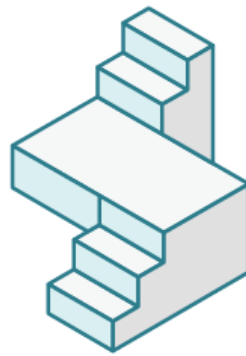# BRIKO WHITEPAPER

Briko

Briko **R**ealistic **I**nternational **K**nowledge **O**rganization

Version:0.5

Draft @ February 22, 2019

Briko Overview

What is Briko?

Briko is a blockchain technology-based open corpora and an Artificial Intelligence (AI) platform for this corpora.

Briko is also a cooperative community that benefits all of its members including corpus contributors, AI model architects, application developers, and users.

What is the Value of Briko?

The significance of Briko platform:

- Introduce a brand-new driving mode for corpora production, drastically reduce the difficulty and the cost of corpora acquisition and maintenance

- Accelerate the deployment of the AI models and algorithms

- Lower the barriers to entry for general developers to use AI models and develop Natural Language Processing(NLP)-related applications

Briko establishes a complete ecosystem for the NLP industry, and all the participants can benefit here:

- General users can provide corpus data, review the submissions and make contributions

- Researchers can provide AI models and algorithms

- Software engineers and developers can provide applications

Who benefits from Briko?

- Briko provides small and medium-sized enterprises and developers with low-cost, customizable corpora and the APIs for NLP machine learning models, helping them minimizing the cost and accelerate the development process for their applications.

- Briko provides the corpus contributors with real returns for their contributions.

- Briko provides an open API for the AI models and algorithms of the NLP and many other fields.

- The Briko project is one of the few blockchain related projects backed by the labor created values.

Background & Problems

Why do we need corpora? -- Deep learning-based Natural Language Processing (NLP)

Natural Language Processing is becoming an important branch in the Artificial Intelligence domain. It is a study of how machines can process and utilize the languages we humans use in our communications. It has been widely adopted to assist the human-to-human communication in the fields such as customer service, email correspondences, and telephone conversations. It is also the key component of machine-to-human communication and is implemented in IoT products and even internet search engines. The NLP development has revolutionized the way we communicate both with machines as well as between each other.

In the past, NLP implementations were scarce due to the fact that human level of natural communication could not yet be achieved with limited technology. With the introduction of Deep Learning, NLP development has ushered into a new era. The performance of NLP is continuously improved with pre-train methods such as word2vec[1,2], GloVe[3], FastText[8], deep learning-based models like seq2seq[6,7], Transformers[4,5], or Transfer Learning[9] and Dual Learning[9].

There are three key components that affect the performance of a deep learning-based NLP model: 1. model architecture and training scheme; 2. corpus resources as train/dev/eval sets; 3. hardware/software deployment environment. As model architectures and training schemes are often elaborated in detail in many technical papers, and the content is either directly opensourced by the authors or realized and shared by other developers based on the publications, users usually have sufficient information to choose the right model for their application. The real obstacles for obtaining a good NLP model, therefore, are the acquisition of a quality dataset and the implementation of the proper software and hardware.

The deep learning-based NLP methods have outperformed the traditional ones, however, their applications in the industry still remain limited. Other than the aforementioned obstacles, nondeterministic difficulties, high computational power requirement and high cost in deep learning-based NLP development have hindered the pivot from traditional methods for many companies.

What corpus do we need? -- Current status of corpora and technical challenges

Language model training relies on the corpus input as the raw material. The performance of the model is directly affected by the quality of the corpus. Other than using the proper model and training scheme, we also need a large amount of corpus data that fit the context, correctly

labeled, selected, verified and are up-to-date to train the model. Only in this way can we obtain a model that can make a precise prediction and process the latest expressions.

Quality corpora, especially the well classified and content specified ones, are usually controlled by large corporations. The cost for accessing them is often extremely high for small and medium-sized enterprises and developers even if they are available. Traditional open corpora, on the other hand, due to the difficulties in user connection and a lack of contribution incentives, are seldom well maintained and kept up-to-date.

Maintaining a large corpus has been proven to be a difficult task. It requires a large amount of human labor to add, filter, revise and review the data, however, the existing open-source corpora are rarely maintained in such manner. We urgently need an adequate incentive system to attract more participants for the corpora building process.

Why do we need blockchain to build corpora?

The emergence of blockchain technology is completely changing the industry and business globally and revolutionizing the economic system. It "achieves the low-cost transfer of the economic values" [11] via the internet and provides a brand new approach to solve the aforementioned problems.

- Data saved on a blockchain is open and immutable, witnessed by everyone and not monopolized. It makes the ownership of a corpus, model or application well documented. The contribution and usage history of corpus data is traceable. Infringement and malicious submission of the corpus can be easily detected and verified.

- The smart contract specifies the distribution of the labor value clearly via code, each execution record is saved on the chain and cannot be changed or withdrawn. The transaction is completely transparent to all the stakeholders. This protects the interests of the contributors and guarantees fair value transfers amongst all parties.

- With the help of open and mutually beneficial ecosystem and innovative value distribution model, we can build a connection, never seen before, between corpus contributors and users, and push forward the new model for the corpus contribution system. This ensures the quality, accuracy, and time-effectiveness of the corpora. Under this new cooperation mode, the contributors and the users will be motivated to work together and therefore, share and further amplify the value created in this ecosystem.

- The blockchain-based architecture is favorable to the third-party developers, can attract more developers get involved to build third-party applications, tools and plug-ins using all kinds of resources on the Briko platform, improve the platform with the Briko team,

expand the diversity of corpora and models in productization, and enable all kinds of possibilities for the entire industry in the long run.

Opportunities and Solutions

The blockchain technology will help us establish a free and open corpus market and AI model platform, and to get corpus and model contributors and users involved in the closed-loop ecosystem for the corpus construction. Through blockchain technology, the Briko platform will redefine the value distribution, integrate the interests between the upstream and the downstream of the NLP-related industries, realize low-cost value transfer between different parties. The platform also maximizes the interests of all participants through fair and transparent smart contracts, incentivize more contributors to get involved and change the driving force of corpus construction.

For corpus and AI model contributors, their ownership of the validated work will be established and they will directly receive BGE (Briko Granted Endorsement) as the evidence of their work. When the corpora and services of the Briko platform are purchased for use, Briko will transfer its income to BGE holders through public repurchase on the secondary market. As a registered non-profit organization (NPO) in Canada, Briko's operations will be under the supervision of the CRA. Briko's financial report will also be disclosed and therefore subject to the supervision of our community. This will ensure the maximized revenue transfer from the platform to our contributors and thus achieve the fairest value transfer between contributors and users. The contributors' profit will be directly tied to the quality and quantity of the contribution. This innovative corpus contribution mechanism with clear and direct incentive distribution system will continue to attract contributors to provide high quality, up-to-date corpus data.

For the users of corpora and models, corpus and model can be selected, customized and purchased based on the labels and the classifications in an easy and self-serving manner. A tailored corpus can be used in the model training process to achieve its optimal performance in the specific scenario. Users will no longer be required to purchase the often expensive corpus from large companies with data monopoly. For developers who want to apply AI in their products, various pre-trained models trained will be available on the Briko platform. A significant amount of saving in development cost can be achieved for small and medium-sized development teams.

At the application level, we hope to integrate state-of-art open-source NLP models and frameworks, take advantage of the abundant corpus resource on the Briko platform, and build an AI platform for users. We will provide easy-to-use, reliable and flexible model

training framework for users to train their own models, and NLP API for users to use directly for fast development and deployment of their NLP-related products.

Potential users of the Briko platform include commercial organizations and academic institutions.

Business organizations can incorporate their own professional corpora into the Briko platform to get more people involved in the corpora constriction while earning BGE rewards. Meanwhile, Briko platform provides them with a testing environment with a low-cost, giving them the opportunity to rapidly build their prototypes, test their concepts and lead the market by swiftly deploy their products. For small and medium-sized enterprises and individual developers, high-quality corpus at a low price are provided for their application development. The Briko-owned models and APIs also lower the entry barrier for the technology, getting more talents involved in the innovation of the NLP field in addition to the large companies and a few core talents.

For academic institutions, Briko platform will provide corpus and model benchmarks, as well as the performance baselines of these models for scholars to compare with. In the meantime, the academics will be able to witness the usage of their AI models and algorithms, identify the challenges in the application and hence get a clearer direction for their research projects.

High-quality, low-cost corpora, as well as the open development platform, will significantly reduce the entry barrier for development, promote innovation and prosperity in applications related to NLP.

Technology and Architecture

System Architecture

Briko system is consist of the Briko corpora, which is based on blockchain technology and collaboratively maintained by the community, and the Briko AI platform built upon the Briko corpora.
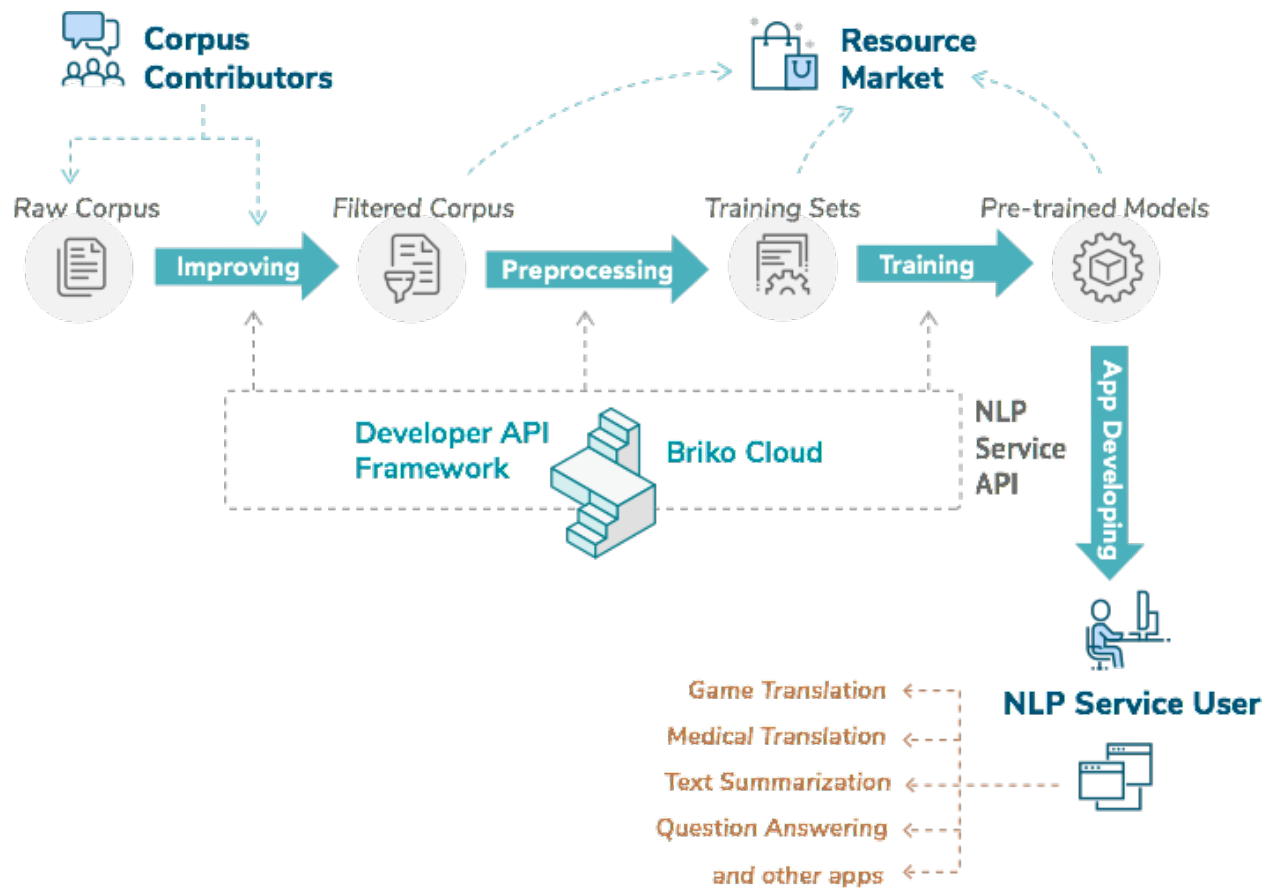
Corpora

Briko will be working a corpus software running on a blockchain system based on the EOS open-source project. The blockchain system acts as a side chain on the EOS main network and enriches the EOS main network ecosystem. The main part of the corpus software will be built based on EOS smart contract technology, to achieve highly secure, efficient and automatic software architecture. Most of the main features, such as the distribution of corpus-related tasks, task acceptance, submission, settlement, BGE repurchase, etc., will all be executed by the smart contract. All the bookkeeping and asset management will also be recorded on the chain for inquiry and publicity. All smart contract source code involved in the Briko project will be open source.

To achieve the following purpose including but not limited to:

- Labor compensation (corpus translation, organizing, scoring, etc.)
- Malicious activity punishment and economic system protection
- Corpus users payment (API usage, translation model usage, corpus download, etc.)

Briko will build an economic system based on BGE.

Briko will provide cross-platform apps as well as RESTful APIs for corpus access. By accessing the corpus API, developers can embed functions such as corpus uploading and

tasks reviewing into their own applications, obtaining BGE rewards while realizing NLP functionalities.

Briko corpus system will provide records of access, contribution, usage and authorization of the corpus on the EOS side chain with the EOS smart contract.

Briko AI Platform

The other component of Briko is the open-source AI development platform based on the on-chain corpora. State-of-art NLP models are integrated into the platform framework and are modularized and generalized to provide a development platform for users to build and train their own customized models. Briko will also open source our own models trained with the corpora, as well as the APIs for common NLP tasks based on these models, so users can directly call the API to realize functionalities such as translation, Q&A and semantic analysis. The Briko platform will greatly lower the technology entry barrier for related applications and promote innovation and development of NLP-related applications.

- **Model Building and Training**

With the continuous development of the deep learning in the NLP field, more and more accurate and efficient models are emerging (such as Transformers[4,5] and seq2eq[6,7], and pre-trained models like BERT and GPT-2). Briko AI platform will continue to provide the latest models for the developers to use. Through the means such as modularization and generalization, we will simplify the process of building and training models, maximizing the time saving for our users.

Briko will be compatible with the major deep learning frameworks, such as TensorFlow, PyTorch, Keras, Caffe, etc. We will also open source our own platform and encourage general developers to participate in the platform construction.

- **Briko pre-trained Models**

The model training process is often time-consuming and requires a significant amount of computational resource, especially the NLP applications that require complicated models and a huge amount of data. It usually takes days even weeks and multiple GPUs to train the model. Therefore, in addition to providing our users with an easy-to-use model-building framework, Briko also the open source pre-trained models. These general pre-trained models can be used directly, it saves users the time and computational power required in the development process. Users can also use transfer learning to fine-tune self-contained models for their own application scenarios. Briko open-source pre-trained models will keep evolving along with the updated corpus.

- **Briko AI API**

Deep learning models are very complicated and hard to understand and therefore not user-friendly for many developers; however, for some non-critical general applications, it's not necessary for developers to build and train their own models. Based on our own open-source pre-trained model, Briko provides APIs that completely mask the model part, so the developers can easily realize many general functions, such as translation, Q&A and sentimental analysis. We will also be continuously adding APIs with new functionalities to better fulfill the need of applications.

Briko Framework

Design concept

Briko framework is an open-source platform based on blockchain technology. It aims to:

- Break the data monopoly of large corporations and cartels and providing high-quality corpora at low cost to the general public.

- Significantly lower the barrier for the NLP application development, reduce development costs, and promote the prosperity of NLP-related applications.

Briko is a collaborative community consists of corpus contributors, model designers and application developers. Through the smart contract based on blockchain technology, we strive to maximize the interests of all the groups in a transparent and efficient manner and promote the prosperity of the entire ecosystem.

Briko platform framework
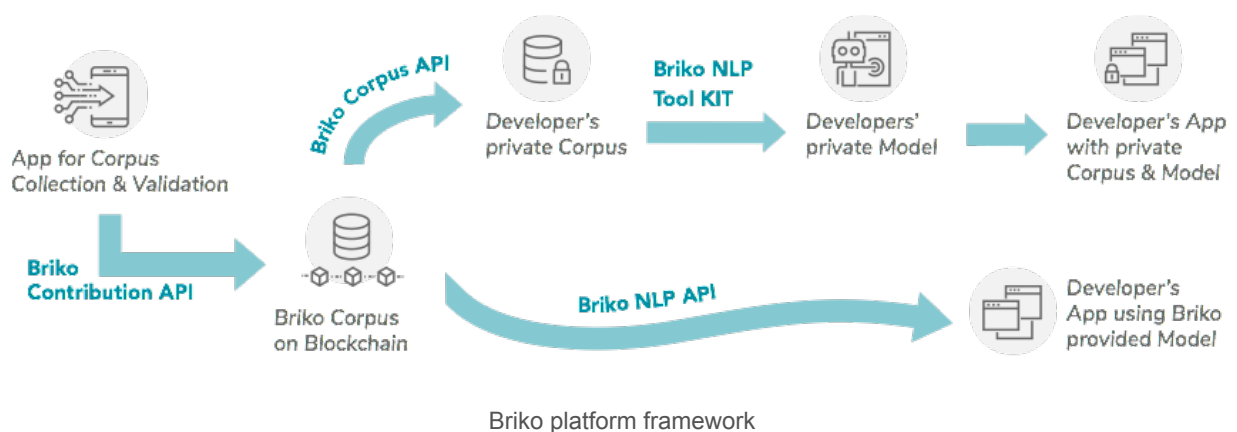
Briko Platform Owned Applications

- **Briko Website**

Briko.org is the official website of the Briko project, it will provide the management feature of Briko corpora, algorithms, and models, including model reviews, corpus classification and management, authorization management and tracking, corpus data uploading, task distribution, processing and submission, blockchain storage access, wallet management and many other basic features. We will post all the tasks on the Briko website first as a starting point.

- **Briko mobile app**

Briko mobile apps make it easier for general users to interact with Briko corpus. Briko mobile apps will be compatible with both Android and IOS. Briko mobile apps will include basic features such as corpus data uploading, task distribution and submission, blockchain storage access and wallet management. The mobile app will be developed after the website is completed.

Third-party DApp

Briko is a non-profit organization, we are committed to lowering the barriers to technology and enabling more developers to use cutting-edge technology to create more value. Therefore, Briko encourages the developers to use our pre-trained models and APIs to develop third-party applications. Here are a few Dapp examples to show the possibility of productization using Briko corpus, models and APIs.

- **Novel Translator**

This translator is optimized based on the category, topic, and domain of the content. It requires minimum to none human efforts in calibration or revision of the translation for the novels in a foreign language. The primary targeted users of the product are consumers of novel content in the entertainment and leisure categories. This application can be realized as APIs for business clients or as web services that is directly open to the end users.

- **Application Document Writing Assistant**

It provides translation and writing assistance for application document compositions. Leveraging the quality and diverse corpus resources and pre-trained models, users can choose the desired language, document type, and application scenario to get format suggestions, translations, grammar correction, formal/informal format, recommended sentences, high-

frequency words, suggestions for ceremonial greetings and other assistance. Not only can it fulfill the personal and business cross-language communication needs, such as the composition of emails, reports, blueprints, notices, dissertations, etc., and improve communication efficiency and effectiveness, but also can it help tourists and immigrants solving their daily problems, like complaints, inquiries, apologies, congratulations, etc.

- **AI Language Tutor**

It can fulfill the common needs in foreign language studies. The UX can be customized for groups of users of different ages. For example, the robot can reply to questions like "how to say 'apple' in French?" or "how to say 'may I use the washroom?' in Chinese?". It can also, for instance, translate conversations should the users need to communicate with foreign language speaker directly. This application combines functionalities like Q&A, machine translation and speech recognition, and embodies more possibilities of the Briko platform - corpus acquisition and optimization for the training of speech recognition models.

Briko Ecosystem

The blockchain as a new technology of disruptive innovation has fundamentally changed people's perspective of individual productivity. It is a brand new world that gives stronger motivation for individuals at work, improves efficiency by reforming the cooperation between organizations, which empowers and redefines the economy.

But where do these changes come from?

- By holding the BGE, participants are the contributors as well as the shareholders in the entire ecosystem.

- The growth of the ecosystem will be fairly distributed to its participants, contributors, and investors through the increase of BGE value.

- It becomes one of the driving forces for the participants to cooperate spontaneously, contribute and create value for the ecosystem, and;

- The distributed, fair, transparent and consensual natures of the blockchain, not only increase the trust amongst people, improve the efficiency of cooperation, but also provide more ways of participation, giving individuals more power as well as opportunities.

Briko's vision is to build a value network within which the general public provides data, researchers provide models and developers provide products. It should function as self-organization and evolve continuously. The value created by members is shared across the

ecosystem. The collaboration approach of the Briko ecosystem encourages diversity, innovation, and builds systematic and orderly symbiotic relationships.

Stakeholder

- Briko organization: Creator of Briko projects.

- Developers: Third-party Dapp, toolkit, extension developers, source code contributors, smart contract developers, and all technical contributors to the Briko ecosystem.

- Resource contributors: Briko project participants who submit, edit or review the text corpora, or provide AI models on the Briko platform.

- Resource users: Anyone who uses the text corpora, AI models, applications and services provided by the Briko platform, including small businesses, individual developers, academic institutions, researchers and students.

Briko Economic System

Briko Granted Endorsement (BGE) -- The token for the Briko platform

BGE is issued on the EOS side-chain. One of the major differences between BGE and other cryptocurrencies(e.g. Bitcoin) is that BGE is not created for general exchange purpose. BGE is used to evaluate the contribution to the Briko ecosystem, such as new AI model and text materials provision, text corpus revision and maintenance.

As the basic form of currency in the Briko ecosystem, participants will need to execute smart contracts to obtain BGE as the recognition and the reward for their work. AI platform services licenses and text corpora can be purchased with EOS currency, and the Briko's income will be used for public repurchase of BGEs on the secondary market. The repurchased BGEs will be destroyed and thus permanently exit the market. Participants are requested to deposit a certain amount BGEs in order to accept Briko issued tasks. The reward will be granted upon task completion and the deposit will be return in full after confirming the non-malicious submission. The deposit BGEs, however, will be confiscated if the submission is deemed as malicious, and will either be permanently destroyed or re-issued as rewards.
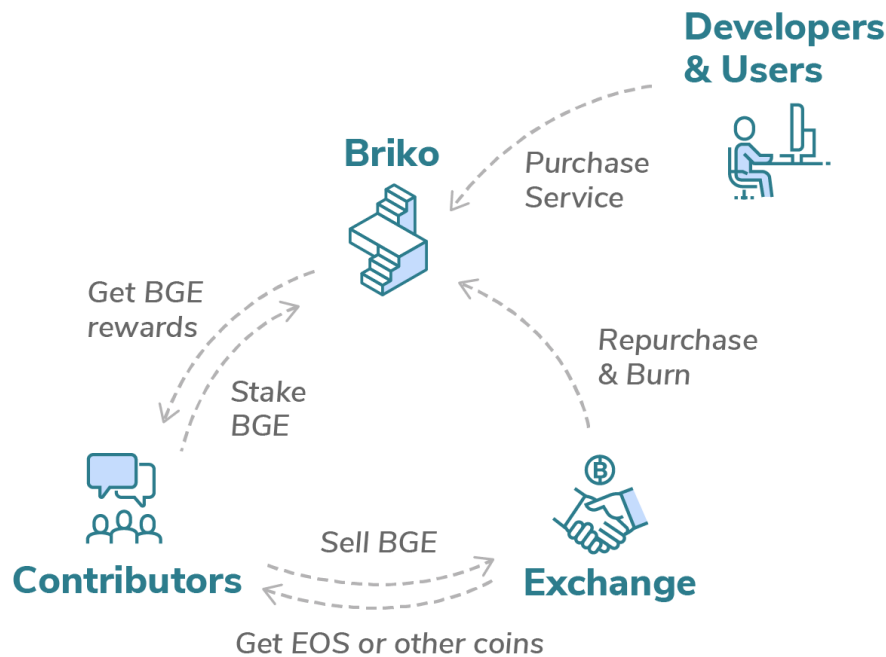
Reward Mechanism -- How to motivate the user contribution?

- **Various forms of rewards to encourage community participation**

BGE is the main form of reward to users who submit valuable text corpus, AI models and reviews. In addition, Briko platform will diversify the reward system by setting various targets for community users. For example, we will be introducing a user-level system that is based on the experience and the time input of each user. The user-level system will work cooperatively with the BGE reward system to incentivize the community users to continuously provide quality contributions. In the meantime, Briko encourages our community to make new rules about the reward system and even establish new systems to promote user contribution according to changes iteratively.

With the development of NLP applications, the expectations of the AI models and text corpora will also evolve. Briko will also upgrade its ecosystem and economic system in order to adapt to the changes. Briko will work with our community and distribute more resource to the sub-projects in need such as specific text corpora and newest and most effective AI models. Economic leverages can be applied in such cases when some less popular text corpora or AI models are missing.

- **Application-driven reward -- more reward for more contribution**

Briko aims to establish a sustainable ecosystem where users with special contributions can be rewarded continuously. More tasks completion recorded on the blockchain by Briko smart contracts equals better chances in receiving the extra reward. The application-driven model of

the reward system takes advantage of blockchain technology and it is expected to solve the hard problem of how to measure user contribution in a value system. Our economic model will effectively motivate users to provide the most advanced AI models and quality text corpora.

Review Mechanism -- How to ensure the quality of user contribution?

The contributions of community users in the form of AI models and text corpora need to be reviewed and approved by the reviewers. Reviewer is a special role among the community users. Reviewers take responsibilities of maintaining the quality of user-submitted text corpus and defending against attacks initiated by malicious users who try to spam the system by submitting incorrect content. Community users can also assume the role reviewer for other submissions while submitting their own text corpora to the Briko platform at the same time, all the work will receive BGE reward.

Briko platform will also establish a mechanism to evaluate the reviewers' work. Similar to the procedures of submitting text corpora, the reviewers have to deposit a certain amount of BGE when submitting their reviews. Deposited BGE will be refunded after reviews get confirmed. Unconfirmed reviews will result in the confiscation of the deposit in order to ensure the quality of reviews.

In addition, to defend against the collusion attacks from contributors and reviewers in which reviewers approves all submission even with the ones with malicious intents, Briko platform will randomly issue challenge tests to the reviewers. The review tasks will be mixed with known and trivial tasks to screen for the malicious or extremely careless reviewers.

Credit System -- How to encourage user participating for the long run?

The Briko economic system encourages users staying in the community and make continuous contributions, and rewards trusted users who participate in building the community in a long run. The user credit system is designed for this purpose. High credit will be given to user submitting text corpora and AI models with high quality. Users with higher credit can receive more BGEs when their tasks are done, and they will be invited to the core community for making rules and community management.

The user credit system provides evidence for rewarding continuous contributions from users, and encourages the participation of skilled users. In the meantime, for newcomers of the community, the credit system can motivate the new users to learn the community rules, operation standards and creating high-quality resources.

Core Team

Miao Cui

More than 10 years of experience in large project management and software development。
Lead projects in many Internet companies, held positions as senior manager of technology.
Master degree from Beijing University of Posts and Telecommunications.

Wei Huang

10 years of experience in system software development and virtual currency system
management. Research in peer-to-peer network and virtual currency systems based on peer-
to-peer network from 2008. Early participant of the Bitcoin network. Master of Computer
Science and Engineering and research experience from the University of Hong Kong and
University of Toronto.

Shuze Zhao

Doctor of Philosophy in EE from University of Toronto. Research areas include data centre,
energy savings in high-performance AI microchips and deep learning based Natural
Language Processing.

Yuhui Feng

Master of Applied Science from University of Toronto. Research areas include human-
machine interactions of robotic systems, Integration and application of intelligent mechanics
in industrial settings and healthcare industries. Years of experience in software development.

Ju Huo

Over 20 years of experience in software development, architecture, and technology
management. Founder of the pioneer enterprise search engine and cloud computing service
company in China: Ginkgotek. Senior researcher of Shengda Innovations. He provides
technical consulting services to international media corporations such as NYT Chinese and FT
Chinese as an independent consultant. Ju is also one of the most influential bloggers in the
technology and hacker/crypto field in China.

Reference

[1] Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space." CoRR abs/1301.3781 (2013).

[2] Mikolov, Tomas et al. "Distributed Representations of Words and Phrases and their Compositionality." NIPS (2013).

[3] Pennington, Jeffrey et al. "Glove: Global Vectors for Word Representation." EMNLP (2014).

[4] Vaswani, Ashish et al. "Attention Is All You Need." NIPS (2017).

[5] Dehghani, Mostafa et al. "Universal Transformers." CoRR abs/1807.03819 (2018).

[6] Sutskever, Ilya et al. "Sequence to Sequence Learning with Neural Networks." NIPS (2014).

[7] Cho, Kyunghyun et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." EMNLP (2014).

[8] Joulin, Armand, et al. "Bag of tricks for efficient text classification." arXiv preprint arXiv:1607.01759(2016).

[9] Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." ACL (2018).

[10] He, Di, et al. "Dual learning for machine translation." Advances in Neural Information Processing Systems. 2016.

[11] Yukio Noguchi. "Blockchain Revolution", publisher Nikkei, 2017