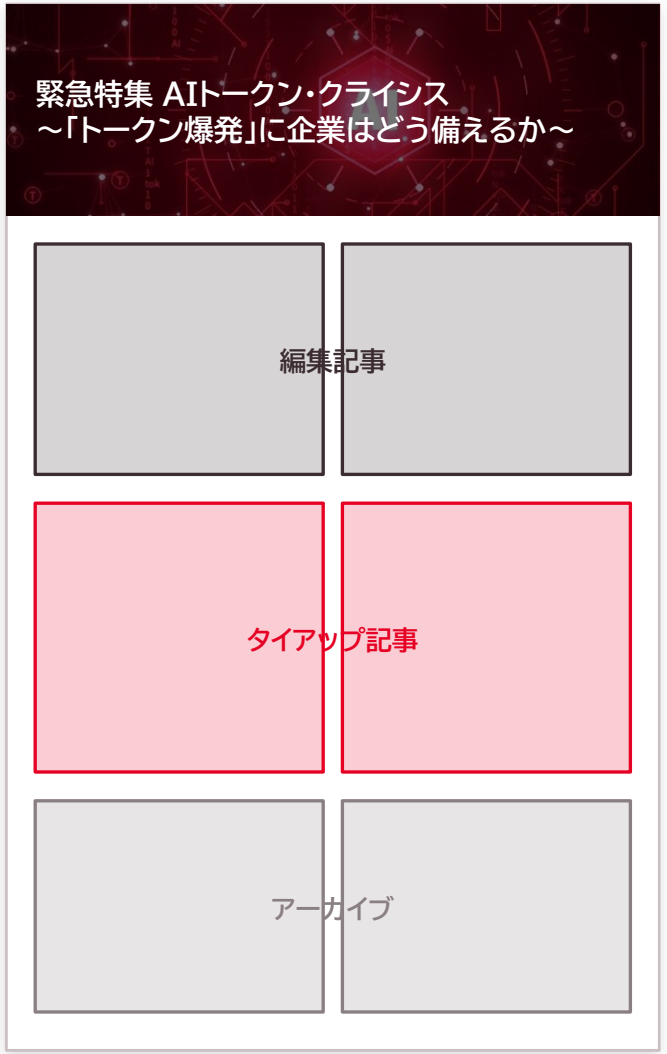
一式300万円（税抜）

② 編集記事連動「AIトークン・クライシス」特設サイト

WEBページ構成

- **実施期間**：特集サイトは12月～半年間掲載。タイアップ記事は任意の掲載開始日で6週間誘導
- **掲載媒体**：WEB：日経ビジネス電子版 & 日経クロステック
- **読者属性**：企業の経営者・マネジメント層、CIO・情報システム部門、AI推進・DX担当者等
- **誘導方法**：両メディアの各種誘導枠・メール、および日経BPの保有するリストへの特別メール配信にて、6週間継続して特集サイトに誘導
- **構成**：編集記事とタイアップ広告記事で構成

1万PV保証：250万円（税抜）



タイアップ

●編集記事
日経ビジネス電子版、日経クロステックの編集記事を掲載

●タイアップ記事広告
インタビュー取材でタイアップ広告にまとめます

●アーカイブ
誘導期間（6週間）が終了した広告記事も下部にまとめて掲載

③ 特別解説番組動画制作

日経BP媒体発行人または日経BP総合研究所研究員が聞き手になり、AIトークン対策をテーマにした解説動画を制作。
特設サイトへの掲載とYouTubeでの広告配信をワンストップで行います。

動画

タイアップ広告（記事）



動画



パッケージに含まれるもの

コンテンツ制作	配信・露出	活用権利
・動画制作 ・タイアップ広告（記事）制作 ・サムネイル・バナー制作	・日経BP各サイト掲載 ・YouTube（限定公開） ・メールマガジン配信…ほか	・二次利用権（1年間） ・自社サイト/SNS利用 ・セミナー・展示会利用

動画配信

YouTubeターゲティング広告

YouTube

日経BP専門媒体の来訪者データを活用し、YouTubeでハイライト動画をリターゲティング配信するサービス。
日経ビジネス電子版、日経クロステック、日経WOMAN、日経クロストレンドの中から選択いただき、その来訪者をYouTube上でリターゲティング配信。

仕組み（3ステップ）

データ蓄積	YouTube上で検知	ハイライト動画配信
日経BP各媒体サイトへの来訪者データを蓄積。配信元メディアは選択可。	蓄積した来訪者がYouTubeにアクセスした際を検知。Cookieベースでのマッチング。	YouTubeインストリーム広告で配信。15秒または30秒の動画を自動再生。



配信仕様

項目	内容
保証再生回数	100,000回
動画尺	15秒 または 30秒
配信形式	YouTube内広告
配信デバイス	PC・スマホ両対応
課金方式	完全視聴完了時 またはクリック
配信期間	約４週間
レポート	再生回数 クリック数

一式800万円（税抜）

日経BP

お問い合わせ

日経BP ソリューションビジネス企画部

sbk@nikkeibp.co.jp

日経ビジネス　日経 XTECH