循环智能
RECURRENT AI

更懂
销售对话场景
语音识别&实时转写

循环智能（Recurrent AI)
2022年2月

循环智能
RECURRENT AI

更准

更快

更可靠

循环智能
RECURRENT AI

「准」，取决于“算法模型”和“行业数据积累”

循环智能
电话录音ASR更准

=

世界级
原创算法模型

×

四百万+小时
金融等领域的数据积累

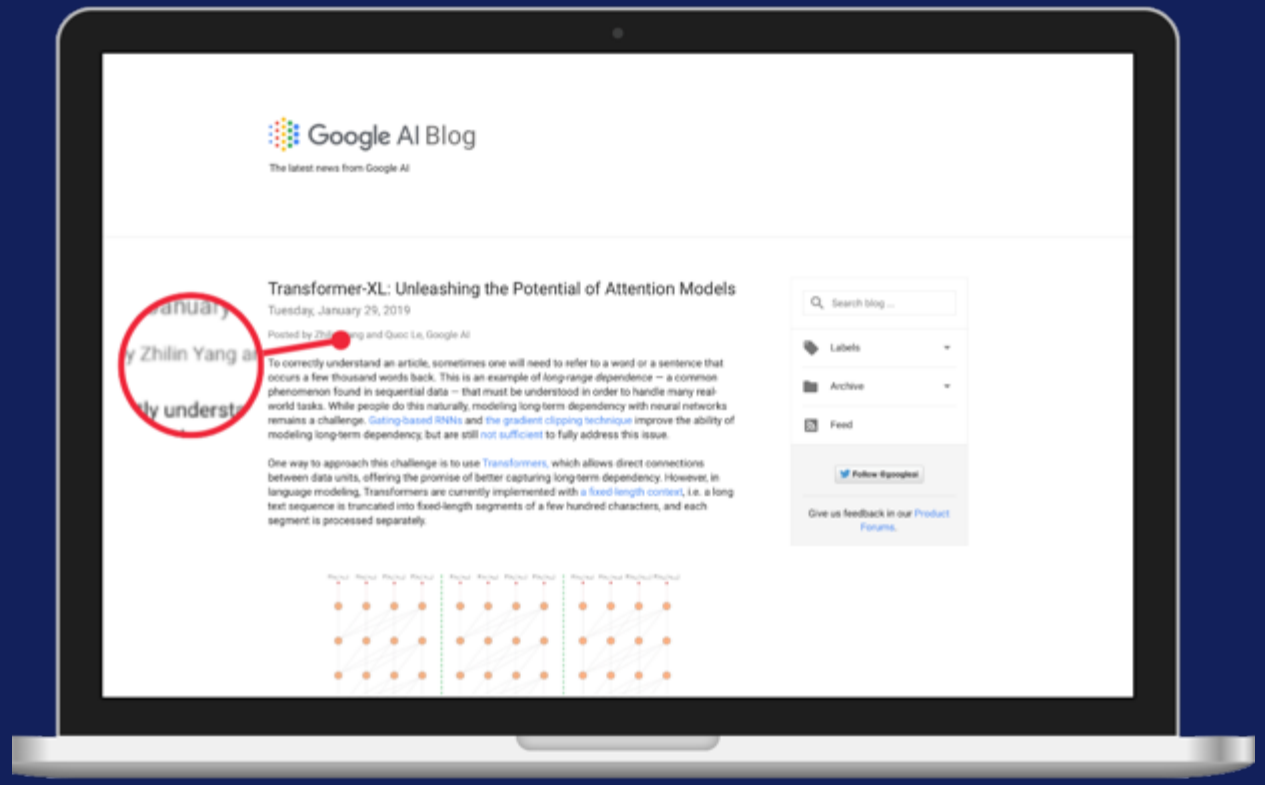
循环智能与Google、CMU联合开发Transformer-XL

循环智能已服务几十家销售超过 1000 人的企业

循环智能
RECURRENT AI

与 Google 联合推出 Transformer-XL 算法模型：在全部 6 个语言建模数据集取得最优结果

2019年初，循环智能联合创始人杨植麟博士与Google、卡内基梅隆大学联合推出 Transformer-XL 模型，在全部 6 个语言建模数据集取得最优结果（state of the art）。

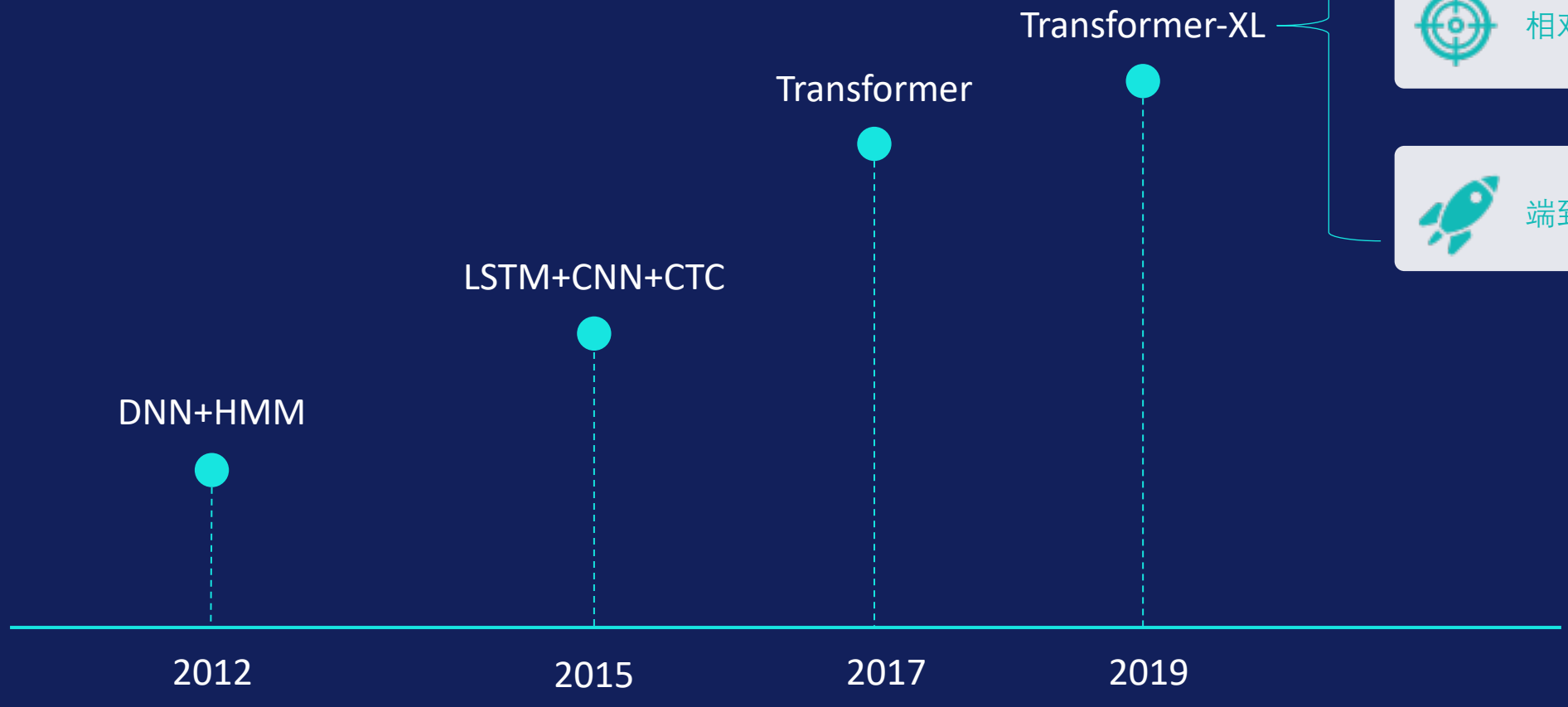


全部 6 个数据集 No.1

循环智能 RECURRENT AI

Transformer-XL 能对更长的上下文建模，拥有更强的语言表达能力

近几年，语音识别领域主流底层算法模型快速演进。Transformer-XL 在 Transformer 的基础上引入循环机制和相对位置编码，进一步提升了语言表达能力。

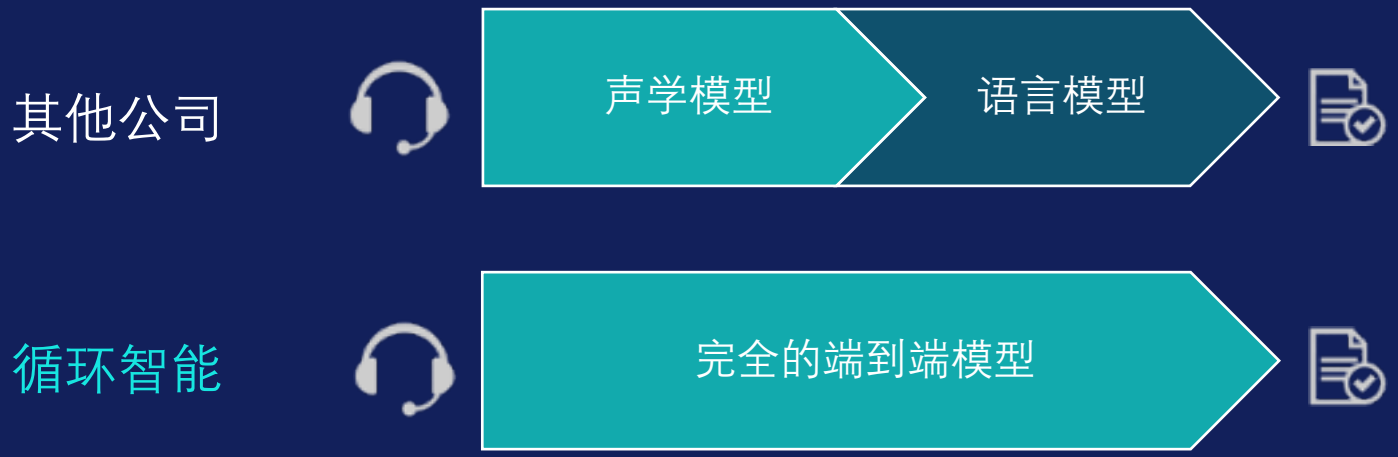


- 循环机制
- 相对位置编码
- 端到端识别

循环智能
RECURRENT AI

循环智能基于 Transformer-XL 模型实现更前沿的“端到端”语音识别

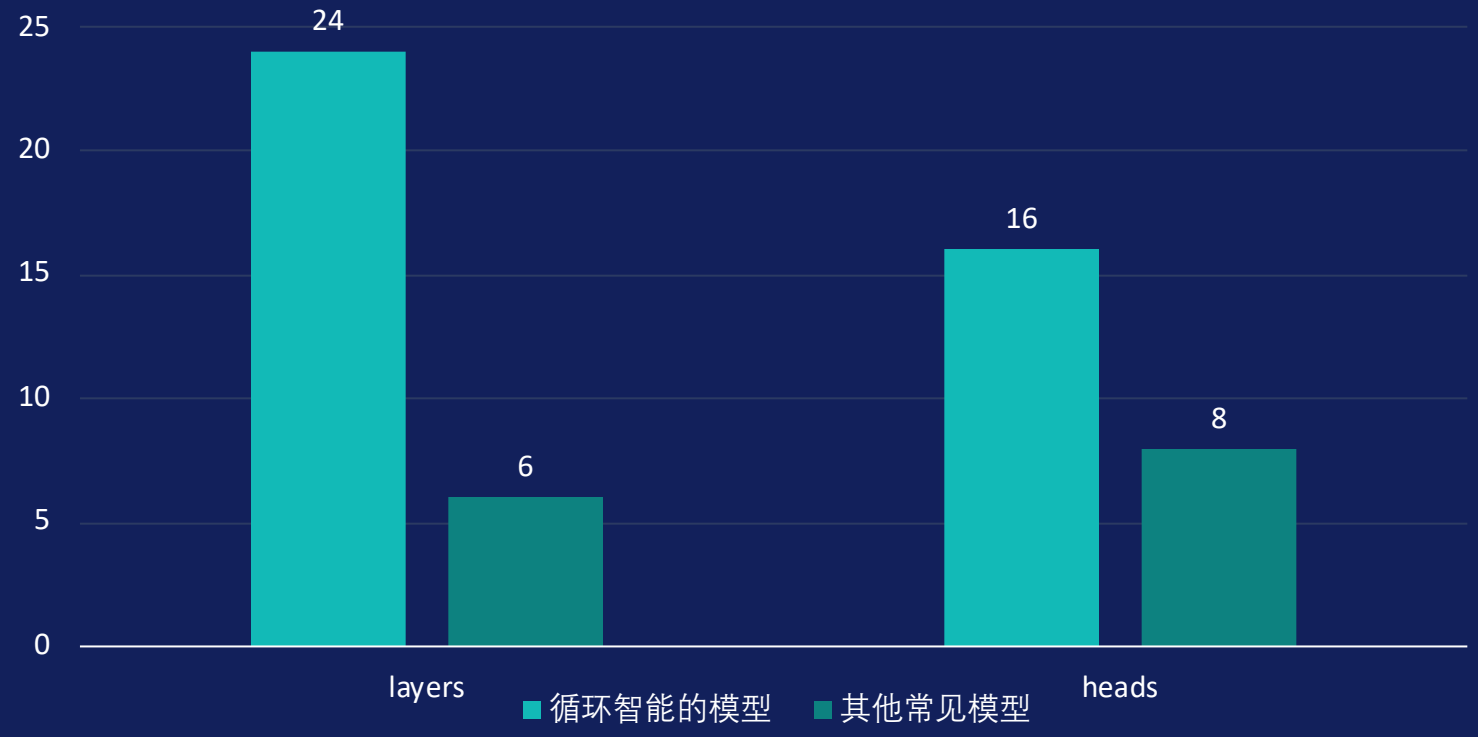
其他公司大多仍在使用基于声学模型+语音模型的传统语音识别方案，两个步骤是独立的，无法联合优化。循环智能则基于Transformer-XL 模型，自研了端到端语音识别系统，系统中的每个组件都可以联合优化，实现更准的识别结果。



循环智能
RECURRENT AI

通过更深的模型层数，捕获远距离文本特征，带来更准的识别结果

更深的模型解决了远距离信息提取的问题，循环智能原创的 Transformer-XL 模型采用 24 layers 和 16 heads，比其他常见模型 6 layers 和 8 heads 模型高出两倍以上。



循环智能 RECURRENT AI

在银行、保险、证券、房产和汽车等领域，经过四百万+小时语料训练

从2018年开始，我们已服务了多家销售人员数量超过 1000 人的金融、房地产和汽车行业公司，电话录音ASR模型经过超过四百万小时的行业语料训练。

中国工商银行	招商银行 CHINA MERCHANTS BANK	中信银行 CHINA CITIC BANK	上海银行 Bank of Shanghai	PICC 中国人民保险	太平洋保险 CPIC
中国太平 CHINA TAIPING	泰康人寿	招商信诺	MetLife 大都会人寿	百年人寿 AEON LIFE	众安保险
vanke 万科	我爱我家 www.5i5j.com	安居客	中原地产 CENTALINE PROPERTY	ziroom 自如	猎聘 Liepin.com
SAIC 上汽集团 SAIC MOTOR		VOLVO	中国東方航空 CHINA EASTERN	捷信 HOME CREDIT	波司登 BOSIDENG

循环智能
RECURRENT AI

实测数据：来自百万级订单客户，循环智能击败巨头公司

2019年起，循环智能将基于Transformer和Transformer-XL的端到端算法模型部署到业务实践中，取得显著成效。在某数百亿市值公司的实际测试中，循环智能的电话录音识别结果，超过国内公认的语音巨头和互联网巨头。

	语音识别错误率	说话人分离错误率
循环智能	-	-
某语音巨头	+3.96%	+27.87%
某互联网巨头	+19.68%	+46.54%

循环智能
RECURRENT AI

循环智能
RECURRENT AI

「快」，取决于“高效的模型架构”和“CUDA底层优化”

Transformer-XL
推理阶段速度
比Transformer
快300~1800倍

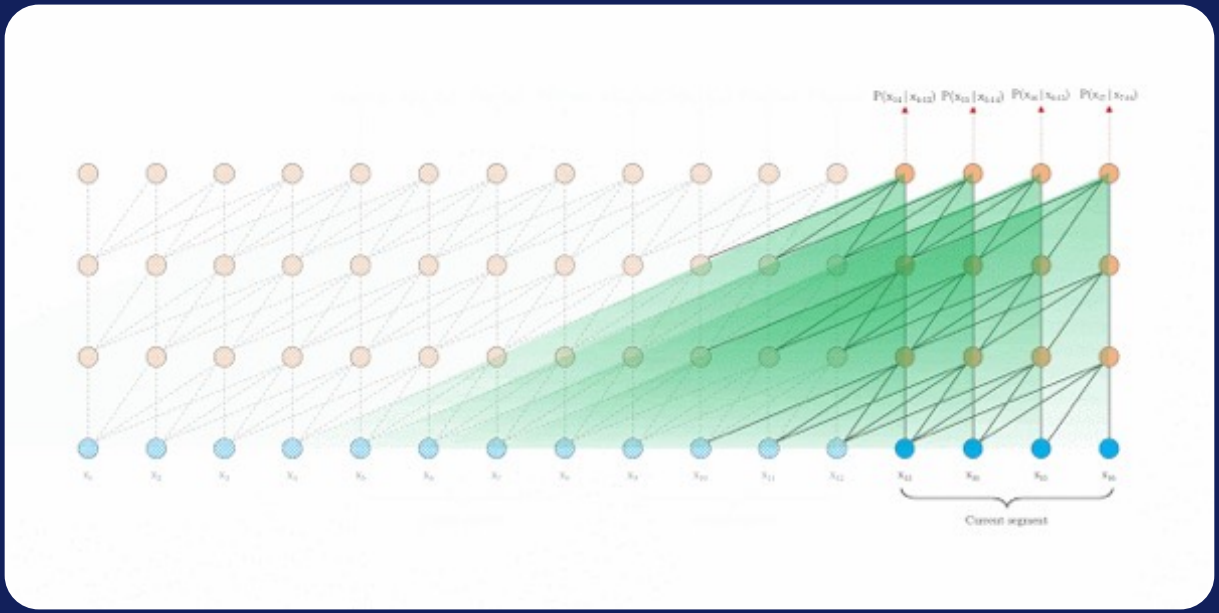
高效
模型网络架构

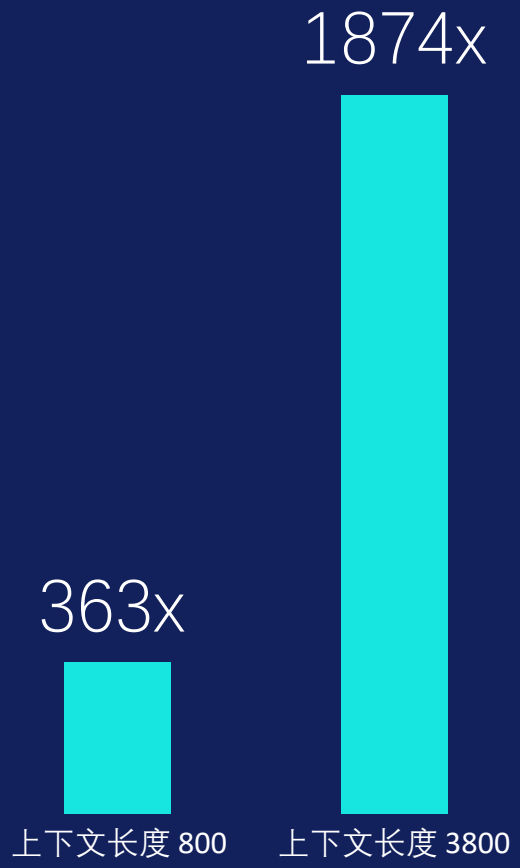
CUDA底层优化
自研CPU、GPU调度算法

最高支持
1CPU+8GPU的
高效运算

Transformer-XL 模型高效的网络架构推理速度是 Transformer 的 300~1800倍

Transformer-XL的推理速度也明显快于 Transformer，尤其是对于较长的上下文。比如，在上下文长度为 800 时，Transformer-XL提速 363 倍；而当上下文长度增加到 3800 时，Transformer-XL提速 1874 倍。

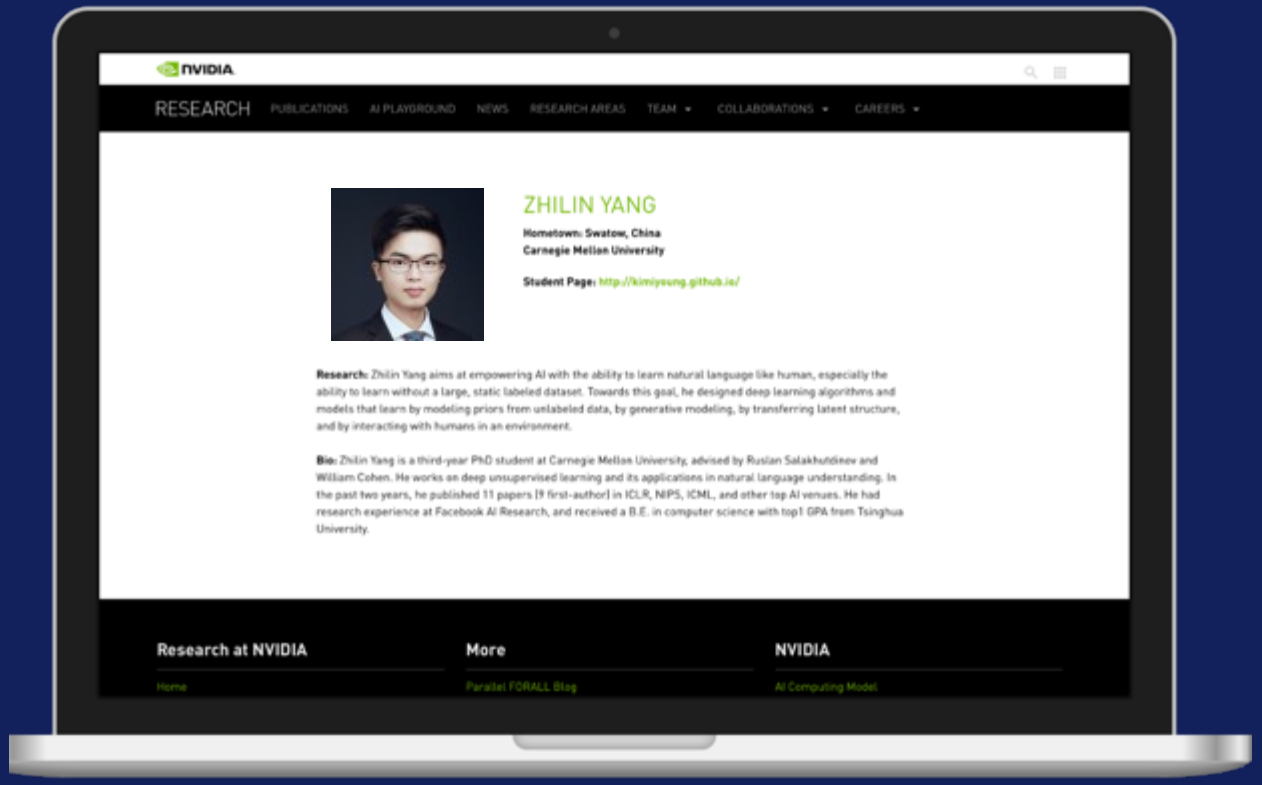




循环智能
RECURRENT AI

杨植麟博士因在 GPU 并行计算领域的研究，获得 NVIDIA Graduate Fellowship

2018年，循环智能联合创始人杨植麟获得 NVIDIA Graduate Fellowship 以表彰和鼓励其在GPU并行计算领域的研究，全球仅 11 人获此殊荣。



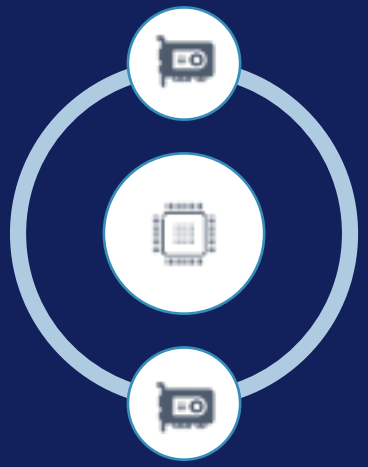
全球仅 **11** 人获此殊荣

NVIDIA
GRADUATE FELLOWSHIP PROGRAM

循环智能
RECURRENT AI

通过CUDA底层优化，最高支持 1CPU+8GPU 的高性能计算组合

循环智能自研了调度算法，对GPU并行计算的CUDA底层进行了充分优化，可以更好的平衡CPU和GPU资源，在几乎不损失性能的情况下，最高支持 1CPU+8GPU的高性能计算组合，行业平均水平1CPU+2GPU。



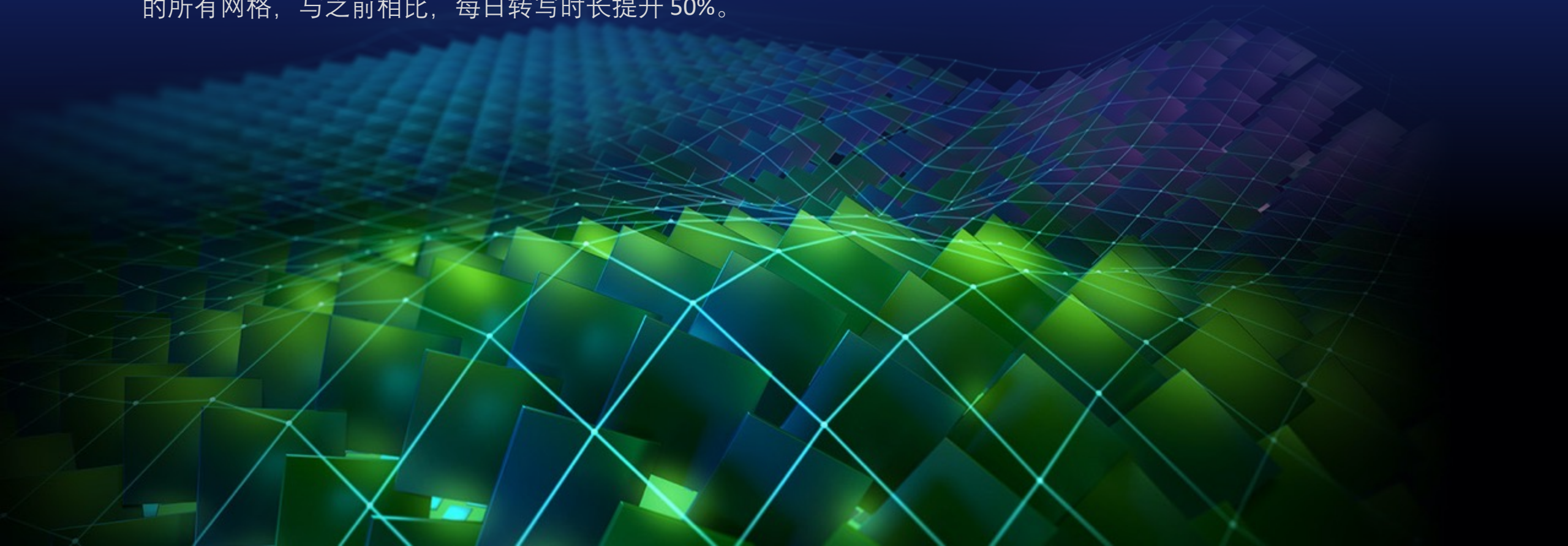
1CPU + 2GPU
行业平均水平

1CPU + 8GPU
循环智能

循环智能
RECURRENT AI

借助空白录音剪除和拼接技术 “填满” GPU计算矩阵的所有网格，性能提升 50%

电话录音的空白通时在GPU并行计算矩阵中会浪费掉资源，循环智能通过独家工程优化，将空白录音剪除，将不同长度的录音进行拼接，从而 “填满” GPU计算矩阵的所有网格，与之前相比，每日转写时长提升 50%。



循环智能
RECURRENT AI

实际数据：单GPU每日转写双轨录音 960 小时

由于采用了高效的网络架构 Transformer-XL，并且对GPU并行计算的CUDA底层进行了充分优化，循环智能目前可以做到单日单个GPU可转写双轨录音 960 小时（基于目前主流的 NVIDIA RTX 2080Ti/Tesla P100/Tesla T4 GPU），行业平均水平不到 800 小时。

循环智能
RECURRENT AI

实际数据：实时转写的端到端延时低于 150ms

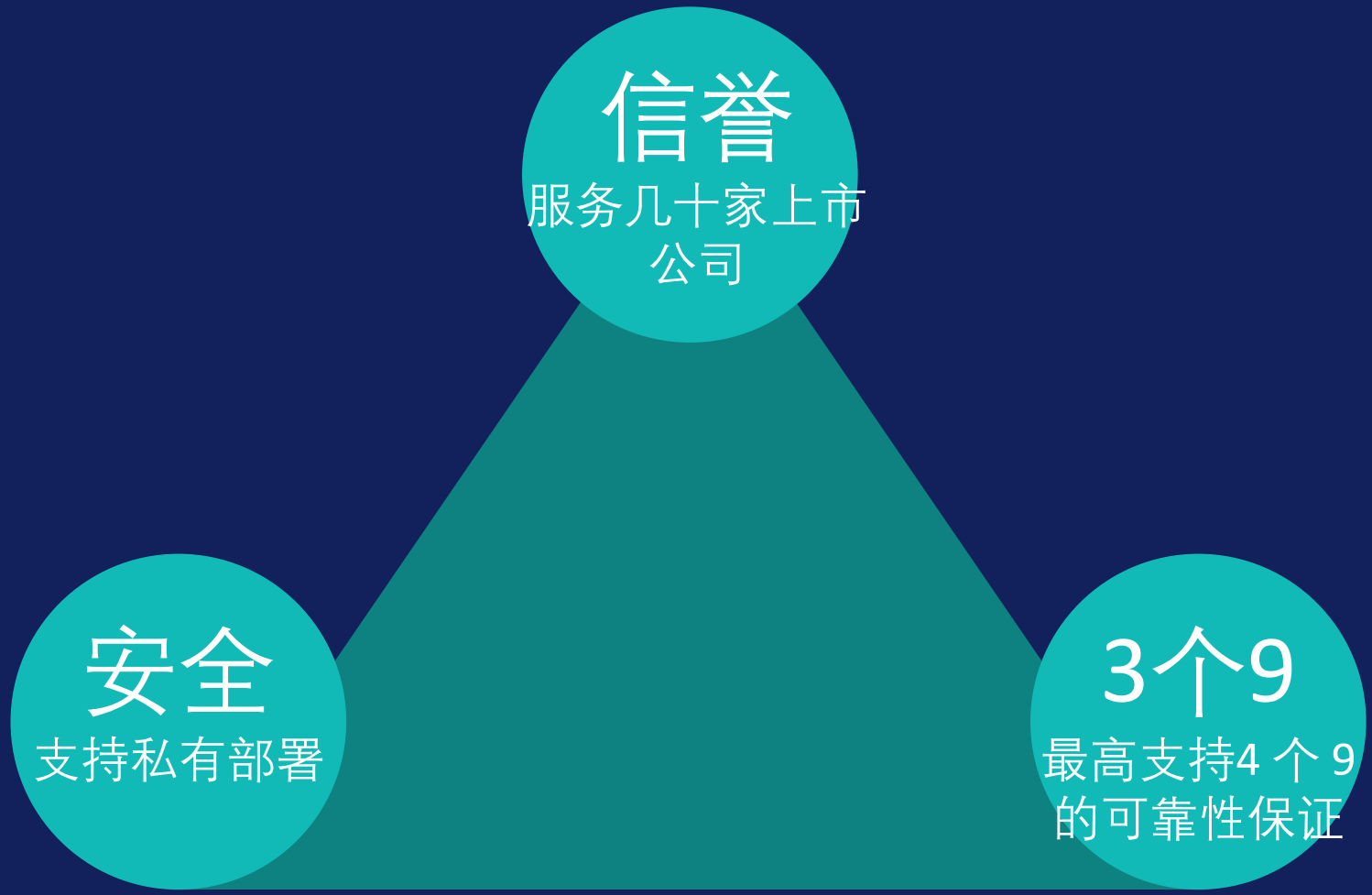
循环智能提供实时的电话语音转写为文字服务，支持 MRCP 等主流传输协议，8K 采样率。在维持高准确率的基础上，我们的端到端延时低于 150ms，当前行业平均水平在 500ms 之上。

循环智能
RECURRENT AI

更可靠

循环智能
RECURRENT AI

「可靠」，源于 “信誉” “安全” 和 “3个9可靠性”



循环智能
RECURRENT AI

安全快速的部署方式：本地私有部署、云端私有部署和SaaS模式三选一

我们将根据实际业务量，即每日电话录音时长，为您推荐合理的云端或本地私有服务器配置。循环智能已经积累了金融、房地产、汽车等多个行业的专属识别模型。

本地私有部署

云端私有部署

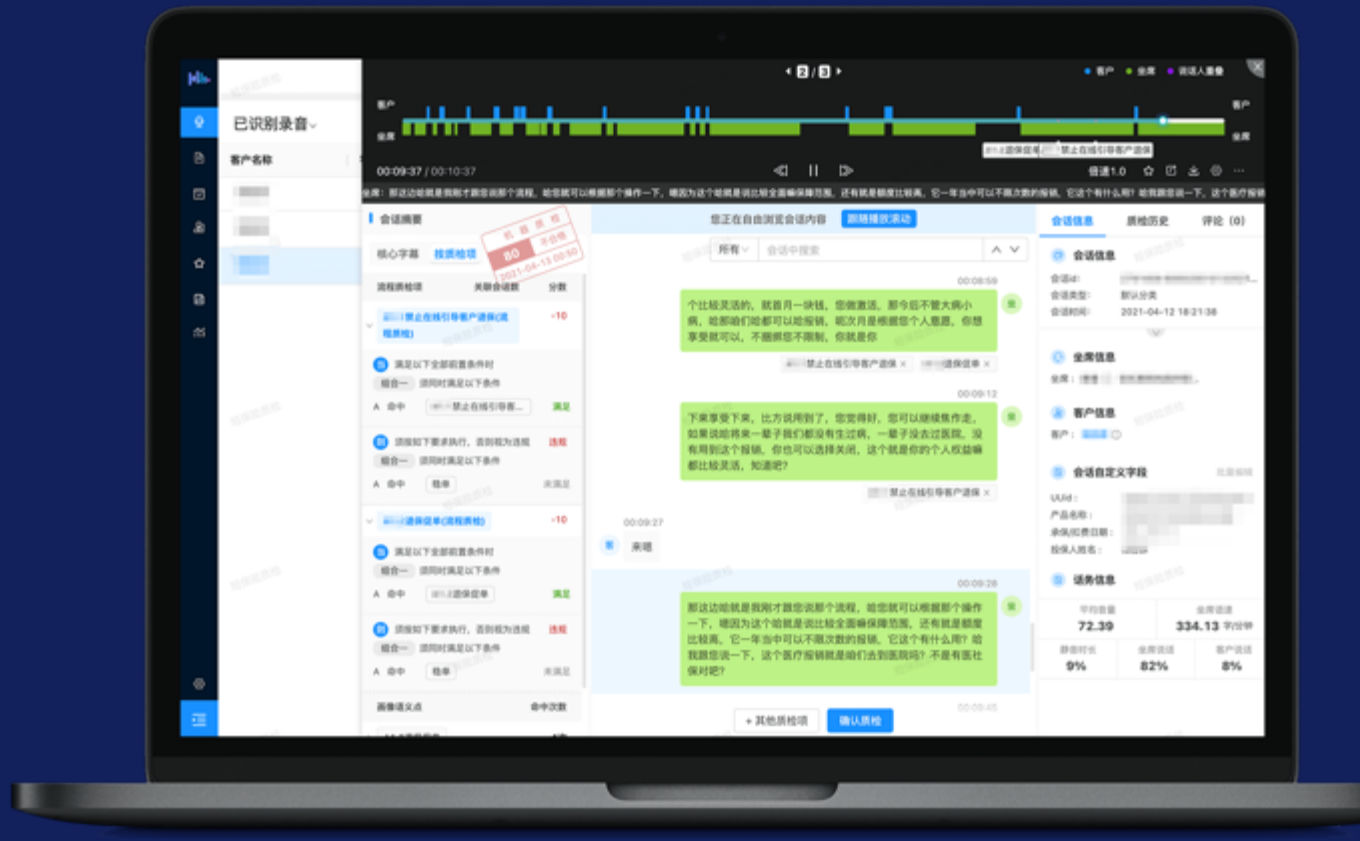
SaaS服务

循环智能
RECURRENT AI

每天转写数十万小时录音，可靠性 99.9%

循环智能每天转写的电话录音时长达到数万小时（仍在持续增长中），为客户提供“3个9”基础可靠性保障，最高可升级到4 个 9。此外，我们的 ASR 引擎支持自定义热词，以便为特定行业或公司的词和词组生成更准确的转录文本，例如产品名称、技术术语或个人姓名等。

99.9 %

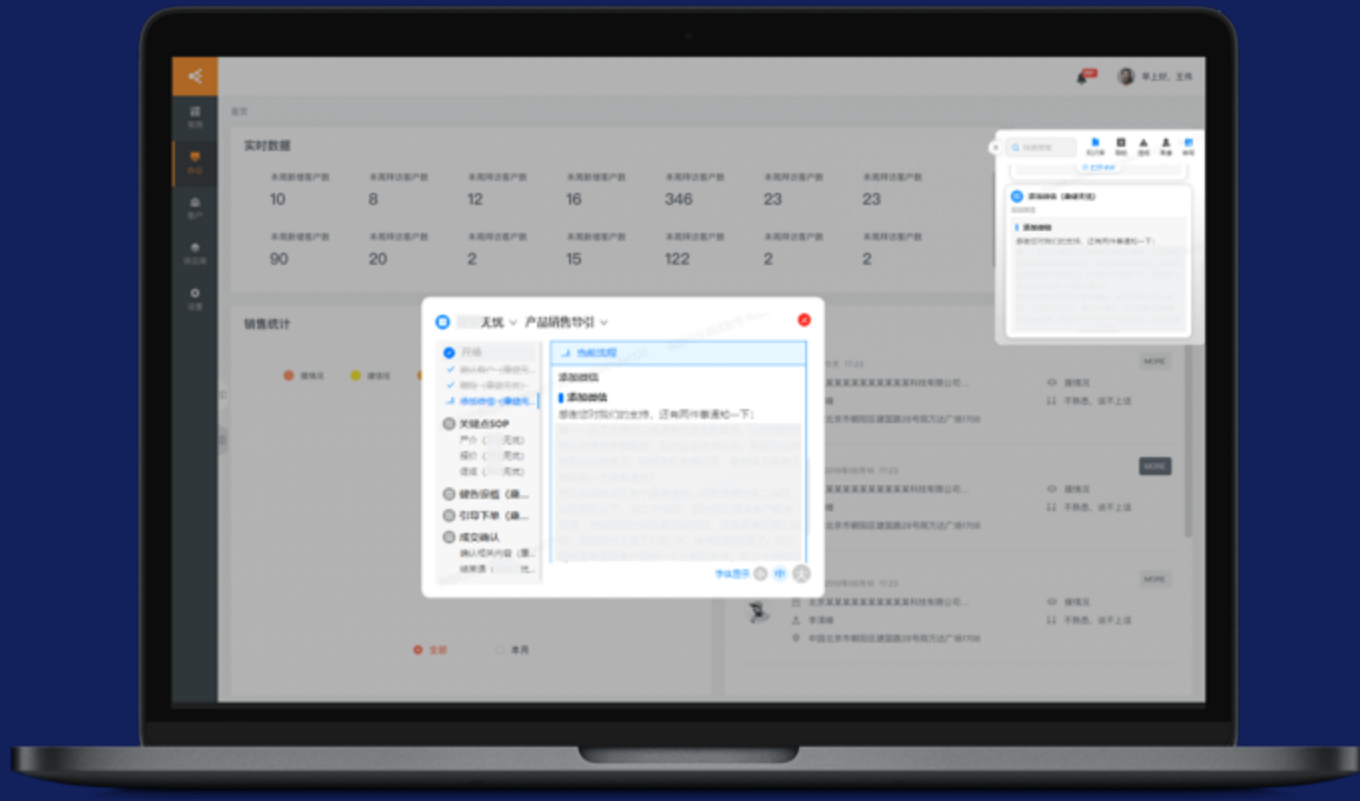


循环智能
RECURRENT AI

实时电话语音转写，成就实时沟通辅助

有了可靠的实时电话录音转写能力，才能在电销和在线客服场景，为销售和客服人员提供智能实时辅助工具，包括实时语音转写、流程导航、客户画像提取、话术推荐、知识点提示、合规质检等功能，降低上岗门槛，助力客户转化率和服务质量的提升。

real-time



循环智能
RECURRENT AI

更准

更快

更可靠

循环智能：银行、保险、房产和汽车等领域
的企业级沟通录音ASR行家